

SPRINGER
REFERENCE

Qingyun Duan
Florian Pappenberger
Andy Wood
Hannah L. Cloke
John C. Schaake
Editors

Handbook of Hydrometeorological Ensemble Forecasting

Handbook of Hydrometeorological Ensemble Forecasting

Qingyun Duan • Florian Pappenberger
Andy Wood • Hannah L. Cloke
John C. Schaake
Editors

Handbook of Hydrometeorological Ensemble Forecasting

With 443 Figures and 59 Tables



Springer

Editors

Qingyun Duan
Faculty of Geographical Science
Beijing Normal University
Beijing, China

Andy Wood
National Center for Atmospheric Research
Boulder, CO, USA

Florian Pappenberger
European Centre for Medium-Range
Weather Forecasts, ECMWF
Reading, UK

Hannah L. Cloke
Department of Meteorology
Reading University
Reading, UK

Department of Environmental Sciences and
Geography
Reading University
Reading, UK

John C. Schaake
U.S. National Weather Service (retired)
Annapolis, MD, USA

ISBN 978-3-642-39924-4 ISBN 978-3-642-39925-1 (eBook)
ISBN 978-3-642-39926-8 (print and electronic bundle)
<https://doi.org/10.1007/978-3-642-39925-1>

Library of Congress Control Number: 2018957996

© Springer-Verlag GmbH Germany, part of Springer Nature 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer-Verlag GmbH, DE, part of Springer Nature.

The registered company address is: Heidelberger Platz 3, 14197 Berlin, Germany

Preface

The first international workshop on the Hydrologic Ensemble Prediction Experiment (HEPEX) was held at the European Centre for Medium-Range Weather Forecasts (ECMWF) in Reading, United Kingdom, in 2004, with the aim to improve operational hydrometeorological forecasts to a standard that can be used with confidence by emergency and water resources managers. This workshop concluded that traditional deterministic, single-valued forecasts are inadequate to meet many needs of emergency and water resources managers and that the emerging ensemble forecasting approach is the way forward. The key advantage of ensemble forecasts over deterministic forecasts is that deterministic forecasts only predict one likely scenario of an impending event, but ensemble forecasts also give associated uncertainty information that is critical for making risk-based decisions.

Many HEPEX workshops have been held in different countries since 2004. These workshops bring meteorologists, hydrologists, and hydrometeorological forecast users together to review state-of-the-art advances in hydrometeorological ensemble forecasting, discuss challenges and future directions, and plan international collaborative activities. At the 2012 HEPEX workshop in Beijing, China, the HEPEX community came to realize that researchers and practitioners alike are yearning for a comprehensive reference book that covers all aspects of hydrometeorological ensemble forecasting by providing theoretical fundamentals for various ensemble forecasting methods and offering practical guidance for real-world ensemble applications.

Hydrometeorological ensemble forecasting is still a relatively young field, especially as it pertains to hydrologic applications. It was not until the turn of the century that the United States National Weather Service (NWS) and ECMWF became the first operational hydrologic forecasting centers to start experimenting with ensemble hydrologic forecasting services. Even today, deterministic forecasting is still the dominant practice in most operational hydrologic forecasting centers. But the impetus for adopting ensemble forecasting approaches is strong and growing stronger. There have been tremendous scientific and technological advances in numerical weather, climate, and hydrologic modeling; new forecasting methodologies; new observational technologies; new data assimilation techniques; and more powerful computational software and hardware. As a result of these advances, literature on ensemble forecasting has proliferated over recent years. But, the literature comprises

mostly specialist publications in various scientific journals. There is no single standard reference book, nor is there a textbook that presents the fundamental theories behind hydrometeorological ensemble forecasting and furnishes good examples of modern applications.

The lack of such a standard reference book presently hinders the advancement of research and practice in the field in several ways. First, researchers and practitioners who are new to the field lack organized source materials to help them understand this relatively new forecasting framework. Second, it creates difficulty in communicating ensemble forecast information to users and decision-makers who need to understand paradigm shifts in forecasting approaches. Third, it hinders the spread of hydrometeorological ensemble forecasting methodologies to regions where hydrometeorological forecasting is still basic or has yet to take root. The need for a good reference book on hydrometeorological ensemble forecasting is obvious and urgent.

Hence, the initiative to edit Springer's major reference book titled *Handbook of Hydrometeorological Ensemble Forecasting* was taken in 2013 to address a full spectrum of topics related to hydrometeorological ensemble forecasting, from the fundamental theories behind hydrometeorological modeling and forecasting to specific approaches to developing different ensemble forecasting components, to the applications, and user perspectives. This book is organized into eleven parts, and each part is further divided into several chapters.

The introductory part discusses broad rationales behind ensemble forecasting, historical perspectives including the HEPEX history, and an overview of hydrological predictability, scales and uncertainty issues. The second part is an overview of meteorological ensemble forecasting. It briefly reviews numerical weather and climate prediction model concepts as well as various procedures used in meteorological ensemble forecast generation. The third part is on calibration and post-processing of meteorological ensemble forecasts. It describes various statistical techniques for calibrating and post-processing (pre-processing for hydrological forecasting) raw meteorological ensemble forecasts. The fourth part is about hydrological methods for converting meteorological forecast outputs into hydrological model inputs. It describes different hydrological modeling approaches and various types of models, including lumped and distributed, conceptual type or physically based. The fifth part presents model parametric uncertainty analysis techniques. It discusses various issues related to model calibration and specific model calibration/validation techniques. The sixth part is about hydrometeorological observations and data assimilation techniques. This part addresses data requirements for hydrometeorological forecasting and presents various data assimilation techniques that can be used to improve the representation of initial and boundary conditions. The seventh part is on post-processing of hydrological model outputs and generation of ensemble hydrological forecast products. It covers different statistical post-processing techniques and various ensemble hydrological forecast products. The eighth part is on verification systems for ensemble forecasts. This part describes various metrics for measuring the performance of ensemble forecast systems against observations. The ninth part focuses on topics related to communicating probabilistic forecasting information to users and decision-makers. The tenth part presents different

applications, showcasing how ensemble systems have been used in practice for a variety of purposes such as meteorological and river forecasts, water resources management, water quality monitoring, and drought monitoring, among others. The last part presents fundamental mathematical and statistical techniques that are needed for ensemble forecasting. The book is intended for didactical as well as reference purposes. The intended audience of this book includes university researchers, graduate students, practicing hydrometeorologists, civil and environmental engineers, and emergency and water resources managers.

This eleven-part, fifty-one-chapter, two-volume book is the result of joint efforts of many authors, reviewers, and editors. Special thanks go to Jutta Thielen, who has played a critical role in the book planning and editing stages. Finally, we would like to convey our sincere appreciation to the editorial support staff from Springer Nature, especially Keerthi Sudevan, Sunaina Dadhwal, and Lijuan Wang, for their excellent collaboration and persistent help during the entire process.

Qingyun Duan
Florian Pappenberger
Andy Wood
Hannah L. Cloke
John C. Schaake
Editors

Contents

Volume 1

Part I Introduction	1
Hydrological Predictability, Scales, and Uncertainty Issues	3
Joshua K. Roundy, Qingyun Duan, and John C. Schaake	
Part II Overview of Meteorological Ensemble Forecasting	33
Overview of Weather and Climate Systems	35
Huiling Yuan, Zoltan Toth, Malaquias Peña, and Eugenia Kalnay	
Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation	67
Zhaoxia Pu and Eugenia Kalnay	
Ensemble Methods for Meteorological Predictions	99
Jun Du, Judith Berner, Roberto Buizza, Martin Charron, Peter Houtekamer, Dingchen Hou, Isidora Jankov, Mu Mu, Xuguang Wang, Mozheng Wei, and Huiling Yuan	
Major Operational Ensemble Prediction Systems (EPS) and the Future of EPS	151
Roberto Buizza, Jun Du, Zoltan Toth, and Dingchen Hou	
Intraseasonal to Interannual Climate Variability and Prediction	195
Malaquias Peña, L. Gwen Chen, and Huug van den Dool	
Part III Post-processing of Meteorological Ensemble Forecasting for Hydrological Applications	237
Hydrological Challenges in Meteorological Post-processing	239
Fredrik Wetterhall and Paul Smith	

Application to Post-processing of Meteorological Seasonal Forecasting	255
Andrew Schepen, Q. J. Wang, and David E. Robertson	
Multi-model Combination and Seamless Prediction	285
Stephan Hemri	
Part IV Hydrological Models	309
Hydrological Cycles, Models, and Applications to Forecasting	311
Sharad K. Jain and Vijay P. Singh	
Black-Box Hydrological Models	341
Chong-Yu Xu, Lihua Xiong, and Vijay P. Singh	
Conceptual Hydrological Models	389
Zhaofei Liu, Yamei Wang, Zongxue Xu, and Qingyun Duan	
Distributed Hydrological Models	413
Yangbo Chen	
Land Surface Hydrological Models	437
Michael B. Ek	
Part V Model Parameter Estimation and Uncertainty Analysis ...	479
Parameter Estimation and Predictive Uncertainty Quantification in Hydrological Modelling	481
Dmitri Kavetski	
Methods to Estimate Optimal Parameters	523
Tiantian Yang, Kuolin Hsu, Qingyun Duan, Soroosh Sorooshian, and Chen Wang	
Uncertainty Quantification of Complex System Models: Bayesian Analysis	563
Jasper A. Vrugt and Elias C. Massoud	
Sensitivity Analysis Methods	637
Yanjun Gan and Qingyun Duan	
Part VI Observation and Data Assimilation	673
Fundamentals of Data Assimilation and Theoretical Advances	675
Hamid Moradkhani, Grey S. Nearing, Peyman Abbaszadeh, and Sahani Pathiraja	
Soil Moisture Data Assimilation	701
Gabrielle Jacinthe Maria De Lannoy, Patricia de Rosnay, and Rolf Helmut Reichle	

Assimilation of Streamflow Observations	745
Seong Jin Noh, Albrecht H. Weerts, Oldrich Rakovec, Haksu Lee, and Dong-Jun Seo	

Volume 2

Part VII Post-processing of Hydrological Ensemble Forecasts ...	781
--	------------

Motivation and Overview of Hydrological Ensemble	
Post-processing	783
Thomas M. Hopson, Andy Wood, and Albrecht H. Weerts	
Short-Range Ensemble Forecast Post-processing	795
Marie-Amélie Boucher, Emmanuel Roulin, and Vincent Fortin	
Seasonal Ensemble Forecast Post-processing	819
Andy Wood, A. Sankarasubramanian, and Pablo Mendoza	

Part VIII Verification of Hydrometeorological Ensemble	
Forecasts	847

Attributes of Forecast Quality	849
A. Allen Bradley, Julie Demargne, and Kristie J. Franz	
Verification Metrics for Hydrological Ensemble Forecasts	893
François Anctil and Maria-Helena Ramos	

Verification of Meteorological Forecasts for Hydrological	
Applications	923
Eric Gilleland, Florian Pappenberger, Barbara Brown, Elizabeth Ebert, and David Richardson	

Verification of Short-Range Hydrological Forecasts	953
Katharina Liechti and Massimiliano Zappa	

Verification of Medium- to Long-Range Hydrological Forecasts	977
Luc Perreault, Jocelyn Gaudet, Louis Delorme, and Simon Chatelain	

Application of Hydrological Forecast Verification Information	1013
Kevin Werner, Jan S. Verkade, and Thomas C. Pagano	

Part IX Communication and Use of Ensemble Forecasts for	
Decision Making	1035

Overview of Forecast Communication and Use of Ensemble	
Hydrometeorological Forecasts	1037
Jutta Thielen-del Pozo and Michael Bruen	

Present and Future Requirements for Using and Communicating Hydrometeorological Ensemble Prediction Systems for Short-, Medium-, and Long-Term Applications	1047
Geoff Pegram, Damien Raynaud, Eric Sprokkereef, Martin Ebel, Silke Rademacher, Jonas Olsson, Cristina Alionte-Eklund, Barbro Johansson, Göran Lindström, and Henrik Spångmyr	
Best Practice in Communicating Uncertainties in Flood Management in the USA	1093
Robert K. Hartman	
Saving Lives: Ensemble-Based Early Warnings in Developing Nations	1109
Feyera A. Hirpa, Kayode Fagbemi, Ernest Afiesimam, Hassan Shuaib, and Peter Salamon	
Communicating and Using Ensemble Flood Forecasts in Flood Incident Management: Lessons from Social Science	1131
David Demeritt, Elisabeth M. Stephens, Laurence Créton-Cazanave, Céline Lutloff, Isabelle Ruin, and Sébastien Nobert	
Overview of Communication Strategies for Uncertainty in Hydrological Forecasting in Australia	1161
Narendra Kumar Tuteja, Senlin Zhou, Julien Lerat, Q. J. Wang, Daehyok Shin, and David E. Robertson	
Part X Ensemble Forecast Application and Showcases	1179
Introduction to Ensemble Forecast Applications and Showcases	1181
Massimiliano Zappa, S. J. van Andel, and Hannah L. Cloke	
Hydrological Ensemble Prediction Systems Around the Globe	1187
Florian Pappenberger, Thomas C. Pagano, J. D. Brown, Lorenzo Alfieri, D. A. Lavers, L. Berthet, F. Bressand, Hannah L. Cloke, M. Cranston, J. Danhelka, J. Demargne, N. Demuth, C. de Saint-Aubin, P. M. Feikema, M. A. Fresch, R. Garçon, A. Gelfan, Y. He, Y. -Z. Hu, B. Janet, N. Jurdy, P. Javelle, L. Kuchment, Y. Laborda, E. Langsholt, M. Le Lay, Z. J. Li, F. Mannessiez, A. Marchandise, R. Marty, D. Meißner, D. Manful, D. Organde, V. Pourret, Silke Rademacher, Maria-Helena Ramos, D. Reinbold, S. Tibaldi, P. Silvano, Peter Salamon, D. Shin, C. Sorbet, Eric Sprokkereef, V. Thiemig, Narendra Kumar Tuteja, S. J. van Andel, Jan S. Verkade, B. Vehviläinen, A. Vogelbacher, Fredrik Wetterhall, Massimiliano Zappa, R. E. Van der Zwan, and Jutta Thielen-del Pozo	
Flash Flood Forecasting Based on Rainfall Thresholds	1223
Lorenzo Alfieri, Marc Berenguer, Valentin Knechtli, Katharina Liechti, Daniel Sempere-Torres, and Massimiliano Zappa	

Medium Range Flood Forecasting Example EFAS	1261
Jutta Thielen-del Pozo, Peter Salamon, Peter Burek, Florian Pappenberger, C. Alionte Eklund, Eric Sprokkereef, M. Hazlinger, M. Padilla Garcia, and R. Garcia-Sanchez	
Seasonal Drought Forecasting on the Example of the USA	1279
Eric F. Wood, Xing Yuan, Joshua K. Roundy, Ming Pan, and Lifeng Luo	
Ensemble Streamflow Forecasts for Hydropower Systems	1289
Marie-Amélie Boucher and Maria-Helena Ramos	
Hydropower Forecasting in Brazil	1307
Carlos E. M. Tucci, Walter Collischonn, Fernando Mainardi Fan, and Dirk Schwanenberg	
New York City's Operations Support Tool: Utilizing Hydrologic Forecasts for Water Supply Management	1329
James Porter, Gerald Day, John C. Schaake, and Lucien Wang	
Probabilistic Shipping Forecast	1371
Dennis Meißner and Bastian Klein	
Probabilistic Inundation Forecasting	1385
A. Mueller, C. Baugh, P. Bates, and Florian Pappenberger	
Challenges of Decision Making in the Context of Uncertain Forecasts in France	1399
Caroline Wittwer, C. de Saint-Aubin, and C. Ardilouze	
Hydrological Ensemble Prediction Applied in China	1413
Guangsheng Wang, Zhijie Yin, Jianqing Yang, and Yuhong Yan	
Part XI Mathematical and Statistical Fundamentals for Hydrometeorological Ensemble Forecasting	1427
Probability and Statistical Theory for Hydrometeorology	1429
Zengchao Hao, Vijay P. Singh, and Wei Gong	
Estimation of Probability Distributions for Hydrometeorological Applications	1463
Grey S. Nearing	
Regression Techniques Used in Hydrometeorology	1485
Wei Gong	
Index	1513

About the Editors



Qingyun Duan is currently a professor and chief scientist of Hydrology and Water Resources in the Faculty of Geographical Science at Beijing Normal University (BNU) in China. Prior to his current position, he worked at US NOAA Hydrology Laboratory from 1991 to 2003 and US Department of Energy's Lawrence Livermore National Laboratory from 2004 to 2009. His research interests include hydrology and water resources, hydrological model development and calibration, hydrometeorological ensemble forecasting, and uncertainty quantification for large complex system models. He is involved with the development of several operational hydrometeorological models used in the US National Weather Service. He is also the developer of the Shuffled Complex Evolution method, one of the most popular optimization methods used in hydrological model calibration today. His recent work includes the development of uncertainty quantification software platform for large complex system models, Uncertainty Quantification Python Laboratory (UQ-PyL), and the BNU Hydrological Ensemble Prediction System (BNU-HEPS). He has authored or co-authored more than 150 peer-reviewed articles, including more than 120 papers in Science Citation Indexed (SCI) journals. Dr. Duan has been active in many international scientific activities, including serving as the co-leader of the Model Parameter Estimation Experiment (MOPEX) and a member of the scientific steering committees of the Global Energy and Water Exchanges (GEWEX) project and the Hydrological Ensemble Prediction Experiment (HEPEX). He was or is serving as an editor or editorial board member for numerous scientific journals, including the *Bulletin of the American Meteorological Society* and *Water*.

Resources Research. Dr. Duan is a recipient of Chinese Government “One Thousand Talents Program” Award and a fellow of the American Geophysical Union and the American Meteorological Society.



Florian Pappenberger is director of Forecasts at the European Centre for Medium-Range Weather Forecasts. The Forecast Department at the ECMWF has a strong user focus and undertakes production of forecasts, forecast evaluation and diagnostics, development of forecast products and applications, software development, catalogue and data services, and outreach and training. Florian has a scientific background in the forecasting of weather-driven natural hazards including floods, droughts, windstorms, forest fires, and impacts on human health. He has over 10 years of expertise in operational probabilistic forecasting, extreme value statistics, and numerical model system development at the ECMWF. He was responsible for the development and implementation of the operational center of the Copernicus Emergency Service – Early Warning Systems (floods). Florian is the author of over 150 publications, has won several scientific awards, and is visiting professor at the University of Bristol. He is an elected fellow of the Royal Geographical Society and the Royal Meteorological Society and a member of several other professional bodies including the HEPEX, British Hydrological Society, EGU, AGU, EMS, and AMS. He is on the editorial board of several international scientific journals and regularly advises on international committees including the WMO.



Dr. Andy Wood
Research Applications Laboratory
National Center for Atmospheric Research
Boulder, Colorado, USA

Andy Wood is a Project Scientist with the National Center for Atmospheric Research (NCAR) Research Applications Laboratory, Boulder, Colorado. His research focuses on water resources, systems engineering, hydrologic modeling, monitoring and prediction,

and the assessment of climatic change impacts on hydrology and water resources. He has over 20 years of experience in these areas in academic, private, and public institutions. He worked as a research assistant professor at the University of Washington (UW) for 4 years, advising graduate students and teaching several classes. He then worked as the lead scientist of private firm 3TIER, Inc., focusing on forecasting and assessment of hydropower, solar and wind energy for US-based and international clients. He later spent 3 years as a Development and Operations Hydrologist with two US National Weather Service River Forecast Centers, where he worked on the transition of research to operations and helped manage the operational forecasting teams. Major accomplishments include the development of the widely used BCSD and quantile mapping techniques for statistical downscaling of climate projections, and the creation of operational streamflow and drought monitoring and forecasting systems (such as the UW Surface Water Monitor) at regional to continental scales.

Dr. Wood chaired the Hydrology Committee of the American Meteorological Society (AMS) from 2011 to 2013 and is currently an Editor with the AMS *Journal of Hydrometeorology*. Since 2012, he has co-chaired the international Hydrologic Ensemble Prediction Experiment (<http://www.hepex.org>), which focuses on advancing the adoption of ensemble streamflow forecasting for water and emergency management, and he has organized over a dozen national and international conferences on hydrology, forecasting, and water resources. Andy is also the co-chair of the US CLIVAR Predictions, Predictability and Applications Interface Panel, serves on the Advisory Board for the ECMWF European and Global Flood Awareness Systems, and is a Task Lead for the WMO global Hydrologic Status and Outlook System. He is also an Adjoint Professor in the University of Colorado Department of Civil, Environmental and Architectural Engineering.

His current work centers on practical applications of scientific advances in hydrologic, weather, and climate modeling, and prediction and projection to improve our understanding and management of water, energy, and terrestrial ecosystems.

Education

B.A., Amherst College, Amherst, MA, English, 1988
MSE, University of Washington, Civil and Env. Engineering, Water Resources and Systems, 1995
Ph.D., University of Washington, Civil and Env. Engineering, Hydrology, 2003



Hannah L. Cloke is a hydrologist and physical geographer specializing in land surface modeling, flood forecasting, applications of numerical weather predictions, and catchment hydrology. Her current research focuses on the theoretical and practical development of early warning systems for natural hazards, particularly for floods and droughts and disaster risk management.

Hannah is visiting professor in the Department of Earth Sciences at Uppsala University, Sweden. She is an elected fellow of the Royal Geographical Society and the Royal Meteorological Society. She works closely with the Environment Agency, the European Centre for Medium-Range Weather Forecasts (ECMWF), and the UK Met Office, as well as a wide range of other national and international partners. Hannah advises government on national and international flooding incidents and provides expert commentary in the media. She is executive editor of *Hydrology and Earth System Sciences* and an active member of the HEPEX initiative.

Hannah was awarded the 2015 NERC Early Career Impact Award, was runner-up in the Guardian University Awards for Impact 2016, and has been awarded the 2018 Plinius Medal of the European Geosciences Union in recognition of her outstanding interdisciplinary research in natural hazards.

Hannah obtained a B.Sc. (1999) and Ph.D. (2003) in Geography from the University of Bristol, UK. She then worked at the European Commission's Joint Research Centre in Ispra, Italy, on the European Flood Alert System and then from 2004 lectured in the Department of Geography at King's College London, UK. In 2012, she moved to the University of Reading, to a joint post between the Department of Geography and

Environmental Science and the Department of Meteorology, where she is now professor of Hydrology and co-director of Water@Reading.



Dr. John C. Schaake Retired NOAA/NWS Senior Executive and Senior Scientist, Retired Independent Hydrological Consultant

Dr. Schaake received his Ph.D. from the Johns Hopkins University in 1965. From 1965 to 1966, he was a postdoc at Harvard University's Harvard Water Program for systems analysis applications in water resources management. From 1966 to 1968, he was a professor of Environmental and Systems Engineering at the University of Florida. From 1968 to 1974, he was a professor of Civil Engineering at the MIT where he worked on systems engineering applications for water management. These experiences led him to realize that water managers needed river forecasts, but these were not widely available.

In 1974, Dr. Schaake joined NOAA's National Weather Service as deputy director of the Hydrologic Research Laboratory where he led the early development of the NWS River Forecast System that was soon implemented nationally in all NWS River Forecast Centers (RFCs) to provide river forecasts needed by water managers. In 1977, he became chief of the NWS Hydrologic Services Division to support NWS RFC operations. Beginning in 1987, as a senior scientist, he led the development of hydrologic ensemble forecast techniques. These were first used in limited RFC operations in the early 2000s. They are continually being improved, are beginning to become an integral part of water system management operations in some parts of the United States, and are being implemented nationwide in all NWS RFCs.

After retiring from the NWS in 2001, Dr. Schaake became an independent hydrologic consultant to clients including the NWS Hydrology Program, the New York City Department of Environmental Prediction, the California Department of Water Resources, and HydroLogics, a consulting company specializing in water resources systems analysis applications.

He is a fellow of the American Geophysical Union and the American Meteorological Society. He was a co-founder of HEPEX at the ECMWF in 2004.

About the Section Editors

James Brown Hydrologic Solutions Limited, Southampton, UK



Michael Bruen
School of Civil Engineering
UCD Dooge Centre for Water Resources Research
Belfield, Ireland



Julie Demargne
HYDRIS Hydrologie
Saint Mathieu de Tréviers, France



Wei Gong
State Key Laboratory of Earth Surface Processes and
Resource Ecology
Beijing Normal University
Beijing, China
Institute of Land Surface System and Sustainable
Development
Beijing Normal University
Beijing, China

**Thomas M. Hopson**

Research Applications Laboratory
National Center for Atmospheric Research
Boulder, CO, USA

**Kuolin Hsu**

Department of Civil and Environmental Engineering
Center for Hydrometeorology and Remote Sensing
University of California
Irvine, CA, USA

**Dmitri Kavetski**

School of Civil, Environmental and Mining Engineering
University of Adelaide
Adelaide, SA, Australia

School of Engineering
University of Newcastle
Callaghan, NSW, Australia

Department of Systems Analysis
Integrated Assessment and Modelling (SIAM)
Swiss Federal Institute of Aquatic Science and Technology,
Dübendorf, Switzerland

Yuqiong Liu NASA Goddard Space Flight Center, Washington, DC, USA

**Lifeng Luo**

Department of Geography
Environment, and Spatial Sciences
Michigan State University
East Lansing, MI, USA

**Hamid Moradkhani**

Director, Center for Complex Hydrosystems Research
Alton N. Scott Professor of Engineering
Department of Civil, Construction and Environmental
Engineering
The University of Alabama
Tuscaloosa, AL, USA

**Maria-Helena Ramos**

IRSTEA – National Research Institute of Science and
Technology for Environment and Agriculture
Antony Cedex, France

**Vijay P. Singh**

Caroline and William N. Lehrer Distinguished Chair in
Water Engineering
Department of Biological and Agricultural Engineering
and Zachry Department of Civil Engineering
Texas A&M University Engineering
College Station, TX, USA

**Zoltan Toth**

Global Systems Division
NOAA/OAR/ESRL
Boulder, CO, USA

**Schalk Jan van Andel**

IHE Delft Institute for Water Education
Delft, The Netherlands

**Nathalie Voisin**

Water Resources Engineer, Hydrology Group
Pacific Northwest National Laboratory
Seattle, WA, USA

Civil and Environmental Engineering Department
University of Washington
Seattle, WA, USA

**Albrecht H. Weerts**

Operational Water Management
Inland Water Systems, Deltares
Delft, The Netherlands

Hydrology and Quantitative Water Management Group
Wageningen University and Research
Wageningen, The Netherlands

**Zongxue Xu**

Beijing Key Laboratory of Urban Hydrological Cycle
and Sponge City Technology
College of Water Sciences
Beijing Normal University
Beijing, China



Huiling Yuan
School of Atmospheric Sciences
Nanjing University
Nanjing, China



Massimiliano Zappa
Mountain Hydrology and Mass Movements
Swiss Federal Institute for Forest
Snow and Landscape Research WSL
Birmensdorf, Switzerland

Contributors

Peyman Abbaszadeh Department of Civil, Construction and Environmental Engineering, The University of Alabama, Tuscaloosa, AL, USA

Ernest Afiesimam Nigerian Meteorological Agency (NiMet), Abuja, Nigeria

Lorenzo Alfieri Directorate for Space, Security and Migration, European Commission – Joint Research Centre, Ispra, VA, Italy

C. Alionte Eklund Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Cristina Alionte-Eklund Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

François Anctil Département de génie civil et de génie des eaux, Université Laval, Québec, QC, Canada

C. Ardilouze Météo France, Toulouse, France

P. Bates Department of Geography, School of Geographical Sciences, University of Bristol, Bristol, UK

C. Baugh European Centre for Medium-Range Forecast (ECMWF), Reading, UK

Marc Berenguer Center of Applied Research in Hydrometeorology, Universitat Politècnica de Catalunya, Barcelona, Spain

Judith Berner National Centers for Atmospheric Research, Boulder, CO, USA

L. Berthet Loire river Flood Forecasting Centre, Orléans, Italy

Marie-Amélie Boucher Civil Engineering Department, Université de Sherbrooke, Sherbrooke, QC, Canada

A. Allen Bradley IIHR–Hydroscience and Engineering, The University of Iowa, Iowa City, IA, USA

F. Bressand Service de Prévision des Crues Grand Delta, Nîmes, France

Barbara Brown Research Applications Laboratory, Weather Systems and Assessment Program, National Center for Atmospheric Research NCAR, Boulder, CO, USA

J. D. Brown Hydrologic Solutions Limited, Southampton, UK

Michael Bruen UCD Dooge Centre for Water Resources Research, UCD School of Civil Engineering, Dublin, Ireland

Roberto Buizza European Centre for Medium Range Weather Forecasts, Reading, UK

Peter Burek Water Program (WAT), International Institute for Applied System Analysis (IIASA), Laxenburg, Austria

Martin Charron Canadian Meteorological Center, Environmental Canada, Montreal, Canada

Simon Chatelain McGill University, Montreal, QC, Canada

L. Gwen Chen Earth System Science Interdisciplinary Center/Cooperative Institute for Climate and Satellites, University of Maryland, College Park, MD, USA
CPC/NCEP/NWS/NOAA, College Park, MD, USA

Yangbo Chen Department of Water Resources and Environment, Sun Yat-sen University, Guangzhou, Guangdong Province, China

Hannah L. Cloke Department of Meteorology, Reading University, Reading, UK
Department of Environmental Sciences and Geography, Reading University, Reading, UK

Walter Collischonn Institute of Hydraulic Research, Federal University of Rio Grande do Sul, Porto Alegre-RS, Brazil

M. Cranston RAB Consultants/University of Dundee, Stirling, Italy

Laurence Créton-Cazanave Université Paris Est Marne-la-Vallée, Labex Futurs Urbains (LATTIS, LEESU, Lab'Urba), Marne-la-Vallée, France

J. Danhelka Czech Hydrometeorological Institute, Prague, Czech Republic

Gerald Day RTI International, Ft. Collins, CO, USA

Gabrielle Jacinthe Maria De Lannoy NASA Goddard Space Flight Center, Code 610.1, Greenbelt, MD, USA

KU Leuven, Department of Earth and Environmental Sciences, Leuven, Belgium

Patricia de Rosnay Data Assimilation Section, European Center for Medium-Range Weather Forecasts, Reading, Berkshire, UK

C. de Saint-Aubin Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI), Toulouse, France

Louis Delorme IREQ Hydro-Québec Research Institute, Varennes, QC, Canada

Julie Demargne HYDRIS Hydrologie, Saint Mathieu de Tréviers, France

David Demeritt Department of Geography, King's College London, Strand, London, UK

N. Demuth Landesamt für Umwelt, Rhineland Palatinate, Mainz, Germany

Jun Du Environmental Modeling Center/National Centers for Environmental Prediction (NCEP), NOAA, College Park, MD, USA

Qingyun Duan Faculty of Geographical Science, Beijing Normal University, Beijing, China

Martin Ebel Bundesamt für Umwelt, Ittigen, Switzerland

Elizabeth Ebert Research and Development Branch, Bureau of Meteorology, BoM, Melbourne, Australia

Michael B. Ek National Center for Atmospheric Research, Boulder, CO, USA

Kayode Fagbemi National Emergency Management Agency (NEMA), Abuja, Nigeria

Fernando Mainardi Fan Institute of Hydraulic Research, Federal University of Rio Grande do Sul, Porto Alegre-RS, Brazil

P. M. Feikema Bureau of Meteorology, Melbourne, VIC, Australia

Vincent Fortin Environment and Climate Change Canada, Dorval, QC, Canada

Kristie J. Franz Department of Geological and Atmospheric Sciences, Iowa State University, Ames, IA, USA

M. A. Fresch Office of Water Prediction, U.S. National Weather Service, Silver Spring, MD, USA

Yanjun Gan State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

M. Padilla Garcia REDIAM, Sevilla, Spain

R. Garçon EDF DTG, Grenoble, France

R. Garcia-Sanchez ELIMCO SISTEMAS S.L., Sevilla, Spain

Jocelyn Gaudet IREQ Hydro-Québec Research Institute, Varennes, QC, Canada

A. Gelfan Water Problems Institute of Russian Academy of Sciences (WPI RAS), Moscow, Russia

Eric Gilleland Research Applications Laboratory, Weather Systems and Assessment Program, National Center for Atmospheric Research NCAR, Boulder, CO, USA

Wei Gong State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China

Institute of Land Surface System and Sustainable Development, Faculty of Geographical Science, Beijing Normal University, Beijing, China

Zengchao Hao College of Water Sciences, Beijing Normal University, Beijing, China

Robert K. Hartman California-Nevada River Forecast Center, NOAA, National Weather Service, Sacramento, CA, USA

M. Hazlinger Slovak Hydrometeorological Institute, Bratislava, Slovakia

Y. He Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, UK

Stephan Hemri Department of Computational Statistics, HITS gGmbH, Heidelberg, Germany

Feyera A. Hirpa European Commission, Joint Research Centre (JRC), Institute for Environment and Sustainability (IES), Climate Risk Management Unit, Ispra, VA, Italy

Thomas M. Hopson Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA

Dingchen Hou Environmental Modeling Center/National Centers for Environmental Prediction (NCEP), NOAA, College Park, MD, USA

Dingchen Hou SAIC at NOAA/NWS/NCEP/EMC, Camp Springs, MA, USA

Peter Houtekamer Canadian Meteorological Center, Environmental Canada, Montreal, Canada

Kuolin Hsu Civil and Environmental Engineering, The Henry Samueli School of Engineering, University of California, Irvine, CA, USA

Y. -Z. Hu German Federal Institute of Hydrology (BfG), Koblenz, Germany

Sharad K. Jain Jal Vigyan Bhawan, National Institute of Hydrology, Roorkee, Uttarakhand, India

B. Janet Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI), Toulouse, France

Isidora Jankov Cooperative Institute for Research in the Atmosphere, Earth System Research Lab (ESRL)/NOAA, Boulder, CO, USA

P. Javelle Irstea, OHAX Hydrology Unit, Aix-en-Provence, France

Barbro Johansson Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

N. Jurdy Service de Prévision des Crues Meuse-Moselle, Metz, France

Eugenia Kalnay Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA

Dmitri Kavetski School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, SA, Australia

School of Engineering, University of Newcastle, Callaghan, NSW, Australia

Department of Systems Analysis, Integrated Assessment and Modelling (SIAM), Eawag, Swiss Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland

Bastian Klein Department Water Balance, Forecasting and Predictions, Federal Institute of Hydrology (BfG), Koblenz, Germany

Valentin Knechtl Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

L. Kuchment Water Problems Institute of Russian Academy of Sciences (WPI RAS), Moscow, Russia

Y. Laborda Service de Prévision des Crues Grand Delta, Nîmes, France

E. Langsholt Ministry of Petroleum and Energy, Norwegian Water Resources and Energy Directorate, Hydrology Department (NVE), Oslo, Norway

D. A. Lavers European Centre for Medium Range Weather Forecasts, Reading, UK

M. Le Lay EDF DTG, Grenoble, France

Haksu Lee National Oceanic and Atmospheric Administration, Silver Spring, MD, USA

Julien Lerat Bureau of Meteorology, Canberra, ACT, Australia

Z. J. Li German Federal Institute of Hydrology (BfG), Koblenz, Germany

Katharina Liechti Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

Göran Lindström Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Zhaofei Liu Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

Lifeng Luo Department of Geography, Michigan State University, East Lansing, MI, USA

Céline Lutolf Université Grenoble 1, PACTE UMR 5194 (CNRS, IEPG, UJF, UPMF), Grenoble, France

D. Manful Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, UK

F. Mannessiez Service de Prévision des Crues Grand Delta, Nîmes, France

A. Marchandise Service de Prévision des Crues Méditerranée Ouest, Carcassonne, France

R. Marty Loire-Cher-Indre Flood Forecasting Centre, Orléans, France

Elias C. Massoud Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA, USA

D. Meinßner German Federal Institute of Hydrology (BfG), Koblenz, Germany

Dennis Meinßner Department Water Balance, Forecasting and Predictions, Federal Institute of Hydrology (BfG), Koblenz, Germany

Pablo Mendoza Advanced Mining Technology Center (AMTC), Universidad de Chile, Santiago de Chile, Chile

Hamid Moradkhani Department of Civil, Construction and Environmental Engineering, The University of Alabama, Tuscaloosa, AL, USA

Mu Mu Institute of Atmospheric Sciences, Fudan University, Shanghai, China

A. Mueller Geography and Environmental Science Department, University of Reading and European Centre for Medium-Range Forecast (ECMWF), Reading, UK

Grey S. Nearing Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA

Sébastien Nobert School of Earth and Environment, University of Leeds, Leeds, UK

Seong Jin Noh Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX, USA

Jonas Olsson Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

D. Organde HYDRIS Hydrologie, Saint Mathieu de Tréviers, France

Thomas C. Pagano Bureau of Meteorology, Melbourne, VIC, Australia

Ming Pan Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA

Florian Pappenberger European Centre for Medium-Range Weather Forecasts, ECMWF, Reading, UK

Sahani Pathiraja Institute for Mathematics, University of Potsdam, Potsdam, Germany

Malaquias Peña Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA

Geoff Pegram Satellite Applications and Hydrology Group, School of Civil Engineering, Surveying and Construction Management, University of KwaZulu-Natal, Durban, South Africa

Luc Perreault IREQ Hydro-Québec Research Institute, Varennes, QC, Canada

James Porter New York City Department of Environmental Protection, Bureau of Water Supply, New York, NY, USA

V. Pourret Météo-France, Toulouse, France

Zhaoxia Pu Department of Atmospheric Sciences, University of Utah, Salt Lake City, UT, USA

Silke Rademacher German Federal Institute of Hydrology (BfG), Koblenz, Germany

Oldrich Rakovec Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

Faculty of Environmental Sciences, Czech University of Life Sciences, Prague, Czech Republic

Maria-Helena Ramos IRSTEA, National Research Institute of Science and Technology for Environment and Agriculture, UR HBAN, Antony, France

Damien Raynaud Université Joseph Fourier, Grenoble, France

Rolf Helmut Reichle NASA Goddard Space Flight Center, Code 610.1, Greenbelt, MD, USA

D. Reinbold Loire-Cher-Indre Flood Forecasting Centre, Orléans, France

David Richardson European Centre for Medium-Range Weather Forecasts, ECMWF, Reading, UK

David E. Robertson CSIRO Land and Water, Clayton, VIC, Australia

Emmanuel Roulin Institut Royal Météorologique de Belgique, Bruxelles, Belgium

Joshua K. Roundy Department of Civil, Environmental and Architectural Engineering, University of Kansas, Lawrence, KS, USA

Isabelle Ruin Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE), CNRS, Grenoble, France

Peter Salamon European Commission, Joint Research Centre (JRC), Institute for Environment and Sustainability (IES), Climate Risk Management Unit, Ispra, VA, Italy

A. Sankarasubramanian Department of Civil Construction and Environmental Engineering, North Carolina State University, Raleigh, NC, USA

John C. Schaake U.S. National Weather Service (retired), Annapolis, MD, USA

Andrew Schepen CSIRO Land and Water, Dutton Park, QLD, Australia

Dirk Schwanenberg Institute of Hydraulic Engineering and Water Resources Management, Universität Duisburg-Essen, Essen, Germany

Daniel Sempere-Torres Center of Applied Research in Hydrometeorology, Universitat Politècnica de Catalunya, Barcelona, Spain

Dong-Jun Seo Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX, USA

D. Shin Bureau of Meteorology, Melbourne, VIC, Australia

Daehyok Shin Bureau of Meteorology, Docklands, VIC, Australia

Hassan Shuaib University of Abuja, Abuja, Nigeria

P. Silvano ARPA Emilia Romagna, Parma, Italy

Vijay P. Singh Department of Biological and Agricultural Engineering and Zachry Department of Civil Engineering, Texas A and M University, College Station, TX, USA

Paul Smith Forecast Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

C. Sorbet Météo-France, Toulouse, France

Soroosh Sorooshian University of California, Irvine, CA, USA

Henrik Spångmyr Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

Eric Sprokkereef Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, The Netherlands

Elisabeth M. Stephens School of Archaeology, Geography and Environmental Science, University of Reading, Whiteknights, Reading, UK

Jutta Thielen-del Pozo European Commission, Joint Research Centre, Ispra, Italy

V. Thiemig European Commission, Joint Research Centre, Ispra, Italy

S. Tibaldi ARPA Emilia Romagna, Bologna, Italy

Zoltan Toth Global Systems Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration/OAR, Boulder, CO, USA

Carlos E. M. Tucci Institute of Hydraulic Research, Federal University of Rio Grande do Sul, Porto Alegre-RS, Brazil

Narendra Kumar Tuteja Bureau of Meteorology, Canberra, ACT, Australia

S. J. van Andel UNESCO-IHE Institute for Water Education, Delft, The Netherlands

Huug van den Dool CPC/NCEP/NWS/NOAA, College Park, MD, USA

R. E. Van der Zwan Principal Water Board of Rijnland, Leiden, The Netherlands

B. Vehviläinen Finnish Environment Institute, Helsinki, Finland

Jan S. Verkade Deltares, Delft, The Netherlands

Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, The Netherlands

Delft University of Technology, Delft, The Netherlands

A. Vogelbacher Bayerisches Landesamt für Umwelt, Augsburg, Germany

Jasper A. Vrugt Department of Civil and Environmental Engineering, University of California Irvine, Irvine, CA, USA

Department of Earth System Science, University of California Irvine, Irvine, CA, USA

Chen Wang South China Botanical Garden, Chinese Academy of Sciences, Richland, WA, USA

Guangsheng Wang Bureau of Hydrology, Ministry of Water Resources, Beijing, China

Lucien Wang Hazen and Sawyer, San Francisco, CA, USA

Q. J. Wang CSIRO Land and Water, Clayton, VIC, Australia

Xuguang Wang School of Meteorology, The University of Oklahoma, Norman, OK, USA

Yamei Wang Faculty of Geographical Science, Beijing Normal University, Beijing, China

Albrecht H. Weerts Operational Water Management, Inland Water Systems, Deltares, Delft, The Netherlands

Hydrology and Quantitative Water Management Group, Wageningen University and Research, Wageningen, The Netherlands

Mozheng Wei Oceanography Division, Navy Research Laboratory, Stennis Space Center, MS, USA

Kevin Werner National Weather Service, National Oceanic and Atmospheric Administration, Salt Lake City, UT, USA

Fredrik Wetterhall Forecast Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

Caroline Wittwer BRGM, Orléans, France

Andy Wood National Center for Atmospheric Research, Boulder, CO, USA

Eric F. Wood Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA

Chong-Yu Xu Department of Geosciences, University of Oslo, Oslo, Norway

Lihua Xiong Department of Hydrology and Water Resources, Wuhan University, Wuhan, China

Zongxue Xu College of Water Science, Beijing Normal University, Beijing, China

Yuhong Yan Bureau of Hydrology, Ministry of Water Resources, Beijing, China

Jianqing Yang Bureau of Hydrology, Ministry of Water Resources, Beijing, China

Tiantian Yang University of California, Irvine, CA, USA

Zhijie Yin Bureau of Hydrology, Ministry of Water Resources, Beijing, China

Huiling Yuan School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather, Ministry of Education, Nanjing University, Nanjing, China

Xing Yuan RCE-TEA, Inst. of Atmosph. Phys., Chinese Academy of Sciences, Beijing, China

Massimiliano Zappa Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

Senlin Zhou Bureau of Meteorology, Docklands, VIC, Australia

Part I

Introduction



Hydrological Predictability, Scales, and Uncertainty Issues

Joshua K. Roundy, Qingyun Duan, and John C. Schaake

Contents

1	Introduction	4
2	History and Application of Ensemble Hydrometeorological Forecasting	7
3	Understanding Prediction and Models	9
4	The Rationale for Hydrometeorological Forecasting	11
4.1	Hydrologic Predictability	13
4.2	Atmospheric Predictability	16
5	Ensemble Hydrometeorological Forecasting and Uncertainty	18
5.1	Input Uncertainty	18
5.2	Hydrologic Model Uncertainty	25
5.3	Verification	27
5.4	Forecast Products and Services	27
6	Summary	28
	References	29

Abstract

The survival and well-being of human civilization depends on water. Human civilization is especially vulnerable to large variations in the water cycle such as flood and drought that disrupts food supplies and can cause havoc to day-to-day operations. Many of these extreme events have occurred in recent years including

J. K. Roundy (✉)

Department of Civil, Environmental, and Architectural Engineering, University of Kansas,
Lawrence, KS, USA
e-mail: jkroundy@ku.edu

Q. Duan

Faculty of Geographical Science, Beijing Normal University, Beijing, China
e-mail: qyduan@bnu.edu.cn

J. C. Schaake

U.S. National Weather Service (retired), Annapolis, MD, USA
e-mail: jcschaake@comcast.net

large droughts and extreme floods in many parts of the world. The looming threat of climate change has the additional potential to make the impacts of extreme water cycle events an even greater threat to society. The ability to have foreknowledge of these extremes in the water cycle can provide time for preparations to reduce the negative impacts of these extremes on society. Predictions of these extreme events require models of the hydrometeorological system, including all its associated uncertainties, and appropriate observations systems to provide input data to these models. Ensemble forecasts using statistical and physically based models that also account for forecast uncertainties have great potential to make the needed predictions of future hydrometeorological events.

This chapter discusses the basis for predictability, predictive scales, and uncertainty associated with hydrometeorological prediction. Although much uncertainty may be associated with some hydrometeorological predictions, ensemble forecasting techniques offer a way to quantify this uncertainty making it possible to have more useful predictions for decision makers and for the ultimate benefit to society.

Keywords

Predictability · Uncertainty · GCM · ESP · Spatial scales · Temporal scales

1 Introduction

The dependence on water for the survival and well-being of human populations has made an inseparable link between human society and the water cycle. In particular, human society is especially vulnerable to large fluctuations in the amount of available water that come in the form of flood and drought and disrupts food supplies and can cause havoc to day-to-day operations. Many such extreme events have occurred in recent years including large droughts and extreme floods in many parts of the world (Marchi et al. 2010; Karl et al. 2012; Villarini et al. 2013; Smith et al. 2013). Furthermore, the looming threat of climate change presents a very real possibility that the fluctuations of available water could intensify (Huntington 2006; Sheffield and Wood 2008) and have a larger impact on society. Although extreme variability in available water is a clear and present threat to society, it is not a modern peril. Since the very beginning of civilization, human beings have had to deal with the elements of nature for survival, especially water, with dynasties literally thriving or collapsing in response to when, where, and how much water came, stayed, and went into the societal living environment.

Skillful foresight of water flow would help not only mitigate the disastrous effects of damaging floods and droughts, but also provide a tool for managing water as a valuable resource. This is particularly relevant for managing residential use, irrigation, power generation, navigation, and environmental and ecosystem protection. Hydrometeorological forecasts predict the amount of water at specific locations and times. It is often associated with prediction of natural hazards such as severe storms, floods, droughts, landslides, and coastal inundation by storm surges.

The term hydrometeorological is derived from a combination of the words hydrology and meteorology, both key components to the earth's water cycle. Hydrology pertains to the components of the water cycle that take place over land. This includes movement of water after it exits the atmosphere as precipitation and flows through the ground, lakes, rivers, and streams until it returns to the atmosphere through evaporation or flows into the ocean. Meteorology, on the other hand, deals with the transport of water within the atmosphere. This includes evaporation from land and sea and movement of evaporated water in the atmosphere until it ultimately falls as precipitation. It is simple to see that the transport of water through the hydrologic system begins where the meteorological system ends and vice versa. This coupling between the hydrology and meteorology is what makes up the water cycle of the earth. The ability to predict the dynamics of this system is essential for any form of water prediction. Therefore, the term hydrometeorological forecasting includes prediction of both the hydrological and metrological components of the water cycle.

Hydrometeorological forecasting is usually done with the aid of process-based or statistical hydrometeorological models. These can be further categorized into meteorological models and hydrological models. Meteorological models are designed to mimic water and energy cycles in the atmosphere and over land. Hydrological models are designed to emulate water and energy cycles that occur over and within the land surface. Hydrometeorological models are generally run in two modes: (i) simulation mode, where hydrometeorological processes of the past are emulated; and (ii) forecast mode, where future hydrometeorological processes are predicted. The commonly simulated/forecasted variables include precipitation, air temperature, river stage or streamflow discharge, snow aerial coverage, snow water content, evaporation, soil moisture, groundwater storage and discharge, river sediment, and other variables with practical application to water systems including upstream anthropogenic activities that affect downstream flows.

Depending on the length of the forecast time horizon, known as lead time, hydrometeorological forecasting may be classified into short-, medium-, and long-range forecasting. Short-range forecasts have lead times of a few hours to a few days into the future. Medium-range forecasts have lead times of a few weeks. Long-range forecasts have lead times ranging from a few weeks to 1 year or more. From a meteorological perspective, short-range predictions are associated with weather and long range is associated with climate. Short- and medium-range forecasts predict how actual hydrometeorological states change in time and space, while the skill in long-range forecasts is in estimates of average hydrometeorological states expected over monthly or seasonal time periods over large areas.

Historically, hydrometeorological forecasts, especially short-range forecasts, have been made as deterministic, single-value estimates of the magnitude, location, and timing of likely hydrometeorological events. Two examples of deterministic hydrometeorological forecasts are: "the temperature will be 35 °C in Hong Kong tomorrow," and "the streamflow discharge of Mississippi River at Vicksburg, Mississippi, will be 516,000cfs on Sunday." Deterministic forecasts have been used in operational hydrometeorological forecasting centers around the world

since before computer-based hydrometeorological models came to existence many decades ago. Even today, they are still widely used for short-range forecasting throughout the world. Deterministic forecasts can often satisfy many public needs.

Recently, forecast users are becoming increasingly aware of the limitations of deterministic, single-value hydrometeorological forecasts, especially under severe or extreme hydrometeorological conditions. As pointed out by Ramon Krzysztofowicz (2001), “A deterministic forecast may create the illusion of certainty in a user’s mind, which can easily lead the user to suboptimal action.”

One example involving forecaster lack of confidence in a single-value prediction that led to issuance of an incorrect warning that resulted in devastating consequences occurred during a torrential storm in Beijing, China. On July 21, 2012, the numerical weather prediction (NWP) model run by Beijing Meteorological Bureau predicted a major storm would occur over Beijing that day. The Beijing Meteorological Bureau only issued a “blue code” warning for Beijing before the storm that lasted from 9:00 am until 3:00 am in the next morning. This warning implied that over 50 mm of rainfall would occur over the city in the ensuing 12-h period. But the actual rainfall amount was over 170 mm for all areas of Beijing City, with downtown receiving 220 mm and over 460 mm in the Southwestern Fanshan District over the storm period. Because of this incorrect, underpredicted warning, inadequate emergency measures were taken by the city government, leading to a virtual shutdown of city traffic soon after the storm began, due to almost all highway underpasses flooded and all public transportation interrupted. A professional football game and a large music concert with tens of thousands of people in attendance went on as scheduled in the evening, compounding the traffic paralysis. The final casualty tally included 79 deaths and property damages in excess of ¥10B in Chinese Yuans.

The irony is that the forecasted rainfall from the numerical weather prediction (NWP) model run by Beijing Meteorological Bureau was closer to the actual rainfall amount than that indicated by the warning level. This severely underestimated warning was issued because the forecaster lacked confidence in the deterministic forecast of this extreme storm that later was determined to be a 1-in-60-year event. This is just one example of the failure of a deterministic hydrometeorological forecast. Reality is that they often fail, especially during extreme events, because a single model prediction lacks information about the certainty or uncertainty of what will actually occur.

In contrast to single-value deterministic forecasts, ensemble forecasts are made using Monte Carlo simulation in which a single model or multiple models are run numerous times to generate multiple samples of future states of the dynamical hydrometeorological system. Different model outcomes are generated by perturbing uncertain factors such as model forcing, initial and boundary conditions, and/or model physics. Ensemble techniques are attractive because they not only offer an estimate of the range of possible future states of the hydrometeorological system, but also offer a way to quantify the risks of a catastrophic hydrometeorological event occurring.

The goal of this chapter is to provide an overview of hydrometeorological ensemble forecasting in terms of understanding the history and rationale of ensemble

hydrometeorological forecasting and uncertainty estimation. This chapter discusses major aspects of predicting hydrometeorological systems and provides a broad overview of the topics covered in this book. This chapter is a mile wide and only an inch deep. But references to other parts of this book will be given to facilitate more in-depth understanding of the concepts presented in this section.

This chapter has five remaining sections. Section 2 provides a brief history and application of hydrometeorological forecasts and is followed by an introduction to the nature of prediction in its many forms and subtleties in Sect. 3. Aspects of prediction discussed in Sect. 3 are then applied to hydrometeorological predictability through discussion of the rationale for hydrometeorological prediction in Sect. 4. Section 5 discusses the components of an ensemble hydrometeorological forecast system emphasizing the uncertainties and skill in hydrometeorological prediction. This is then followed by a brief summary in Sect. 6.

2 History and Application of Ensemble Hydrometeorological Forecasting

Ensemble approaches to scientific understanding can be traced to the emergence of Monte Carlo simulation as a practical way to solve complicated computational problems. The term “Monte Carlo simulation” was coined by scientists working on the Manhattan Project in Los Alamos, New Mexico, in the 1940s, who used random numerical techniques to obtain solutions for highly complex fluid dynamical computational problems. Monte Carlo simulation is, in a way, similar to many games popularly played in the gambling city Monte Carlo in Europe. Today, Monte Carlo simulation usually refers to the use of random number generation together with numerical models to obtain approximate solutions to complicated computational problems.

Edward Lorenz’s discovery of the chaotic nature of the atmospheric system in 1965 provided theoretical justification that ensemble hydrometeorological forecasting is the only practical way to deal with the uncertainty in the atmospheric initial conditions and the inherent nonlinearity of the atmospheric system. Edward Epstein recognized in 1969 that the atmosphere could not be completely described with a single forecast run due to inherent uncertainty, and proposed a stochastic dynamic model that produced means and variances for the state of the atmosphere (Epstein 1969). Although these Monte Carlo simulations showed skill, in 1974 Cecil Leith revealed that they produced adequate forecasts only when the ensemble probability distribution was a representative sample of the probability distribution in the atmosphere (Leith 1974).

Epstein presented a theoretical Stochastic-Dynamic approach, which adds stochastic terms to dynamical equations to account for various uncertainties (Epstein 1969). However, his approach was not practical in operational applications as the number of model equations is too many to be implemented numerically given the limited computation resources at the time. Later, Leith proposed a Monte Carlo approach as a practical implementation of Epstein’s Stochastic-Dynamic approach,

which represented the first attempt at ensemble forecasting (Leith 1972). In this approach, a hydrometeorological model was run in an ensemble mode. In the 1970s, only the uncertainty in initial condition was represented with only a few ensemble members due to the computational constraint. As computational power has grown exponentially over the years, uncertainty in initial condition, model dynamics, model structure, and even model parameters are often considered simultaneously in ensemble forecasting today. For example, ensemble Kalman filter (ETKF) has been used to provide probabilistic estimate of both the state variables as well as model parameters in many meteorological applications (Cornick et al. 2009; Evensen 2009).

It was not until 1992 that ensemble forecasts began being prepared by the European Centre for Medium-Range Weather Forecasts and the National Centers for Environmental Prediction (Molteni et al. 1996; Toth et al. 1997). The ECMWF Ensemble Prediction System uses singular vectors to simulate the initial probability density, while the NCEP Global Ensemble Forecasting System uses a technique known as vector breeding. Now ensemble forecasting is an essential and routine component of the suite of meteorological forecast products for many countries.

Realizing the importance of ensemble forecasting, the World Meteorological Organization (WMO) sponsored a 10-year international program in 2003 – The Observing System Research and Predictability Experiment (THORPEX) to accelerate improvement in ensemble meteorological forecasting for lead times ranging from 1 day to 2 weeks. As part of the THORPEX, ten numerical weather prediction centers around the world have provided their medium-range ensemble meteorological forecasts to THORPEX Interactive Grand Global Ensemble (TIGGE) database since the end of 2006.

From a hydrological perspective, ensemble forecasting strategy has been used since the 1970s. The U.S. National Weather Service (NWS) produced Extended Streamflow Prediction (ESP) by using historical temperature precipitation data from different years to represent future precipitation and temperature forcing (Day 1985). Climatological ESP has been used extensively in reservoir and water supply operations applications. US nation-wide ensemble hydrological forecasting began in the late 1990s with the advent of the Advanced Hydrologic Prediction Service (AHPS) that uses climatological ESP. At approximately the same time, ECMWF began to produce hydrological ensemble forecasts by making use of meteorological ensemble forecasts from NWP models.

In the 2000s, the NWS California Nevada River Forecast Center (CNRFC) began issuing short-range hydrologic ensemble forecasts that used preprocessed single-value precipitation and temperature forecasts from atmospheric models and from operational hydrometeorological forecasters. In 2010, NWS began implementing a national Hydrological Ensemble Forecast System (HEFS) that included atmospheric forecast preprocessing techniques to provide reliable forcing, at both short range and long range, for hydrologic forecast models at the space and time scales required by these models.

A major driver for global adoption of ensemble hydrological forecasting is the founding of the Hydrological Ensemble Prediction Experiment (HEPEX) in 2004.

The main objective of HEPEx is to bring the international hydrological community together with the meteorological community to demonstrate how to produce reliable “engineering quality” hydrological ensemble forecasts that can be used with confidence to assist the emergency management and water resources sectors to make decisions that have important consequences for the economy and for public health and safety.

3 Understanding Prediction and Models

To fully understand ensemble hydrometeorological prediction, it is first necessary to grasp the basics of predictions and models. The Webster online dictionary defines prediction as “a statement about what will happen or might happen in the future” (Merriam-Webster). Predictions can take many forms and be founded on any reasoning. But predictions are only useful if they are made from a sound understanding that provides a strong base for the prediction. Often, the basis of the prediction is called the premise of predictability. The premise is the rationale for the prediction and ultimately limits the potential success of the prediction. Regardless of the premise, prediction usually takes one of two forms: deterministic or probabilistic.

Deterministic prediction gives a single outcome of the thing being predicted. In many societal applications, a deterministic prediction is wanted in order to make a decision that is related to the event being predicted. If the prediction were perfect, then a deterministic statement would provide all the information needed to make an informed decision. But there are very few perfect predictions. Therefore, it becomes important to account for the uncertainty in the prediction. A probabilistic prediction does not give a single outcome, but provides a range of possible outcomes and an estimate of the likelihood of their occurrence. Although probabilistic statements do not provide single outcomes, they enable informed decisions based on the additional information about the predictive uncertainty.

The importance of predictability and the difference between deterministic and probabilistic predictions can be illustrated through the simple example of flipping a coin. Consider, for example, a coin toss where a competitor predicts before the toss which side of the coin (heads or tails) will come out upright. If that happens, they win; otherwise, they lose. In such situations, it is in the person’s best interest to be able to correctly predict the outcome from flipping a coin to ensure they receive the maximum benefit or avoid the penalty.

In this example, the competitor’s choice of heads or tails is an example of a deterministic, single-value prediction. Alternatively, the statement that there is a 50% chance of heads and a 50% chance of tails would be an example of a probabilistic prediction for the same event. The probabilistic prediction does not give a direct outcome, but gives an assessment of all outcomes. The competitor must then make their own deterministic prediction, but the competitor’s choice can be informed by the probabilistic prediction.

The basis or premise of predictability of the probabilistic prediction is key for the ultimate utility of the prediction. For example, the 50-50 chance of the coin flip outcome could be dictated by the event space and the assumption that each side of the coin is equally likely. Although this assumption of a fair sided coin is reasonable for this simple example, a prediction could be refined by observations and incorporating these observations into a simple statistical model. An example of developing such a model would be to flip the coin multiple times and recording the outcome, with the outcomes being used to calculate the relative frequency of each outcome. For example, if after four coin flips there was one head and three tails, the probabilistic prediction of all future coin flips would be 25% heads and 75% tails. For this prediction, the premise of predictability would be based on four observed experiments. Although this is an assessment of all possible outcomes, it would not be considered statistically robust given the few samples used to derive the prediction. On the other hand, if the same prediction of 25% heads and 75% tails were based on 1000 samples, this would be considered a statistically robust prediction (but unlikely one) and would be a clear indicator of a biased coin. Such foreknowledge about the biased coin would provide considerable advantage to the competitor in this example.

Statistical models based on observations are not the only way to estimate the likelihoods of the eventual outcomes of an event. Physically based models based on the underlying physics are also very useful. An example of a physically based model for the coin flip outcomes would be a mathematical model that numerically simulates the rotation of the coin and its interaction with the surrounding environment throughout the trajectory of the coin: based on properties of the coin itself, like its size and weight as well as the initial height of the coin and the magnitude and location of the force exerted on it. The premise of predictability for such a model would be that there is enough information about the properties of the coin and physical representation of its environment that could be represented mathematically to understand the outcome of flipping a coin. Using this model for a single prediction would be an example of a deterministic prediction. But the only way this model would consistently give the correct outcome of flipping a coin would be if the physics and inputs in the model were exact. Because the mathematical formulas and measurements for a coin toss simulation would be imperfect, there is little chance a single run of the model would provide more utility than a simple guess. In this case, the inability of the model to provide useful information about the coin flip would come from the uncertainty in model input and model physics that would likely overwhelm any predictive power of knowing the properties, initial height, or force exerted on the coin.

Additional model runs provide a more robust assessment of the possible outcomes than a single realization. Different model inputs based on uncertainties could be used to generate multiple model outcomes that could be combined to give a probabilistic description of the event. Multiple runs of a model form an ensemble and are necessary when making predictions using models that have underlining uncertainties.

4 The Rationale for Hydrometeorological Forecasting

Predicting the hydrometeorological system is similar to predicting the result of flipping a coin in that statistically or physically based numerical models can be used to make deterministic or probabilistic forecasts, but unlike flipping a coin, which only has two outcomes, the hydrometeorological system has many outcomes. Mathematical and statistical models used to predict the hydrometeorological system are extremely complex. Although they incorporate numerical representations of the physics that drive the water cycle, there are still many uncertainties in the inputs, parameters, and models themselves. Therefore, it is essential that these models use ensemble techniques to provide the maximum amount of information about the hydrometeorological system. But even with ensemble techniques, any prediction is only as good as the basis for prediction. This section discusses the premise of predictability in hydrometeorological models and outlines the rationale for ensemble hydrometeorological forecasting.

The premise of predictability for the hydrometeorological system is twofold. First is the predictability of the atmosphere that drives the hydrologic cycle. Second is the memory in the initial hydrologic state of the land. The premise of predictability also depends on both temporal and spatial scales of prediction. The temporal scale of hydrometeorological prediction can be separated into short-range (weather) and long-range (climate) predictions. This premise of predictability and the roles of space and time scales for both weather and climate predictions are illustrated in Fig. 1.

In this figure, short-range predictions are for the period ranging from zero out to about 15 days. Hydrometeorological predictability in the short range is impacted by the initial hydrologic conditions, i.e., the current state of the hydrologic system.

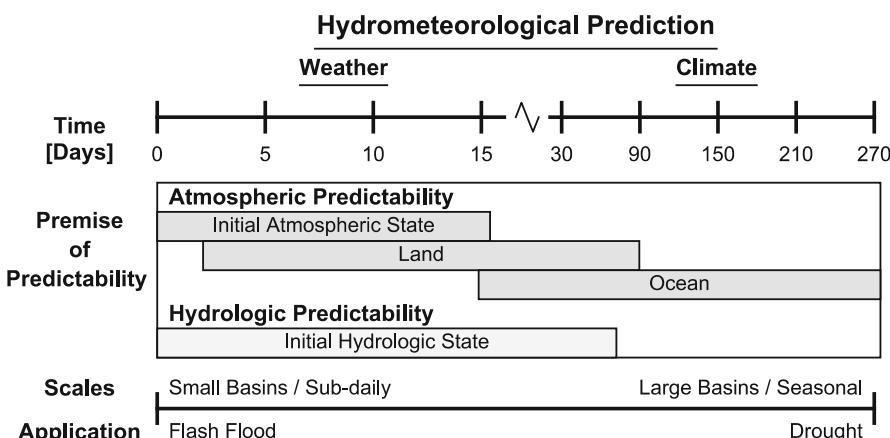


Fig. 1 Summary characteristics separated into two main categories (weather and climate) for hydrometeorological prediction

However, short-range hydrometeorological predictability also greatly depends on the predictability of the atmosphere. Short-range predictability of the atmosphere is primarily driven by atmospheric initial conditions whose effect is generally considered to be limited to the first 15 days (Lorenz 1969, 1982). Predictability of the atmosphere is strongest in the first few days when predictions can be very precise and are often applied to small basins and at subdaily time-steps (de Roo et al. 2003). Because of the high level of predictability at short lead times, short-range predictions are particularly applicable to flood events that can develop quickly at specific locations. Short-range predictions of low flows can also be useful for other purposes such as controlling reservoir operations to ensure ecological habitat.

Climate hydrometeorological predictions are for time scales ranging from about 2 weeks to many months in the future or longer. At these time scales, the main basis for atmospheric predictability is the influence of boundary conditions (land and sea) on the atmosphere. At climate scales, slowly varying atmospheric boundary conditions influence the transport of water in the coupled ocean-atmosphere-land system. Land and sea surface boundary conditions both contribute to predictability at the climate scale, although the sea surface plays a large role at longer lead times (Palmer and Anderson 1994; Goddard et al. 2001).

Hydrologic initial conditions also play a critical role in hydrometeorological predictions at climate scales, but their role diminishes rapidly with forecast lead time (Li et al. 2009). One major exception is the influence of snow. Snow water initial conditions can be a basis for long-range hydrologic predictability in spring and summer for areas with substantial snow accumulation.

The precision (i.e., level of detail) of climate scale predictions is usually much coarser than short-range predictions. This is because atmospheric predictions are influenced by atmospheric noise and require temporal and spatial averaging in order to detect the predictive signal. Therefore, hydrometeorological predictions at climate time scales are most applicable for large basins on seasonal or longer time-scales. The coarser level of detail of climate scale predictions make them most applicable to drought that typically forms over long time periods and large areas. Although droughts are the main use of climate scale hydrometeorological predictions, the ability to predict anomalously wet time periods at climate scales is also beneficial.

Hydrometeorological predictions for both short- and long-time scales are typically made as separate atmospheric and hydrological predictions. Because forecast uncertainty is important, these predictions, both atmospheric and hydrologic, are often made as ensemble predictions with ensemble member time series having daily, or possibly, subdaily time steps.

The remainder of this section discusses some examples of the premise of hydrometeorological prediction. First, the premise of predictability of the hydrologic system due to initial conditions is discussed and is followed by a discussion of the predictability of the atmosphere and the combined impact of hydrologic and atmospheric predictability. Examples of hydrometeorological prediction given in this section are for the climate scale. However, similar methods and applications could

easily be demonstrated for short-range forecasts. For a showcase of other applications of ensemble hydrometeorological predictions, see Part 10 of this book.

4.1 Hydrologic Predictability

Hydrologic predictability depends on initial hydrologic conditions such as soil moisture, groundwater, snow and current streamflow, as well as on atmospheric predictability because hydrologic processes depend on precipitation and temperature forcing as well as initial conditions.

To illustrate how initial hydrologic conditions affect predictability, consider the example of a large rain event over a basin. If the previous month before the rain event were abnormally dry, then soils would be dry and baseflow would be low. Under these conditions, a heavy rain event would result in high infiltration rates and a smaller amount of direct runoff. The small amount of direct runoff would combine with low but increasing baseflow to produce the total discharge that would not likely cause extreme flooding. In contrast, if conditions prior to the heavy rainfall event were abnormally wet, then the basin would have wet soils at or near saturation and a larger initial baseflow. In this scenario, the same storm event would produce low infiltration and a large amount of direct runoff. Combining the large amount of direct runoff with the initially high baseflow could result in a very large discharge that might induce extreme flooding. This heavy rain event could lead to even more extreme conditions if snow were present. This could produce even more streamflow and would be called a “rain on snow event.”

This example illustrates why knowing initial hydrologic conditions is a crucial part of hydrometeorological prediction. This requires measuring or estimating soil moisture, snowpack, and streamflow throughout the basin. These variables are routinely measured in some locations. But it is not feasible to make such measurements everywhere hydrologic predictions are needed. For this reason, hydrologic models have been developed to provide temporally and spatially continuous estimates of initial hydrologic conditions as well as hydrologic forecasts.

Hydrologic models can use a variety of methods to transform atmospheric forcing variables (precipitation, temperature, humidity, radiation) into estimates of soil moisture, snowpack, runoff, and streamflow. When observations or estimates of atmospheric forcing are used to drive hydrologic models, this is often referred to as offline modeling, because the atmospheric model is not coupled to the hydrologic model. Traditionally, hydrologic forecast models have been used in an off-line mode for hydrologic forecasting.

Hydrologic forecast models can be applied in several different modes. One is the forecast mode where a hydrologic forecast model is used to make hydrologic forecasts for some lead time. Another is a simulation mode where a hydrologic model is run, possibly for many years, using historical observations. Short-term simulations run in parallel with forecast operations and provide initial hydrologic conditions for

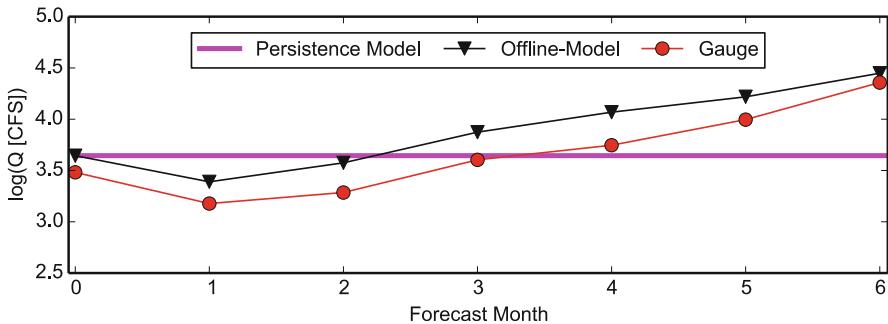


Fig. 2 Example of a seasonal hydrologic streamflow prediction using the simple persistence model along with the offline model and gauge observations

forecast runs. Long-term simulations give information about how well a hydrologic model can predict streamflow when input forcing is based on observations. These simulations can be used to make estimates of hydrologic model error. Finally, hydrologic models can be used in a reforecast mode to generate hydrologic reforecasts from atmospheric reforecasts. Such reforecasts can be used to develop and test water management and flood control models. For a full treatment of hydrologic models and their use in forecasting, see Part 4 of this book.

The simplest prediction model that uses only initial hydrologic conditions is the persistence model. The persistence model simply predicts future streamflow to be the same as the current streamflow. Although this is a very simple model, the persistence model represents the minimum benchmark for more sophisticated hydrological models. An example of hydrologic forecasts from a persistence model is given in Fig. 2 together with an example of a single-value forecast from a hydrologic forecast model and corresponding gauge observations. In this case, the persistence model provides reasonable predictions for the first few months but lacks variability and cannot predict extreme events. There is little practical utility in persistence model forecasts: they are purely deterministic and provide no information about uncertainty. An important limitation of the persistence model is that it does not fully utilize the information provided by the initial hydrologic conditions. Another is that it does not use information about future forcing.

Most regions have large seasonal variability in the hydrologic cycle that can further inform hydrologic forecasts. A forecast technique that uses atmospheric climatology together with estimates of initial hydrologic conditions is known as Extended Streamflow Prediction or ESP (Day 1985). This technique has been used by the U.S. National Weather Service since 1974 (shortly after the first continuous hydrologic simulation models became available and began to be used operationally by NWS).

ESP was first implemented in the western US to make seasonal snowmelt runoff forecasts as an alternative to existing regression-based statistical forecasts of seasonal snowmelt runoff volumes. It is the earliest known example of ensemble prediction being used for operational streamflow forecasting. The original,

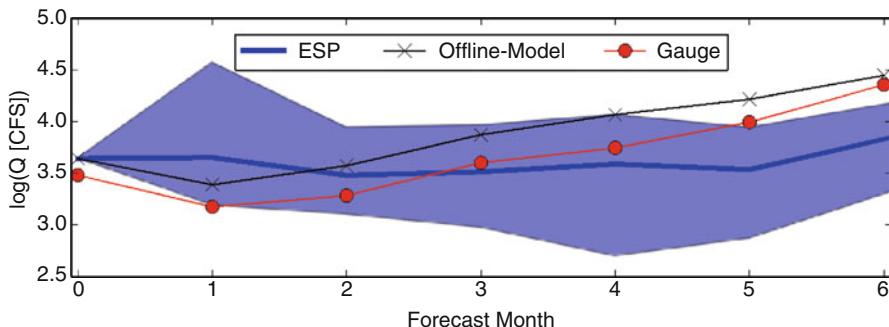


Fig. 3 Example of a seasonal hydrologic streamflow prediction using the ESP method with the ensemble mean (dark line) and the uncertainty bounds from all ensemble members along with the offline model and gauge observations

climatological ESP uses initial hydrologic conditions estimated by running the forecast model forward each day using the previous day's initial conditions together with recent observations of atmospheric forcing to generate continuing day-to-day updates of initial hydrologic conditions. It uses future atmospheric forcing selected from the climatological record. Different ensemble members are generated from precipitation and temperature time series taken from different past years of the climatological record beginning on the current forecast day of the year.

As different approaches to hydrologic ensemble prediction developed, the original meaning of the acronym ESP has evolved to mean Ensemble Streamflow Prediction. These new forms of ESP have also been called conditional ESP (Hamlet and Lettenmaier 1999). Nevertheless, the key element of ESP is use of ensemble of atmospheric forcing (either from climatology or from atmospheric forecast models) to drive hydrologic forecast models to produce forecasts of future hydrologic variables. An example of an original, climatological ESP forecast for the same event seen in Fig. 2 is given in Fig. 3. The shaded blue area in Fig. 3 is the spread of the individual ensemble members and gives an indication of uncertainty in the persistence of the initial conditions. The range of the ESP ensemble brackets both the deterministic hydrologic forecast (from Fig. 2, also shown in Fig. 3) as well as the gage observations through month 4 but does not bracket them for months 5 and 6. The inability to bracket the gauge observations is due to the initial discrepancy between the ESP hydrologic forecast model and the corresponding gage observation.

ESP has more utility than either a simple persistence model or a single-value hydrologic forecast model. If a single-value forecast were wanted, that could be derived as the ensemble mean. The dark blue line in Fig. 3 is the ensemble mean for the ESP forecast. The ensemble mean includes variability that is associated with the seasonal variability of the atmospheric forcing. Climatological ESP forecasts tend to the climatology of the streamflow as the influence of the initial condition dissipates. This limits the predictability of the climatological ESP method at longer forecast lead times. Despite this limitation, the climatological ESP method is still widely used in operational hydrologic forecasting.

4.2 Atmospheric Predictability

Atmospheric predictability depends on two factors: the internal dynamics of the atmosphere and the variability of atmospheric boundary conditions (with land and sea surfaces). It also may depend on knowledge of past climatology. A lower limit of atmospheric predictability is a prediction that the future will be like the past climatology.

How internal atmospheric dynamics limits atmospheric predictability was studied by Edward Lorenz using a toy model of the atmospheric system. Lorenz made a seminal discovery that equations describing the atmospheric system exhibit chaotic behavior, in which a minute perturbation in the initial condition would result in dramatically different future states (Lorenz 1965). This effect, known as the “Lorenz attractor” or “butterfly effect,” implies that deterministic forecasts of atmospheric events beyond a few days are not possible because the prescribed initial condition used to initialize a forecast cannot be certain, regardless how accurate the observations may be.

This does not mean that atmospheric forecasts beyond a few days are totally impossible. By accounting for uncertainty in initial conditions, it is possible to provide a probabilistic estimate of future atmospheric states for lead times up to 2 weeks (Froude et al. 2013). An ensemble approach is well suited to represent uncertainty in initial conditions. This can be achieved by running models in an ensemble mode with different initial conditions sampled from a probability distribution function (PDF) of initial conditions. Because of the predictability limit of the atmospheric system due to chaos (Straus and Shukla 2005), deterministic details of hydrometeorological events cannot be forecast beyond 2 weeks. Accordingly, long-range hydrometeorological forecasts are generally produced as probabilistic forecasts of events or as ensemble time-series forecasts (Hoskins 2013).

Predictability of the atmosphere beyond 2 weeks comes from models called coupled General Circulation Models (GCMs) that include the dynamics of ocean and land boundary conditions as well as the physics of the atmosphere and uncertainty in initial conditions. Accordingly, the premise for predictability of the atmosphere has two parts, weather time-scales and climate time-scales. As shown in Fig. 1, weather scale predictions of the atmosphere is generally considered to be the first 15 days and the predictability primarily lies in the initial conditions of coupled ocean-atmosphere-land system. As the influence of initial atmospheric conditions deteriorates, boundary conditions become increasingly important for maintaining predictability of the atmosphere. The influence of the land boundary conditions on the predictability of the atmosphere can be important at all time-scales but is strongest for longer weather scales and shorter climate scales. For very long climate scales, the predictability of the atmosphere is primarily driven by sea surface boundary conditions and by ocean dynamics that affect the sea surface. Atmospheric predictions increasingly rely on ensemble techniques using multiple realizations generated by numerical models.

As atmospheric models were improved and began to be run in ensemble mode, statistical techniques were developed to use atmospheric precipitation and temperature forecasts to specify future precipitation and temperature forcing for hydrologic

forecast models. Atmospheric model forecasts were not used directly as input to hydrologic models because: (i) atmospheric forecasts are for much larger spatial scales than hydrologic models; and (ii) the climatology of atmospheric forecasts is different than the climatology of the atmospheric forcing needed by the hydrologic forecast models. Despite these differences, atmospheric forecasts still have useful information about future hydrologic forcing. Therefore, statistical techniques were developed to account for uncertainty in the relationship between the atmospheric forecasts and the temperature and precipitation events that actually occurred over the local hydrologic forecast areas.

The first operational statistical processing techniques to generate ensemble forcing for hydrologic ensemble forecast models from the atmospheric forecasts were developed by the U.S. National Weather Service (Schaake et al. 2007). This required a record of past atmospheric reforecasts using the current atmospheric forecast model (Hamill et al. 2007). The first atmospheric reforecasts became available in the early 2000s from an ensemble forecast version of the NWS Global Forecast System (GFS). Only GFS ensemble mean forecasts were used in the statistical processing techniques (MEFP) that are part of the National Weather Service Hydrologic Ensemble Forecast System (HEFS).

An example of a hydrometeorological forecast that utilizes ensemble predictions of the atmosphere from a GCM to drive a hydrological model is given in Fig. 4 for the same period as Figs. 2 and 3. Like the ESP method, the GCM-based forecast provides estimates of uncertainty of the evolution of the initial hydrologic state; however, the GCM method uses forecasts of atmospheric variables instead of a random selection from the climatology. The forecasts of the atmospheric variables have the potential to provide a better prediction than the ESP method. This can be seen in Fig. 4 for the first 2 months, which show a smaller range of uncertainty that still brackets the hydrologic simulation model. This indicates a higher degree of confidence in the model. This is particularly seen in the 1st month, where the GCM prediction has a much smaller uncertainty and the ensemble mean closely matches the offline model. The increased skill in the 1st month is due to the influence of the

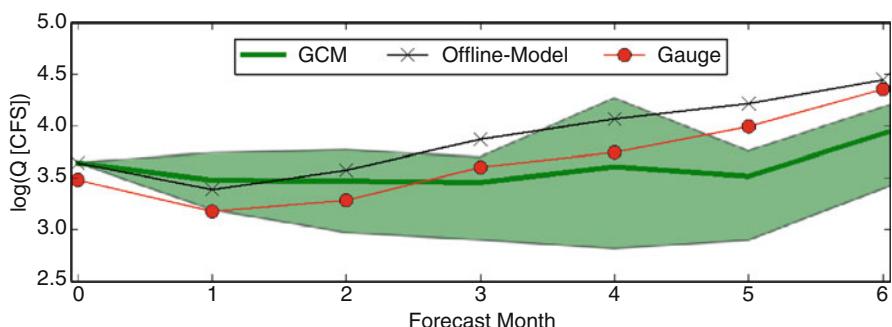


Fig. 4 Example of a seasonal hydrologic streamflow prediction using postprocessed atmospheric forcing from a GCM with the ensemble mean (dark line) and the uncertainty bounds from all ensemble members along with the offline model and gauge observations

atmospheric initial conditions on the weather time scale. As the forecast time increases, the predictability of the GCM-based forecast diminishes. In particular, month 3 for the GCM-based forecast has a smaller uncertainty estimate than the ESP prediction; however, the uncertainty fails to bracket the offline model. This indicates that the GCM prediction was overconfident in its prediction of the 3rd month and shows that a GCM-based forecast can provide better predictions than the ESP in some situations, but there is still a large amount of uncertainty in the GCM predictions.

5 Ensemble Hydrometeorological Forecasting and Uncertainty

Ensemble hydrometeorological forecasting systems are complex and consist of many components that are imbedded with various levels of uncertainty. Different systems will account for different levels of uncertainties, which have direct impacts on the usefulness of the predictions. For example, there is a large amount of uncertainty in the persistence model prediction (Fig. 2). But it is not quantified, which limits the usefulness of the prediction. One of the major benefits of using ensemble predictions is that it provides a means to quantify uncertainty. The climatological ESP method (Fig. 3) quantifies the uncertainty of the evolution of initial hydrologic conditions by accounting for some aspects of the uncertainty of the atmospheric forcing. Similarly, the GCM-based hydrometeorological ensemble prediction (Fig. 4) also accounts for uncertainty in future atmospheric forcing of the hydrologic forecast model and provides a prediction with smaller uncertainty bounds than the climatological ESP for the first few months. Since there likely will always be uncertainty in hydrometeorological predictions, it is necessary to make sure all aspects of uncertainty are accounted for. If all uncertainty is not accounted for, then there is a risk of the forecast being over confident. This appears to have happened for the climatological ESP and GCM predictions in Figs. 3 and 4 for the forecasts of months 5 and 6. This illustrates that although both the ESP and GCM-based forecasts account for some aspects of uncertainty in the prediction, other aspects of uncertainty are not completely represented.

Accounting for all uncertainties in hydrometeorological forecasting is extremely challenging and complex. To simplify the conceptual understanding of an ensemble hydrometeorological system, Fig. 5 provides a schematic of the main components of a hydrometeorological forecasting system, which includes four main components: Input Uncertainty, Hydrologic Uncertainty, Ensemble Verification, and Products and Services. These four components will be discussed throughout the remainder of this section, with a particular emphasis on input and hydrologic uncertainty.

5.1 Input Uncertainty

There is much uncertainty in the inputs (atmospheric forcing) that drive hydrologic models that needs to be accounted for in a hydrometeorological forecast system.

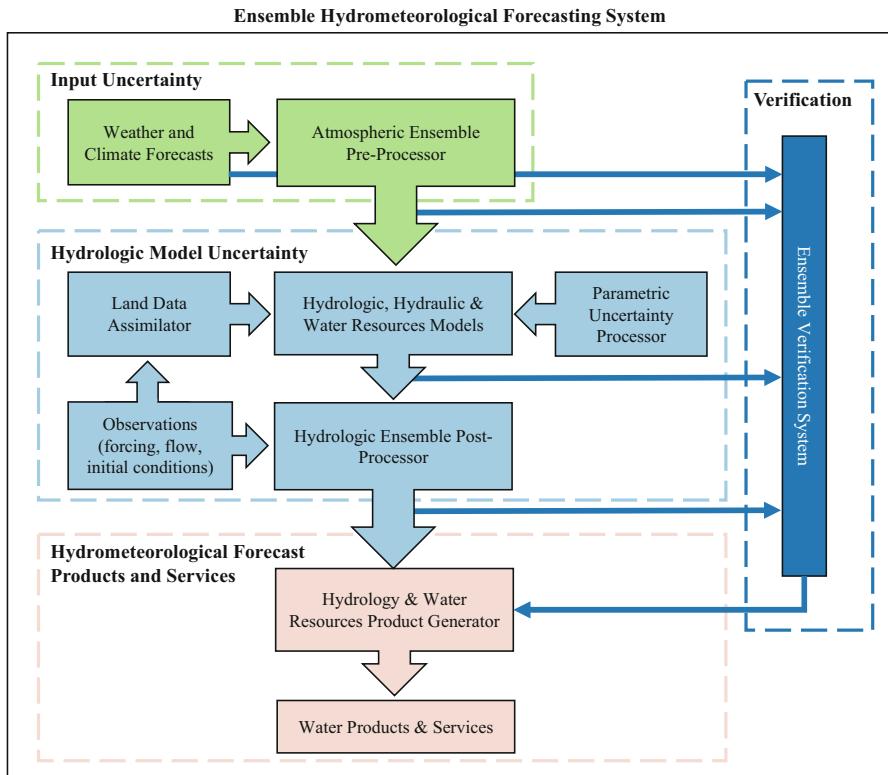


Fig. 5 Schematic of an Ensemble Hydrometeorological Forecasting System broken up into four main components, input uncertainty, hydrologic model uncertainty, verification, and forecast products and services

There are several ways to account for this uncertainty. One is the climatological ESP method. Another is to use preprocessed ensemble forecasts from a GCM. Nevertheless, there is much uncertainty associated with the GCM prediction related to the scale of the prediction, as illustrated in Fig. 6. This is one of the reasons preprocessing the meteorological predictions are so important (see Part 3 of this book). By no means, the methods discussed in this chapter fully account for all of the uncertainty associated with atmospheric forcing. However, they illustrate the importance of considering this uncertainty in hydrologic prediction. This section focuses on predictive scales for hydrometeorological forecasts by considering the change in skill with temporal and spatial scales in precipitation forecasts from GCM. One measure of skill for GCM precipitation forecasts is the correlation coefficient between forecasts and corresponding observations (Zar 2005). Two definitions of the correlation coefficient can be used. One is the Spearman rank correlation coefficient that considers only the rank of the values being compared. The other is the Pearson correlation coefficient that uses the actual values of the variables being compared. Both measures are used in the sections below.

5.1.1 Temporal Skill in GCM Short-Range Precipitation Forecasts

Forecast skill in short range GFS precipitation forecasts is temporally scale dependent. This is illustrated in Fig. 6 for winter season precipitation events in the North Fork of the American River in California. Figure 6 is an example of how forecast skill for events of increasing duration can exceed forecast skill of shorter duration events.

The measure of forecast skill used in Fig. 6 is the Pearson coefficient of correlation between GFS ensemble mean forecasts and observed values, including all zero values, for all forecasts made during a 30-day window on either side of January 15 between 1979 and 2002. The lowest curve in this figure is the correlation between 6 h forecasts and observed values as a function of the valid time (in future 6 h periods) of the 6 h forecast. The correlation values tend to decrease with lead time and a small daily diurnal cycle persists throughout the 14-day total forecast period. The highest curve in this figure is the correlation between the average forecast and observed precipitation values where the duration of the average begins at forecast creation time and ends at the indicated validation time. This curve shows that the GFS-based skill in predicting the total precipitation to be expected for the next 14 days is the same as the skill in predicting the amount of precipitation that will occur in the first 6 h. Clearly most of the information about future precipitation in this example is in the average or accumulated amount of precipitation over various future time periods. GFS forecast skill behaves this way in this location because most of the winter precipitation along the west coast is caused by large-scale storms that may last for several days. It is very difficult to predict the magnitude, timing, and exact location of every burst of precipitation during these events. A major source of error is temporal phase error: precipitation may occur before or after the time it was predicted. But averages over time, smooth the effect of these phase errors.

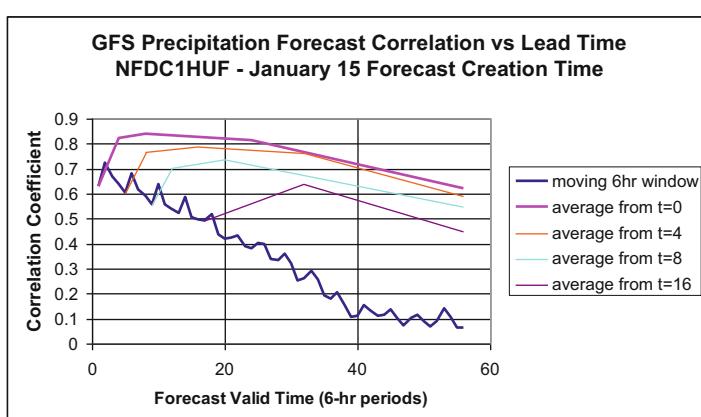
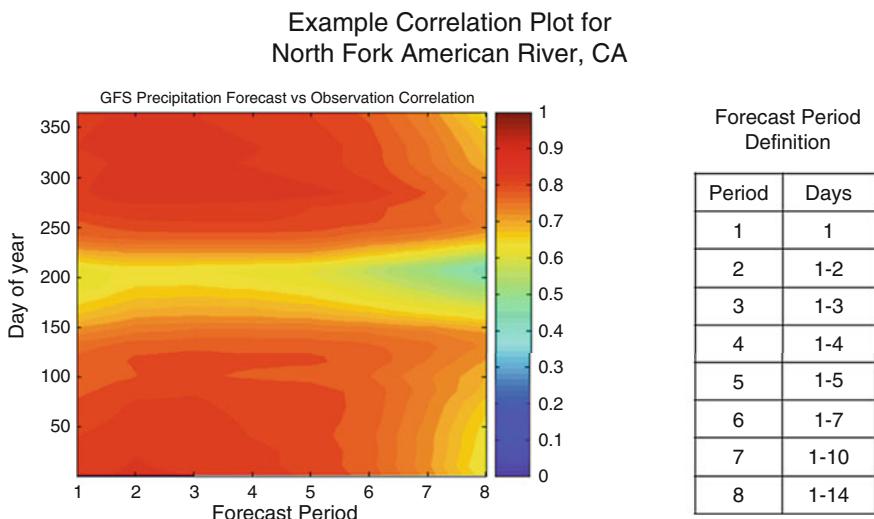


Fig. 6 Example of temporal scale-dependent uncertainty in precipitation forecasts

5.1.2 Analysis of Short-Range GFS Precipitation Forecast Skill in the US

This section provides a general overview of the skill of GFS-ensemble mean precipitation forecasts for different forecast lead times at different times of the year and at different locations throughout the US. The measure of forecast skill selected for this study is the Pearson coefficient of correlation between observations and GFS ensemble mean forecasts. Data presented here are for 24 basins distributed across the US. The forecast variable chosen for this study is the cumulative precipitation amount for different numbers of days in the future. This variable was chosen because the forecast skill in future cumulative precipitation amounts often tends to increase with time from the forecast creation date before it tends to decrease.

An array of correlation coefficients between observed and GFS raw ensemble mean forecasts of this variable were computed for different days of the year and different forecast lead times ranging from 1 to 14 days. An example plot of this array of correlation coefficients is illustrated in Fig. 7 for the North Fork of the American River, California. Figure 8 presents correlation plots for precipitation forecasts for each of the 24 basins. These plots show that forecast skill varies during the year, with forecast lead time and with location in the US. Correlation tends to be much stronger in the west during the cool season. It also seems to be stronger closer to the sea than



This plot shows how the coefficient of correlation between GFS forecast precipitation and observed precipitation varies during the year, depending on the event being forecast. This plot was constructed for each of the 24 selected MOPEX basins. Separate correlation plots were made for precipitation, tmin and tmax

Fig. 7 Variation of GFS Precipitation Forecast Skill during the year as a function of forecast duration for the North Fork of the American River, California

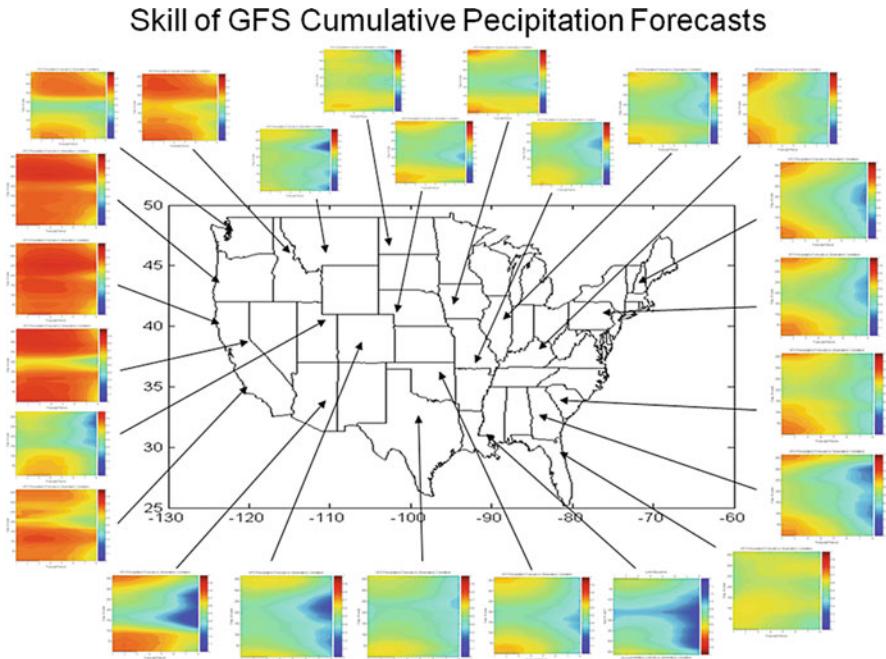


Fig. 8 Variation of GFS Precipitation Forecast Skill during the year as a function of forecast duration for 24 basins throughout the US

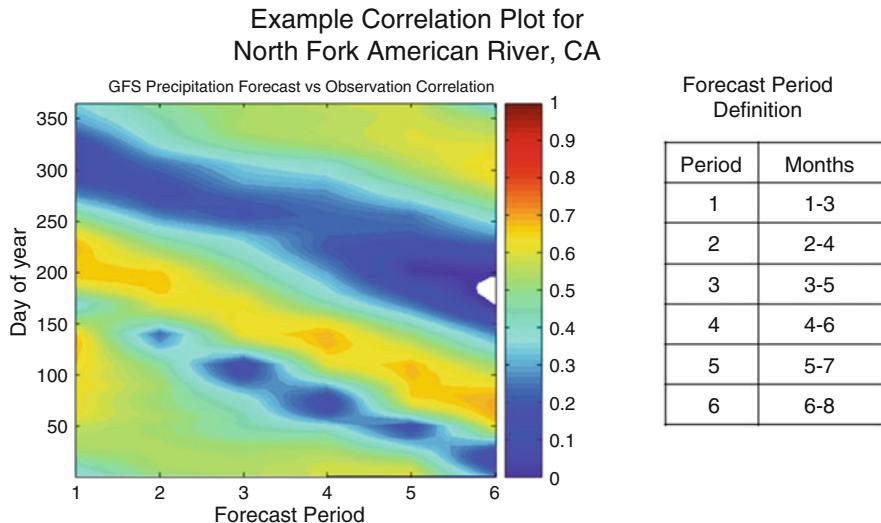
in the interior of the country where there seems to be almost no skill at any forecast lead time during the warm season.

5.1.3 Analysis of Seasonal CFS Precipitation Forecast Skill in the US

This section provides a general overview of the skill of CFS seasonal precipitation forecasts for different forecast lead times at different times of the year and at different locations throughout the US. The measure of forecast skill used is the Pearson coefficient of correlation between observations and GFS ensemble mean forecasts. Data presented here are for 24 basins distributed across the US.

The forecast variable chosen for this study is the mean precipitation amount over seasonal periods ending up to eight 30-day months in the future. These periods were chosen because the primary purpose of this version of CFS is for seasonal prediction and there is little forecast skill in CFS reforecasts for periods shorter than one season. An array of correlation coefficients between observed and CFS raw ensemble mean forecasts of this variable were computed for different days of the year and different forecast lead times (to the end of the forecast period) ranging from 3 to 8 months. An example plot of this array of correlation coefficients is illustrated in Fig. 9 for the North fork of the American River, California.

Presented in Fig. 10 are correlation plots for precipitation forecasts for each of the 24 basins. The most obvious result in these plots is that forecast skill is much more



This plot shows how the coefficient of correlation between CFS forecast precipitation and observed precipitation varies during the year, depending on the event being forecast. This plot was constructed for each of the 24 selected MOPEX basins. Separate correlation plots were made for precipitation, tmin and tmax.

Fig. 9 Variation of CFS Precipitation Forecast Skill during the year as a function of forecast season for the North Fork of the American River, California

dependent on the event being forecast than on the forecast lead time. Some events in the west can be predicted much better than may have been recognized using other approaches to seasonal forecasting. Figure 10 clearly shows that correlation tends to be much stronger in the west. But the correlation in the west seems to be stronger during the beginning of the winter precipitation season than during the main part of the season that occurs from December to February. It also seems to be stronger closer to the sea (or along the northern US border) than in the interior of the country. But there also seems to be a phase shift from west to east toward greater skill occurring in the east for events that occur later in the year than in the west.

5.1.4 Analysis of Uncertainty in GCM Seasonal Forecasts

The Spearman correlation coefficient has been computed for year-to-year time variations in precipitation covering a 28-year time period. This measures how well forecasts predict year-to-year variability in the precipitation. To measure prediction skill at different temporal and spatial scales, the Spearman correlation coefficient was computed for 9 temporal scales and 9 spatial scales. The temporal scales range from the first 15 days to 8 months. The spatial scales range from approximately a 1×1 to a $17 \times 17^\circ$ box that always maintains the same central grid. The results from this experiment for two different locations, a and b, for forecasts starting in January, April, July, and October are given in Fig. 11. The two locations lie on the same longitude but are separated by about 10° in the latitude.

CFS Seasonal Forecast Skill - Precipitation

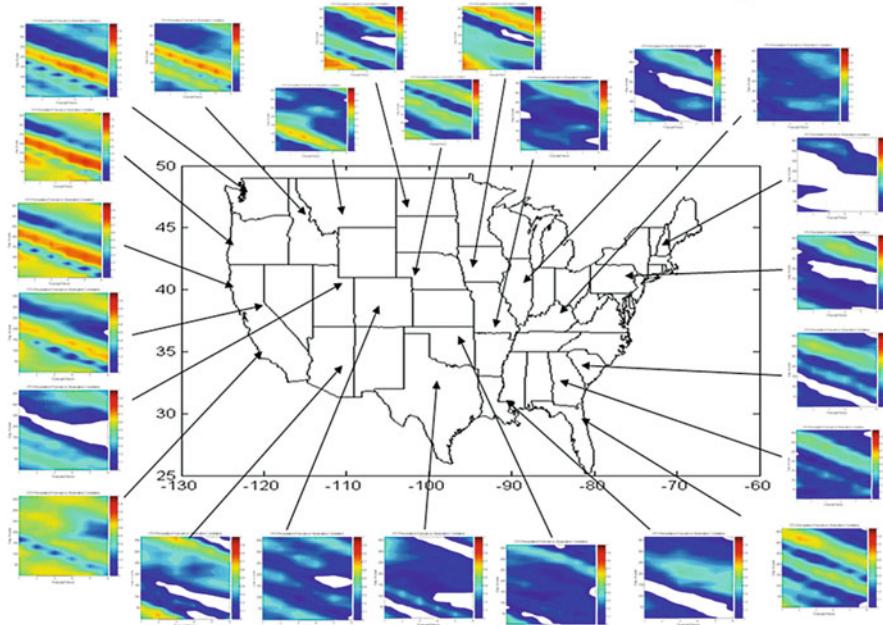


Fig. 10 Variation of CFS Precipitation Forecast Skill during the year as a function of forecast season for 24 basins throughout the US

At location (a), the correlation is the strongest in January followed by October and the correlation is weakest during April and July. Location (b) has the highest skill in July followed by April and January. The correlation is weak in October. This demonstrates large differences in GCM predictability of precipitation with only relatively small changes in forecast spatial location. Such seasonal differences in GCM skill will cause seasonal variability of skill in hydrometeorological forecasts that utilize the GCM.

In addition to seasonal differences in predictability, the two locations also exhibit differences in skill at different temporal and spatial scales. For example, at location (a), for January there is a clear decrease in skill from 15 days to the 1 month forecast. Furthermore, the highest correlations are at large temporal scales 0–6 months and at large spatial scales (17°). In comparison, location (b) has the highest correlations in July of which the strongest correlation occurs at the finer spatial and temporal scales. This behavior is flipped in April when the correlation is low at finer temporal and spatial scales and is higher at coarser spatial and temporal scales.

The variation of correlation with spatial scale and season is likely connected with the attribution or the premise of predictability for location and time of year. For example, the predictability for location (a) in January is likely linked to slowly varying processes that impacts larger areas. This would be consistent with the predictive signal being related to spatial and temporal averaging. The obvious

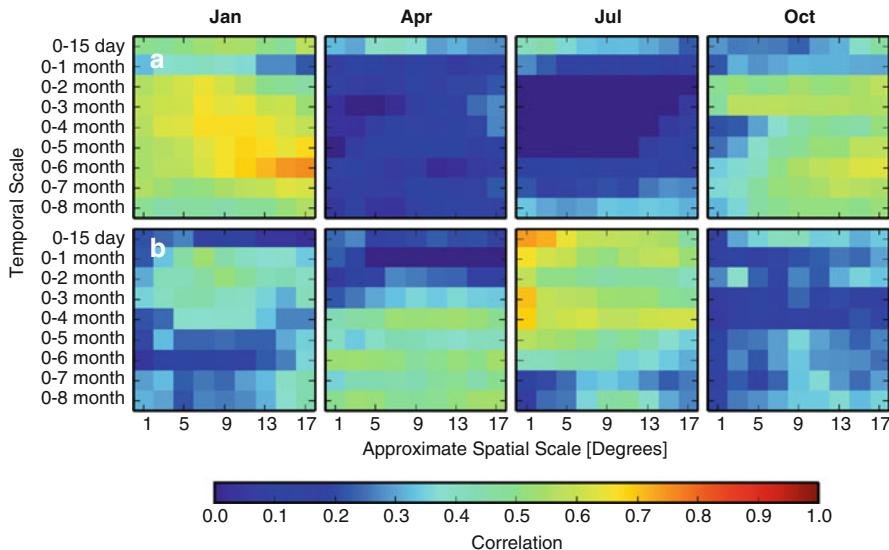


Fig. 11 Example of the year-to-year skill (Spearman correlation) of precipitation forecasts from a GCM at various spatial and temporal scales for two locations with the same longitude but with a 10° difference in latitude (a) and (b)

connection would be the slowly varying sea-surface boundary conditions in the GCM. In contrast, the acute predictability in July for location (b) would indicate that there is a more local aspect to its attribution. This could be due to the land boundary conditions in the GCM and the predictability they provide. This illustrates the importance of understanding how location, season, and scale influence GCM forecasts that subsequently influence hydrometeorological forecasts. Furthermore, aspects of location, season, and scale have implications for the application of these predictions. For example, the precipitation forecast in Fig. 11 for location (a) would be more suited for long-term drought prediction as opposed to short-range flood forecasts. In contrast, in July for location (b), there is greater potential for application to short-term flood forecasting. It is also evident there is little benefit for using the GCM for short-term forecasts at location (b) for January and April and seasonal prediction at location (a) for April and July. It is important to recognize that these results are specific to these locations and to this GCM. Other forecast models and locations could have very different predictive characteristics.

5.2 Hydrologic Model Uncertainty

Another major source of uncertainty in hydrometeorological predictions is associated with the hydrologic forecast models. An example of hydrologic model uncertainty can clearly be seen in Figs. 2, 3, and 4 by comparing the hydrologic model simulations (that use observed precipitation forcing) and the gauge observations. In

In this example, the hydrologic model is well correlated with the gauge observations, but there is a clear bias that varies with season. Although not shown here, there are other seasons and locations where correlations between the model simulations and the gauge observations are very low. Since these hydrologic model simulations incorporate observations of atmospheric forcing, these results are not limited by predictability of the atmosphere. Some of the uncertainties in the hydrologic model simulation shown in Figs. 2, 3, and 4 are associated with the atmospheric observations used to drive the hydrologic model as well as uncertainties in the hydrologic model itself.

Given these differences between hydrologic model simulations and observations, there are several ways to improve the predictions. One way is to calibrate the hydrologic model against observations. Hydrologic model calibration involves changing key model parameters to better match the observations. This can be done manually, but is often most effective when using an optimization algorithm that searches parameter space to find optimal parameter sets. Due to model complexity and structural model errors, usually no single parameter set gives the optimal solution. Instead, several parameter sets provide similar levels of consistency between the model and observations. These parameter sets can be used to run an ensemble of hydrologic model predictions using the same initial conditions and atmospheric forcing in order to quantify the uncertainty of the prediction associated with the estimation of the model parameters. Calibrating a model can be challenging due to the large amount of computations to effectively search the parameter space and requires a good observational data set. Part 5 of this book discusses model calibration and addresses many of the challenges with estimating model parameters and provides case studies of model calibration.

Even after calibration, there can still be local biases and errors between the hydrologic model simulation and observations. These errors can often be reduced by postprocessing the hydrologic prediction. Postprocessing requires a record of observations as well as a record of hydrologic simulations. Instead of changing model parameters, the observations and simulations are used to create a statistical relationship between the model and the observations. This relationship can remove bias in the predictions and can be used to generate ensemble traces of possible values of the observations to account for uncertainty in the predictions. Postprocessing of hydrometeorological predictions is discussed in Part 7 in this book.

The other important input to the hydrologic model is initial hydrologic conditions. This is very important for prediction but much uncertainty is associated with it. Key elements of initial hydrologic conditions include soil moisture, snow, and streamflow. There are techniques for measuring or estimating all of these variables through both in situ and remotely sensed methods. Although there is a perception that observations are perfect, there is uncertainty associated with them due to measurement methods and the fact that these measurements are not actual measurements of initial values of model state variables. Therefore, the best estimates of hydrologic initial conditions combine all observations with model estimates and information about the model errors and observational uncertainty using data assimilation techniques. This can provide temporally and spatially continuous estimates of

initial hydrologic conditions that incorporate uncertainty in the estimates. Part 6 of this book discusses data assimilation in hydrometeorological ensemble forecasting.

5.3 Verification

Although there are different models and methods for making ensemble hydrometeorological predictions, not all are equal in terms of their ability to provide reliable and skillful predictions. So, it becomes important to identify the best models and methods for each application (Roundy et al. 2015). Forecast verification is an important part of hydrometeorological prediction. There are many ways to verify forecasts. They range from simple deterministic measures to complex metrics that incorporate uncertainties. Many studies have compared ensemble hydrometeorological forecasts both from a climatological ESP approach and GCM-based atmospheric predictions (Mo et al. 2012; Yuan et al. 2013). These studies show that the GCM-based predictions provide some benefit over climatological ESP, but the skill of GCM-based forecasts varies in space (with location) and time of year. Part 8 of this book discusses forecast verification techniques and their communication.

5.4 Forecast Products and Services

Preparedness and response actions of emergency management authorities and the general public are highly dependent on the availability and dissemination of timely, skillful, and reliable hydrometeorological forecasting information. The usefulness of forecasts depends on how much confidence the forecast user has in them. Deterministic single-value hydrometeorological forecasts lack the uncertainty information that is needed for formulating proper actions. On the other hand, ensemble hydrometeorological forecasts offer numerous potential benefits (Krzysztofowicz 2001): (1) they are scientifically more “honest” than deterministic forecasts as they contain uncertainty information, which allows the forecast user to take risk information into account and make rational decisions; (2) they enable risk-based criteria for issuing disaster watches and warnings and for formulating emergency responses that are based on explicitly stated detection probabilities; and (3) they bring potential economic benefits of forecasts to society as a whole, which are achieved by initiating the necessary preventive measures or avoiding unnecessary overreactions to potential disasters. As the skill in hydrometeorological forecasting increases, society will continue to reap rich benefits.

One specific example of application of hydrometeorological ensemble forecasts is through the regulation and control of reservoirs. Reservoirs serve multiple purposes, including flood protection, electricity generation, water supply, recreation, and environmental and ecological protection. Skillful and reliable hydrometeorological forecasts are becoming increasingly important for reservoir operations to improve their socioeconomic, environmental, and ecological values. Deterministic single-value hydrometeorological forecasts are not suitable for developing reservoir

operations rules, which have generally been developed in an ensemble (i.e., probabilistic) framework using ensemble inputs to drive the reservoir operation models. Traditionally, a reservoir operation rule was developed by treating historical hydro-meteorological data from different years as ensemble inputs. With the availability of real-time ensemble forecasts based on NWP and climate models, reservoir operation is an area well positioned to take advantage of the predictive uncertainty information in ensemble hydrometeorological forecasts. Krzysztofowicz showed that the economic gain from a probabilistic temperature increases with the error in deterministic forecast (see Fig. 1.1 in Krzysztofowicz 1983). Stalling performed an evaluation which showed the economic benefit of hydrological forecasts in reservoir operation exceeds \$1B per annum (Stallings 1997). A challenge today is to develop improved operating rules that use ensemble hydrologic forecasts in ways that also satisfy constituent expectations for system operation.

6 Summary

Continuous hydrometeorological prediction of the water cycle is driven by estimates of initial hydrologic conditions and atmospheric forecasts of forcing variables such as precipitation and temperature. The premise of predictability for hydrometeorological predictions relies on the initial state of the atmosphere and the prediction of the land and sea boundary conditions as well as the initial state of the hydrologic system. Hydrometeorological models based on physical processes can be used to make hydrometeorological forecasts. But these predictions have strong variability in their skill depending on scale, season, and location as well as other factors associated with model parameterization, model structural errors, and input uncertainty.

There are many sources of uncertainty in hydrometeorological forecasting including uncertainties associated with forcing, initial and boundary condition, model structure and parameters, and observational datasets. A schematic illustrating an ensemble hydrometeorological forecasts system and its major uncertainties is presented Fig. 5. These uncertainties are introduced at different stages in the forecasting process and propagate through the model, and eventually are manifested as uncertainty in final forecast products. It would be ideal to be able to account for these uncertainties in the model equations, solve those equations numerically, and render a probabilistic forecast.

Forecast uncertainty can be accounted for by using an ensemble of hydrometeorological models with multiple parameter sets to account for parameter estimation and multiple hydrometeorological models to account for model structure and errors, incorporating in situ and remotely sensed observations with the off-line model simulations to account for the initial hydrologic state and utilizing observed climatology or predictions from GCMs to account for the uncertainty associated with the atmospheric forcing. It can also include ensemble postprocessing techniques to improve the reliability of ensemble forecasts. Combining all these methods to estimate the uncertainty associated with the hydrologic prediction could result in a

large number of ensemble members that would be computationally intensive and could result in very large uncertainty bounds.

The latest trend includes grand ensemble forecasting, i.e., the ensemble of ensemble forecasts generated by different weather or climate models. Many ensemble strategies are emerging. These range from “poor-men’s ensemble,” a simple combination of all deterministic forecasts from multiple models, to ensemble forecasts from multiple models, to “super-ensemble,” in which ensemble members are taken from ensemble forecasts from different models using a regression approach to favor members with higher correlation to observations. More recently, other grand ensemble strategies include a Bayesian model-averaging approach, which creates weighted probabilistic forecasts from the probabilistic forecasts generated by individual models. The weighting is assigned based on the likelihood of a model being correct in representing the real world. In the end, the most efficient way of dealing with uncertainty in hydrometeorological forecasting is through ensemble techniques. Finally, another strategy to improve hydrologic ensemble predictions is to post-process individual hydrologic ensemble members. That approach was used to produce reliable ensemble inflow forecasts for the New York City water supply “Operations Support Tool.” This is discussed in Part 9 of this book.

Although uncertainty will likely never completely disappear, reduction of uncertainty through model improvements, better estimation of initial hydrologic conditions, and improvements in GCM forecast and their use in hydrometeorological models will reduce uncertainty and improve the skill and reliability of hydrometeorological predictions. As uncertainty decreases, forecast usefulness will increase and provide decision makers with information needed to prepare for hydrometeorological extremes and ensure continued value for the life and the well-being of society.

References

- M. Cornick, B. Hunt, E. Ott, H. Kurtuldu, M.F. Schatz, State and parameter estimation of spatiotemporally chaotic systems illustrated by an application to Rayleigh–Bénard convection. *Chaos. Interdiscip. J. Nonlinear. Sci.* **19**, 13108 (2009). <https://doi.org/10.1063/1.3072780>
- G.N. Day, Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan. Manag.* **111**, 157–170 (1985)
- A.P.J. de Roo, B. Gouweleeuw, J. Thielen, J. Bartholmes, P. Bongioannini-Cerlini, E. Todini, P.D. Bates, M. Horritt, N. Hunter, K. Beven, F. Pappenberger, E. Heise, G. Rivin, M. Hils, A. Hollingsworth, B. Holst, J. Kwadijk, P. Reggiani, M. Van Dijk, K. Sattler, E. Sprokkereef, Development of a European flood forecasting system. *Int. J. River Basin Manag.* **1**, 49–59 (2003). <https://doi.org/10.1080/15715124.2003.9635192>
- E.S. Epstein, Stochastic dynamic prediction. *Tellus* **21**, 739–759 (1969). <https://doi.org/10.1111/j.2153-3490.1969.tb00483.x>
- G. Evensen, The ensemble Kalman filter for combined state and parameter estimation. *IEEE Control. Syst. Mag.* **29**, 83–104 (2009). <https://doi.org/10.1109/MCS.2009.932223>
- L.S.R. Froude, L. Bengtsson, K.I. Hodges, Atmospheric predictability revisited. *Tellus Ser. A Dyn. Meteorol. Oceanogr.* **65**, 19022 (2013). <https://doi.org/10.3402/tellusa.v65i0.19022>

- L. Goddard, S.J. Mason, S.E. Zebiak, C.F. Ropelewski, R. Basher, M.A. Cane, Current approaches to seasonal to interannual climate predictions. *Int. J. Climatol.* **21**, 1111–1152 (2001). <https://doi.org/10.1002/joc.636>
- TM. Hamill, Comments on “Calibrated Surface Temperature Forecasts from the Canadian Ensemble Prediction System Using Bayesian Model Averaging.” *Mon. Weather Rev.* **135**, 4226–4230 (2007).
- A. Hamlet, D. Lettenmaier, Columbia River streamflow forecasting based on ENSO and PDO climate signals. *J. Water Resour. Plan. Manag.* **125**, 333–341 (1999). [https://doi.org/10.1061/\(ASCE\)0733-9496\(1999\)125:6\(333\)](https://doi.org/10.1061/(ASCE)0733-9496(1999)125:6(333))
- B. Hoskins, The potential for skill across the range of the seamless weather-climate prediction problem: a stimulus for our science. *Q. J. R. Meteorol. Soc.* **139**, 573–584 (2013).
- T.G. Huntington, Evidence for intensification of the global water cycle: Review and synthesis. *J. Hydrol.* **319**, 83–95 (2006). <https://doi.org/10.1016/j.jhydrol.2005.07.003>
- T.R. Karl, B.E. Gleason, M.J. Menne, J.R. McMahon, R.R. Heim, M.J. Brewer, K.E. Kunkel, D.S. Arndt, J.L. Privette, J.J. Bates, P.Y. Groisman, D.R. Easterling, U.S. temperature and drought: Recent anomalies and trends. *Eos. Trans. Am. Geophys. Union* **93**, 473–474 (2012). <https://doi.org/10.1029/2012EO470001>
- R. Krzysztofowicz, Why should a forecaster and a decision maker use Bayes theorem. *Water Resour. Res.* **19**, 327–336 (1983). <https://doi.org/10.1029/WR019i002p00327>
- R. Krzysztofowicz, The case for probabilistic forecasting in hydrology. *J. Hydrol.* **249**, 2–9 (2001). [https://doi.org/10.1016/S0022-1694\(01\)00420-6](https://doi.org/10.1016/S0022-1694(01)00420-6)
- C.E. Leith, Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418 (1974). [https://doi.org/10.1175/1520-0493\(1974\)102<409:TSOMCF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1974)102<409:TSOMCF>2.0.CO;2)
- C.E. Leith, R.H. Kraichnan, Predictability of Turbulent Flows. *J. Atmos. Sci.* **29**, 1041–1058 (1972).
- H.B. Li, L.F. Luo, E.F. Wood, J. Schaake, The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.* (2009). <https://doi.org/10.1029/2008jd010969>
- E.N. Lorenz, A study of the predictability of a 28-variable atmospheric model. *Tellus* **17**, 321–333 (1965). <https://doi.org/10.1111/j.2153-3490.1965.tb01424.x>
- E.N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.* **26**, 636–646 (1969). [https://doi.org/10.1175/1520-0469\(1969\)26<636:APARBN>2.0.CO;2](https://doi.org/10.1175/1520-0469(1969)26<636:APARBN>2.0.CO;2)
- E.N. Lorenz, Atmospheric predictability experiments with a large numerical model. *Tellus* **34**, 505–513 (1982). <https://doi.org/10.1111/j.2153-3490.1982.tb01839.x>
- L. Marchi, M. Borga, E. Preciso, E. Gaume, Characterisation of selected extreme flash floods in Europe and implications for flood risk management. *J. Hydrol.* **394**, 118–133 (2010). <https://doi.org/10.1016/j.jhydrol.2010.07.017>
- Merriam-Webster Prediction, In: Merriam-Webster.com. <http://www.merriam-webster.com/dictionary/canonicalform>
- K.C. Mo, S. Shukla, D.P. Lettenmaier, L.-C. Chen, Do climate forecast system (CFSv2) forecasts improve seasonal soil moisture prediction? *Geophys. Res. Lett.* **39**, L23703 (2012). <https://doi.org/10.1029/2012GL053598>
- F. Molteni, R. Buizza, T.N. Palmer, T. Petroliagis, The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996). <https://doi.org/10.1002/qj.49712252905>
- T.N. Palmer, D.L.T. Anderson, The prospects for seasonal forecasting – A review paper. *Q. J. R. Meteorol. Soc.* **120**, 755–793 (1994). <https://doi.org/10.1002/qj.49712051802>
- J.K. Roundy, X. Yuan, J. Schaake, E.F. Wood, A framework for diagnosing seasonal prediction through canonical event analysis. *Mon. Weather Rev.* **143**, 2404–2418 (2015). <https://doi.org/10.1175/MWR-D-14-00190.1>
- J. Schaake, J. Demargne, R. Hartman, M. Mullusky, E. Welles, L. Wu, H. Herr, X. Fan, D.J. Seo, Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth. Syst. Sci. Discuss.* **4**, 655–717 (2007).

- J. Sheffield, E.F. Wood, Global trends and variability in soil moisture and drought characteristics, 1950–2000, from observation-driven simulations of the terrestrial hydrologic cycle. *J. Clim.* **21**, 432–458 (2008). <https://doi.org/10.1175/2007JCLI1822.1>
- J.A. Smith, M.L. Baeck, G. Villarini, D.B. Wright, W. Krajewski, Extreme flood response: The June 2008 flooding in Iowa. *J. Hydrometeorol.* **14**, 1810–1825 (2013). <https://doi.org/10.1175/JHM-D-12-0191.1>
- E.A. Stallings, *The Benefits of Hydrologic Forecasting* (Silver Spring, Maryland, 1997)
- D.M. Straus, J. Shukla, The known, the unknown and the unknowable in the predictability of weather. **175**, 20 (2005) [Available from the Center for Ocean–Land–Atmosphere Studies, 4041 Powder Mill Rd., Suite 302, Calverton, MD 20705].
- Z. Toth, E. Kalnay, Z. Toth, E. Kalnay, Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.* **125**, 3297–3319 (1997). [https://doi.org/10.1175/1520-0493\(1997\)125<3297:EFANAT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2)
- G. Villarini, J.A. Smith, R. Vitolo, D.B. Stephenson, On the temporal clustering of US floods and its relationship to climate teleconnection patterns. *Int. J. Climatol.* **33**, 629–640 (2013). <https://doi.org/10.1002/joc.3458>
- X. Yuan, E.F. Wood, J.K. Roundy, M. Pan, CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. *J. Clim.* **26**, 4828–4847 (2013). <https://doi.org/10.1175/JCLI-D-12-00683.1>
- J.H. Zar, Spearman rank correlation, in *Encyclopedia of Biostatistics* (Wiley, 2005). <https://doi.org/10.1002/0470011815.b2a15150>

Part II

Overview of Meteorological Ensemble Forecasting



Overview of Weather and Climate Systems

Huiling Yuan, Zoltan Toth, Malaquias Peña, and Eugenia Kalnay

Contents

1	Introduction	36
2	The Basis of Weather and Climate Predictability (Dynamical Systems)	37
2.1	Historical Perspective	37
2.2	Dynamical Systems	38
2.3	Deterministic Systems	39
3	The Limits of Predictability (Chaotic Dynamics)	40
3.1	Aperiodic Deterministic Systems	40
3.2	Chaos	41
3.3	Linear and Nonlinear Perturbation Characteristics	41
3.4	An Example of a Simple Chaotic System	44
3.5	Predictability	45
4	Features of Weather and Climate Systems from Mesoscale to Global Scales	47
4.1	Spatial and Temporal Scales	47
4.2	Large-Scale Precipitation	49
4.3	Convective Systems	53
4.4	Tropical Cyclones	55

H. Yuan (✉)

School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather, Ministry of Education, Nanjing University, Nanjing, China

e-mail: yuanhl@nju.edu.cn

Z. Toth (✉)

Global Systems Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration/OAR, Boulder, CO, USA

e-mail: zoltan.toth@noaa.gov

M. Peña

Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA

e-mail: mpena@uconn.edu

E. Kalnay

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA

e-mail: ekalnay@atmos.umd.edu

4.5	Monsoon	57
4.6	Low-Frequency (MJO, ENSO) Phenomena	58
5	Coupled Ocean-Atmosphere-Land Systems	60
5.1	Earth System	60
5.2	Global Circulation	62
	References	63

Abstract

Weather and climate phenomena develop as part of the coupled ocean-atmosphere-land-ice system. To understand the nature of the coupled system and its constituent processes, as well as the basis for and the limits of their predictability, some important concepts are reviewed, including determinism, chaotic error growth, and linear as well as nonlinear perturbation dynamics. It is shown that weather is predictable but only for finite times. Initial and model errors amplify, eventually rendering weather and climate forecasts useless. First, skill is lost in forecasts of fine-scale features while larger-scale phenomena remain predictable for longer periods of time. Processes and other characteristics associated with different scales of motion are discussed next, proceeding from the finest to the largest, coupled global-scale ocean-atmosphere phenomena. The potential use of ensemble forecast techniques to quantify scale and case-dependent predictability in the context of hydrologic forecasting is emphasized throughout.

Keywords

Predictability · Chaotic and dynamical systems · Nonlinear interactions · Scales of motion · Ensemble prediction systems (EPS) · Weather and climate systems · Coupled Ocean-atmosphere-land system

1 Introduction

In the context of hydrological prediction, meteorological ensemble prediction systems play an important role since their outputs are used to drive hydrological models. It is important then to understand the fundamental concepts behind meteorological ensemble forecasting. This section reviews meteorological ensemble forecasting as applied to weather and climate systems. The basic concepts of numerical weather prediction (NWP), including data assimilation, observing, and numerical modeling systems (chapter ▶ “[Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation](#)”), ensemble methods for meteorological prediction (chapter ▶ “[Ensemble Methods for Meteorological Predictions](#)”), current and future operational ensemble prediction systems (EPSs) (chapter ▶ “[Major Operational Ensemble Prediction Systems \(EPS\) and the Future of EPS](#)”), and climate variability and seasonal to interannual predictions (chapter ▶ “[Intraseasonal to Interannual Climate Variability and Prediction](#)”). Statistical post-processing of meteorological ensemble forecasts and verification methods of hydrometeorological ensemble forecasts are discussed in separate sections (“[Post-processing of Meteorological](#)

Ensemble Forecasting for Hydrological Applications” and “Verification of Hydro-meteorological Ensemble Forecasts”) of the handbook.

2 The Basis of Weather and Climate Predictability (Dynamical Systems)

2.1 Historical Perspective

Before the advent of numerical weather prediction (NWP) in the late 1950s and 1960s, skillful weather forecasts were not possible beyond 2 days lead time (Kalnay 2003). Many scientific and technological problems had to be overcome to obtain the significant accuracy seen in current prediction products such as forecasts generated, for example, during Hurricane Sandy in 2012, whose track was predicted 7 days ahead (Magnusson et al. 2014). This progress took several decades and is called by some “the quiet revolution of numerical weather prediction” (Bauer et al. 2015).

During the Cold War, the development of electronics first used in weapons systems led to the emergence of electronic computers. These machines allowed for the fast execution of computational tasks at speeds not possible previously. Von Neumann was the first to use computers in scientific applications – for weather forecasting (Kalnay 2003). Since then, atmospheric scientists have been early users of every newly available computational resources. Over time, the power of computers increased by orders of magnitudes since then, allowing the modeling of global atmospheric circulation at increasingly high resolutions. The analytical formulas describing the thermodynamical relationships between atmospheric variables in space and time were established in the first half of the twentieth century (Kalnay 2003). In the 1950s and 1960s, these relationships were discretized to facilitate their use for electronic computers. By integrating the resulting differential prognostic equations over a selected geographical domain, one could, with the use of computers, advance the state of the atmosphere from one time to a subsequent time level. The combination of scientific accomplishments with computational advances thus led to revolutionary changes in the field of weather forecasting.

In addition to the scientific and numerical formulation of the governing equations, and computational resources to solve them, successful weather prediction needed a third critical element. This is the capability to accurately estimate the initial state of the atmosphere from which the future states can be projected via the prognostic equations. The estimation of the initial state requires a thorough sampling of the atmosphere through observing networks. Traditional observing networks were designed to capture the synoptic-scale behavior of the atmosphere. Smaller scales however are typically sparsely sampled both in time and space, creating large uncertainties in the initial conditions of NWP systems. *Data assimilation* has emerged as a new science that can combine disparate observing networks to produce a physically consistent 3D state of the atmosphere, which is called an analysis.

The unprecedented advances in the development of the main pillars of NWP (computing facilities, scientific readiness, and observing and data assimilation systems) made successful weather prediction possible. Weather affects our society and environment constantly and, in some instances, dramatically. Societal impacts of these meteorological developments are unparalleled among other Earth sciences. The advances that made all this possible are available to other branches of Earth Sciences. Yet predictions of some other phenomena, such as earthquakes, are significantly less advanced than those of weather. What makes weather more predictable?

2.2 Dynamical Systems

In the following we refer to concepts of dynamical systems, a theory concerned about the time evolution of systems in multidimensional space, governed by a variety of rules, including ordinary differential equations. An example of a dynamical system is given by:

$$F = -kx \quad (1)$$

that describes the displacement x (with left/right displacements having negative/positive signs) of a rigid ball by a force F exerted by a spring characterized by a constant coefficient k and connected to a wall, from its original position (zero displacement, see Fig. 1).

Equation 1 follows Newton's second law $F = ma$, where m is the mass of the ball and a is the acceleration of the ball. Thus, the second derivative of x (with respect to time t) is equal to the acceleration of the ball. The simple Harmonic motion of the system illustrated by Fig. 1 is characterized by the differential equation:

$$m \frac{d^2x}{dt^2} + kx = 0 \rightarrow \frac{d^2x}{dt^2} = -\left(\frac{k}{m}\right)x \quad (2)$$

The solution of Eq. 2 can be written in a general form:

Fig. 1 A rigid ball connected to a wall via a spring



$$x(t) = C_1 \sin(\omega t) + C_2 \cos(\omega t) = A \sin(\omega t + \varphi) = A \sin\left(\sqrt{\frac{k}{m}}t + B\right) \quad (3)$$

The time evolution of weather and climate parameters (such as temperature, wind, and humidity) of course is described with different differential equations. Nevertheless, both the ball-spring system of Fig. 1 and the more complicated numerical models of the atmosphere are called dynamical systems as their time evolution follows a fixed rule (Eq. 3 for the ball-spring system and the differential equations used in weather models).

We note in passing that for computational applicability, numerical models of the atmosphere are discretized not only in space (i.e., gridpoints or other spatial representations, see Sect. 2.2) but also in time (i.e., using discrete time steps). To avoid numerical instabilities, finer spatial resolutions usually require more frequent time steps as well (Kalnay 2003). A general representation of a weather forecast model can take the form of:

$$X_{n+1} = f(X_n) \quad (n = 0, 1, 2, \dots), \quad (4)$$

where $n = 0$ represents the initial time and X_0 represents the atmospheric state at initial time, while X_1, \dots, X_n represent the weather conditions (atmospheric state) at different subsequent time steps. Since time is discretized, such dynamical systems are called *discrete* dynamical systems.

2.3 Deterministic Systems

A dynamical system is called deterministic if its time evolution is not influenced by random processes, and therefore its evolution can be exactly reproduced when the state of the system is perfectly known at any point in time. In other words, the future states of a deterministic system are uniquely determined by its initial condition. Therefore the evolution of a deterministic system can always be exactly reproduced if both its initial state and governing rules are perfectly known. Many deterministic systems exhibit periodic or quasi-periodic behavior. For example, the ball-spring system described by Eq. 2 exhibits a periodic time behavior (Eq. 3).

Examples of periodic or quasi-periodic deterministic systems with a harmonic oscillator like the ball-spring system include the simple pendulum and other rigid body problems such as solar and other astronomical systems. The periodic behavior in such systems arises from a balance or dynamic interplay between two or more forces such as the gravitational attracting force of the Sun and the centrifugal inertial force of the Earth traveling around it. Though due to limitations and errors in our measurements, the initial condition of real systems is never known exactly; thanks to

their periodic and therefore stable behavior, small errors in the initial condition of such systems remain small (i.e., the actual and forecast evolution or trajectories of such systems remain close forever). As a result, the behavior of such systems is predictable for long periods of time. As an example, the length of an Earth day or the annual cycle, for example, can be predicted with high precision (seconds) for long periods of time (million years). As we will see in the next section, not all deterministic systems, however, are as predictable.

3 The Limits of Predictability (Chaotic Dynamics)

3.1 Aperiodic Deterministic Systems

Not all deterministic systems are periodic or quasi-periodic. Notably, certain changes to systems with a periodic motion can render them aperiodic. One example of such changes is when the cord of a pendulum holding a weight is made elastic (Lynch 2002a, b):

$$k(l - l_0) = mg \quad \text{or} \quad l = l_0 \left(1 + \frac{mg}{kl_0} \right) \quad (5)$$

where l_0 marks the unstretched value of the length of the spring l (for more details, see Lynch 2002a, b). The dynamical system of the elastic pendulum can be written in the general form of:

$$\dot{X} + LX + N(X) = 0 \quad (6)$$

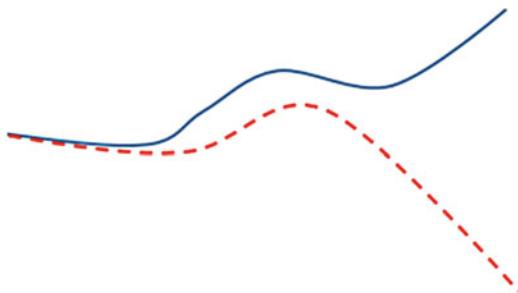
where X denotes the state vector of the dynamical system (characterized by the angle of the pendulum, radial and angular momenta, etc.) and L represents linear terms, while N stands for a nonlinear vector function.

Irregular behavior in such deterministic systems arises when certain forces (e.g., the centrifugal force of the pendulum) exceed some threshold values that the other forces of the system (the elastic cord) can no longer balance. Once these thresholds are reached, the simpler balance observed in periodic systems such as Eq. 2 is replaced with more complex behavior and the periodic motion changes to aperiodic. (For example, the acceleration of the elastic pendulum's weight (growth) is slowed and stopped when the elastic cord is stretched enough to provide sufficient force to slow and reverse the weight's acceleration.)

Interestingly, with the choice of a high enough viscosity parameter, laboratory or numerical computer models of the Earth's atmosphere exhibit close to that of periodic westerly propagating large-scale waves. When the viscosity is lowered and a critical threshold is reached, the behavior of these systems becomes unstable (i.e., aperiodic or irregular) in time.

Fig. 2 Schematic illustrating the growth of perturbations.

The blue line, from left to right, represents the time evolution of a chaotic system. The evolution of the system from a slightly perturbed state (red-dashed line) diverges with time from the blue trajectory



At the point where the dynamical systems turn aperiodic (e.g., viscosity or elasticity parameter thresholds are exceeded) new features, more complex interactions, and “dynamical” balances appear – e.g., we observe an ever-changing *dynamical* balance between the gravitational, centrifugal, and elastic suspension forces with the elastic pendulum. The underlying forces behind the evolving dynamical balances are often referred to as “instabilities” of the systems. It is important to point out that notwithstanding their aperiodicity, the behavior of such unstable systems is still deterministic – i.e., their behavior can be exactly reproduced or predicted as long as their state at a time and their governing rules are both exactly known. Determinism gives rise to predictability.

3.2 Chaos

Any dynamical system that is sensitive to initial conditions is chaotic. The atmosphere is one example of a chaotic system. Due to the instabilities, there is at least one perturbation pattern that if present as an error in the state of aperiodic deterministic systems will cause the perturbed state to diverge from the unperturbed evolution of the system (Lorenz 1963, 1969, 1972, see Fig. 2). Therefore, unlike periodic systems where initial errors persist but do not grow, in aperiodic deterministic systems errors amplify with increasing forecast lead time. This behavior, called chaotic error growth, is characteristic of aperiodic deterministic systems also called chaotic systems. The profound implications of chaotic behavior on the predictability of aperiodic or chaotic deterministic systems will be explored below.

3.3 Linear and Nonlinear Perturbation Characteristics

An actual error in the initial or forecast state of a chaotic system is just one of many possible perturbations one can make to the state of such a system. The study of perturbation behavior, therefore, is critical for understanding how errors may evolve. A convenient and relatively simple approach is the addition of infinitesimally small perturbations to the time-evolving state of a system. For this, governing equations

like Eq. 6 are “linearized” by dropping all nonlinear interactions (e.g., 3rd term of Eq. 6).

With the help of the linearized (or linear) equations and their inverse or adjoint version (Errico et al. 1993; Errico 1997), a host of metrics have been developed to analyze the behavior of simple or more complex dynamical systems. Out of all possible perturbations in a system (that is equal to the number of free variables in the system), we can determine, for example, the time-dependent perturbation vector exhibiting the largest amplification. This can be done (a) in an asymptotic sense, as the trajectory over which perturbations are evaluated covers the entire phase space of the system (i.e., leading Lyapunov vectors characterized by the Lyapunov exponents (LE) measuring their amplification in time, representing “sustainable” growth), or (b) over a selected segment of a trajectory (singular vectors (SV) and values, describing finite-time growth). It is important to point out that the perturbations are evaluated over a time-evolving flow. Therefore their growth (or contraction), measured, for example, by the instantaneous expansion of the local Lyapunov vectors (LLVs) corresponding to the global LEs, varies as a function of instabilities supported by the flow at any point on the trajectory of the system in the phase space (Fig. 3).

Importantly, the LEs characterize the long-term stability of a dynamical system. In an n -dimensional system analogous to that shown in Fig. 4, each axis j of the ellipsoid grows or decays over the long term by amounts given by $e^{\lambda_j t}$, where the λ_j are the Lyapunov exponents ordered by size $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$. The number of exponents larger/smaller than 0, for example, indicates how many expanding/shrinking perturbation patterns a system has that correspond with distinct instabilities and diffusive processes, respectively (Fig. 4). (The total volume of the ellipsoid in Fig. 4 evolve like $V_0 e^{-(\lambda_1 + \lambda_2 + \dots + \lambda_n)t}$; therefore, the sum of the Lyapunov exponents for a Hamiltonian (volume conserving)/dissipative system is zero/negative (Kalnay 2003).) Systems with negative-only LEs are stable (i.e., differences between states asymptotically do not amplify), whereas systems with at least one positive LE are unstable and chaotic. (Interestingly, finite-size chaotic systems necessarily also have an LE with zero value. Such LE corresponds to a Lyapunov vector that equals the time derivative of the system. Since the difference between two nearby points on a trajectory of a finite system remains finite and bounded, the corresponding LE must

Fig. 3 Lorenz’s experiment:
the difference between the
start of these curves is only
0.000127. (Adapted from
Stewart 1989, p. 141)

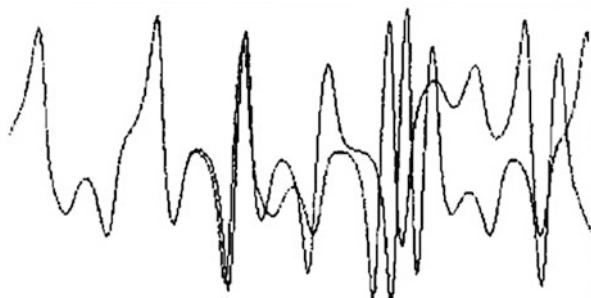
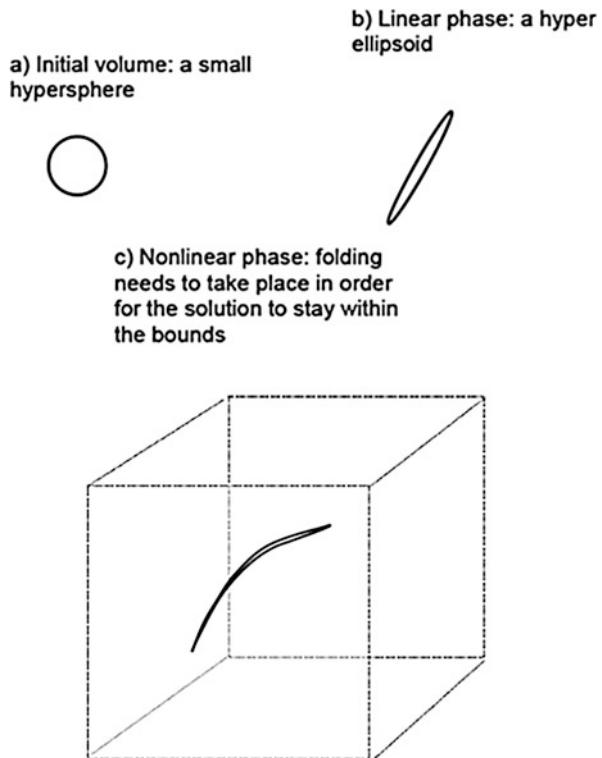


Fig. 4 Schematic of the evolution of states in a chaotic system's phase space
(a) initially contained in a spherical volume. States along expanding/shrinking perturbation directions are initially stretched apart and collapsed together linearly in a symmetric manner (**b**). The bounded solution space effecting nonlinear interactions is schematically indicated by the hypercube. The ellipsoid continues to be stretched in the unstable directions even as nonlinearities emerge, though it folds since the solution phase space is bounded (**c**). (Adapted from Kalnay 2003, Fig. 6.2)



be equal to zero (Kalnay 2003)). Since an initial error field, in general, has a non-zero projection in all perturbation directions, the error along the growing perturbation pattern in forecasts of such systems will necessarily amplify and dominate the total error until nonlinearities curtail perturbation growth.

Linear perturbations, however, indicate only how very small perturbations (and errors) evolve in the full nonlinear model. The growth of finite-sized perturbations in the full nonlinear model are limited by nonlinear interactions with other evolving processes in finite-sized chaotic systems (Fig. 4c). The nonlinear interactions, as discussed earlier, make the systems more complex and consequently the study of associated perturbations more challenging. Nonlinear interactions, for example, modulate both the orientation and growth rate of perturbations. In finite-sized systems, nonlinearities curtail the initial exponential growth that is characteristic of the leading LLVs by limiting the amplitude of perturbations to a so-called saturation level that corresponds to the size of the system in question. The behavior of linear perturbations, therefore, is often unrepresentative of how finite-sized nonlinear perturbations or errors behave in the full system.

Nonlinear perturbations can be evaluated by studying the difference between two nonlinear model integrations that started from two nearby points on the attractor (i.e., a collection of trajectories that the system can visit in the phase space of the model).

This “nonlinear perturbation” concept allows the generalization of some linear characteristics such as the Lyapunov vectors and exponents. Though these studies are computationally more intensive and scientifically less exhaustive, the bred vectors (BV, Toth and Kalnay 1993, 1997), finite-sized Lyapunov vectors (Boffetta et al. 1998), ensemble transform (ET, Wei et al.), and nonlinear local Lyapunov vectors (NLLV, Feng et al. 2017) have been proposed, among others as analogs of the linear Lyapunov vectors for the study of *sustainable* nonlinear perturbation growth. Nonlinear singular vectors and values (Mu 2000) and the conditional optimal nonlinear perturbations (Mu et al. 2003), on the other hand, are designed to capture finite-time nonlinear perturbation behavior, analogous to the linear SVs.

3.4 An Example of a Simple Chaotic System

Here we use a simple three-variable dynamical model system designed by Edward Lorenz (1963) who first discussed chaotic systems, to illustrate some other characteristics of such systems:

$$\begin{aligned}\frac{dx}{dt} &= \delta(y - x) \\ \frac{dy}{dt} &= rx - y - xz \\ \frac{dz}{dt} &= xy - bz\end{aligned}\tag{7}$$

where δ , b , and r are constant parameters and x , y , and z represent the three free variables of the system. This set of equations, often referred to as the Lorenz-63 is nonlinear as the time evolution of one of the free variables depends on the product of the other variables. To forecast the position $x(t)$, $y(t)$, and $z(t)$ of the system in phase space, we integrate Eq. 7 from a given initial condition. As seen from Fig. 3, with Lorenz' (1963), the system behaves chaotically, with its solution sensitive to the choice of initial conditions.

Under certain values of the Lorenz-63 model parameters, the solution is dissipative. This can be shown by expressing as the divergence of the system (Kalnay 2003):

$$\frac{\partial \dot{x}}{\partial x} + \frac{\partial \dot{y}}{\partial y} + \frac{\partial \dot{z}}{\partial z} = -(\delta + b + 1)\tag{8}$$

This equation shows that an original volume V in the Lorenz system (Eq. 8) contracts with time to $V e^{-(\delta + b + 1)t}$. This proves that a bounded and globally attracting set of zero volume trajectories called attractor exists that all nearby states will approach.

The Lorenz-63 model can also be expressed in the form of differential equations:

$$\dot{x} = F(x) \quad (9)$$

The Jacobean form of Eq. 9,

$$DF(x,y,z) = \begin{vmatrix} -\delta & \delta & 0 \\ r-z & -1 & x \\ y & x & -b \end{vmatrix} \quad (10)$$

can be used to study how the behavior of the system varies as a function of parameter choices. In particular, the local stability can be investigated by linearizing the flow. At the equilibrium or origin point $(0, 0, 0)$, the Jacobean Eq. (10) has the eigenvalues $\left(-b, \frac{1}{2}\left(-1 - \delta \pm \sqrt{(1 - \delta)^2 + 4r\delta}\right)\right)$. When parameter $r < 1$ (i.e., $\sqrt{(1 - \delta)^2 + 4r\delta} < 1 + \delta$), all three eigenvalues of the linear flow are negative $\left(-b, \frac{1}{2}\left(-1 - \delta - \sqrt{1 - 2\delta + 4r\delta + \delta^2}\right), \frac{1}{2}\left(-1 - \delta + \sqrt{1 - 2\delta + 4r\delta + \delta^2}\right)\right)$, indicating that all infinitesimally small perturbations decay and the origin is a stable and stationary point. At $r = 1$, the evolution of the flow changes abruptly (i.e., bifurcation, Kalnay 2003). When $r > 1$, one eigenvalue becomes positive, signifying the presence of an expanding perturbation pattern. This indicates that the flow diverges locally from the origin in one direction (Kalnay 2003), and hence the system becomes unstable. As r becomes larger than 1, two new stationary points also occur in the phase space with coordinates C_+ and C_- ($\pm\sqrt{b(r-1)}$, $\pm\sqrt{b(r-1)}$, $r-1$). The $r = 1$ and another bifurcation point in the Lorenz-63 system (see in Kalnay 2003 and Knill 2005) are examples for the separation of parameter regions mentioned in Sect. 3.1 where forces within the system can maintain a *simple* (stable solution) versus a *dynamical* balance (i.e., emergence of instabilities, leading to an unstable solution).

3.5 Predictability

Numerical models of the atmosphere are certainly deterministic systems that exhibit chaotic behavior. Today's models well simulate and, at short lead times, well predict atmospheric motions that they resolve, demonstrating their realism. Considering also a host of studies with simple and more complex models, we can postulate that the atmosphere as a whole and its subsystems (see Sect. 4) are chaotic dynamical systems, each with at least one expanding perturbation pattern or direction. And as Lorenz (1963) pointed out, the growth of perturbations in chaotic systems has serious implications for the predictability of weather.

Predictability is defined here as the lead time at which forecast error exceeds (or forecast skill drops below) a prespecified threshold. Assuming a numerical model can perfectly reproduce the behavior of a dynamical system, *intrinsic* predictability

(e.g., Lorenz 1969) is governed by the following factors: (a) the level of instabilities in the system, measured, for example, by the global Lyapunov exponents; (b) how instabilities vary along the trajectory (e.g., measured by the instantaneous expansion of the local Lyapunov vectors); (c) the error variance in the initial state of a forecast (i.e., initial error variance); and (d) how initial error variance projects onto and later evolves along dynamically characteristic vectors (e.g., LLVs). Numerical models of complex natural systems are never perfect, so *external* (or “practical”) predictability (Lorenz 1969) of such systems with imperfect models will also depend on the nature and level of model-induced errors.

As pointed out in Sect. 3.3, any error pattern, whether introduced into the initial state by the lack of sufficient observations or approximations in data assimilation or into the forecast state by an imperfect numerical model, will have a finite projection on the growing perturbations of a chaotic system. These errors projecting onto the growing perturbation directions will eventually render predictions useless. In particular, when errors saturate around the level typical of differences among randomly chosen states of the system, the forecast loses all predictive information. While statistics of forecast error variance as a function of lead time reveal the *average* rate forecast skill is lost, reflecting the presence of stronger or weaker instabilities, large case, and regime-dependent variations are observed in forecast performance. Exploiting that the evolution of perturbations is a good proxy for the behavior of errors, nonlinearly evolving perturbations in ensemble forecasts are used to gage variations in case-dependent predictability (see Sect. 2.3).

It is well established that spatial and temporal variability in the atmosphere are strongly connected. Larger-scale processes vary slowly not only in space but also in time. Correspondingly, the life cycle of larger-scale features is typically longer than their smaller counterparts. This is due to less intense instabilities that support only slower evolution of the larger-scale features themselves and slower error growth in their forecasts (e.g., Boer 2003).

Consistent with expectations from dynamical systems theory, the skillful forecast period has gradually expanded over the past decades as the error in the initial state of the atmosphere used to initialize the NWP forecasts and errors induced by imperfections in the NWP model have been reduced. In the early days, NWP forecasts could successfully predict only larger-scale features that already existed at analysis time and were captured in the initial condition of the models. As the data assimilation and forecast systems improved, smaller-scale motions and the emergence and evolution of second generation features (i.e., those not present at the initial time of the forecast) could also be successfully captured in NWP forecasts. But even today, only broad, conditional climatology-type statistics can be predicted for third-generation features that appear three life cycles into the future. No matter how much they are reduced – in chaotic systems like the atmosphere – with sufficient time even the smallest initial or model errors lead to a complete loss of predictability. As various atmospheric phenomena are reviewed in the following section (Sect. 4), it will be important to keep in mind that the finer the scale of a features, the more limited its predictability is.

As discussed in Sects. 3.1 and 3.2 of this chapter, determinism gives rise to both the existence of and the limits to predictability. Perfect predictability implies strict

determinism; however, as we have seen, limited predictability does not necessarily mean lack of determinism but, as in the case of the atmosphere, can indicate the presence of growing perturbations, a sign of deterministically based chaos.

4 Features of Weather and Climate Systems from Mesoscale to Global Scales

4.1 Spatial and Temporal Scales

Considering spatial scale, the most widely used classification of atmospheric scales is the system developed by Orlanski (1975), although Fujita (1981) modified the classification schemes (Markowski and Richardson 2010). According to Orlanski (1975), the atmospheric scale can be divided into microscale (less than 2 km), mesoscale (2–2000 km), and macroscale (large scale, larger than 2000 km) (Fig. 5; Markowski and Richardson 2010). Orlanski (1975) also indicated that the mesoscale meteorology can be divided into these subclasses: Mesoscale-alpha (meso- α , 200–2000 km), mesoscale-beta (meso- β , 20–200 km), and meso-gamma (meso- γ , 20–20 m), and meso-gamma (meso- γ , 20–20 m), and meso-gamma (meso- γ , 20–20 m).

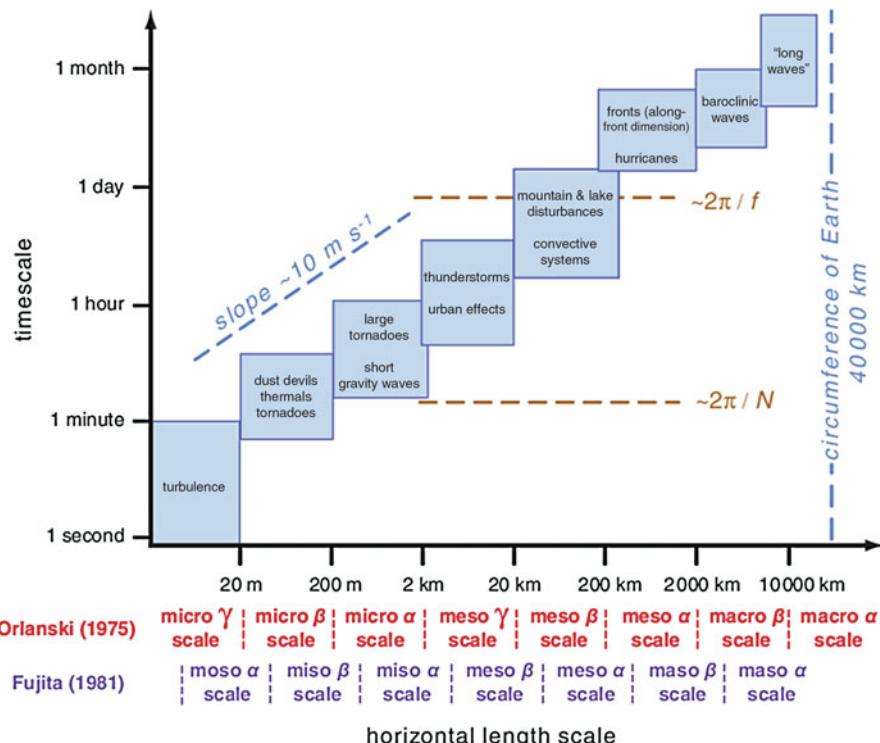


Fig. 5 Scale definitions and the characteristic time and horizontal length scales of a variety of atmospheric phenomena. Orlanski's (1975) and Fujita's (1981) classification schemes are also indicated. (Adapted from Markowski and Richardson 2010, Fig. 1.1)

(meso- γ , 2–20 km). Occasionally, the synoptic scale refers to the range of macro-scale atmospheric phenomena in space and time. The smaller spatial scales usually correspond to a shorter temporal scale. For example, the meso- α and meso- β systems can develop and dissipate in a few days. For different scales, the governing equations of atmospheric dynamical systems (see Sect. 2.2 of the handbook) sometimes can be simplified based on the relative importance for that scale. For example, the friction force can be neglected for high-level atmospheric movement above the planetary boundary layer, while the Coriolis force can often be neglected for the microscale phenomena (e.g., tornado). While the vertical accelerations are much smaller than the horizontal accelerations, the vertical pressure gradient force is balanced with gravitational force, and thus the atmospheric motion can be treated as hydrostatic equilibrium on the synoptic scale (Markowski and Richardson 2010), such as hurricanes and fronts. But on the microscale, vertical accelerations, such as those in a tornado, cannot be neglected, representing nonhydrostatic motions. For most of the mesoscale systems, especially for convective systems, both vertical and horizontal movements are important; the full governing atmospheric equations are needed to deal with energy and moisture variations.

Typical extratropical cyclones can be considered as of synoptic scale (Markowski and Richardson 2010). Mesoscale phenomena usually contain much shorter-scale motions with shorter Lagrangian timescales (the amount of time required for an air parcel to pass through the phenomenon) compared to the life cycle of extratropical cyclones (Markowski and Richardson 2010). For midlatitude meso- γ phenomena, the Lagrangian timescales can be influenced purely by buoyancy oscillations of about 10-min frequency (Fig. 5, $2\pi/N$; N is the Brunt-Väisälä frequency, i.e., buoyancy frequency). Such oscillations can be forced, for example, by short gravity wave motions. For midlatitude meso- α phenomena, the Lagrangian timescales can be extended to about 17 h ($2\pi/f$; f is Coriolis parameter), roughly a pendulum day, such as the low-level wind maximum near the top of nocturnal boundary layers, forced by the inertial instability.

As the atmospheric systems are across the scales, various weather and climate systems span different spatial (horizontal and vertical) scales and temporal scales. Consequently, the atmospheric predictability varies for different weather and climate systems. Sects. 2 and 3 of this chapter indicate that the atmosphere is a chaotic system and atmospheric predictability highly depends on initial conditions. The chaos theory (Lorenz 1963, 1969) laid a solid foundation for ensemble weather forecasting. Early studies (e.g., Epstein 1969; Leith 1974; Hoffman and Kalnay 1983; Kalnay 2003; Lewis 2005) suggested that an ensemble forecast with several members provide more skillful predictions than a single deterministic model. In the early stage, adding perturbations to the initial conditions of atmospheric NWP models was the predominant way to configure ensemble forecasts. Later, model related uncertainties were also simulated by varying different dynamical cores, physical parameterization packages, and/or physical parameter values across the ensemble members (see chapters ▶ “Ensemble Methods for Meteorological Predictions” and ▶ “Major Operational Ensemble Prediction Systems (EPS) and the Future of EPS”). The advancement of the computational resources and the

development of NWP models have facilitated the operational ensemble weather forecasts. Since the early 1990s, both the National Centers for Environmental Prediction (NCEP; Tracton and Kalnay 1993; Toth and Kalnay 1993; Toth et al. 1997) and the European Centre for Medium-Range Weather Forecasts (ECMWF; Palmer et al. 1993; Mureau et al. 1993; Molteni et al. 1996) have operationally provided global ensemble forecasts. Motivated by the success of global ensemble systems, many researchers and operational centers have developed ensemble forecasts with limited-area models at finer resolutions, to improve short-range weather forecasts, especially to extend the predictability of precipitation forecasting (Brooks et al. 1995; Du et al. 1997; Hamill and Colucci 1998; Buizza et al. 1999; Mullen et al. 1999; Stensrud et al. 1999; Toth et al. 2002). In Sect. 2.4 of the handbook, operational ensemble prediction systems (EPSs) currently run at major operational centers will be discussed in detail.

Based on temporal and spatial scale, the main types of EPS in weather forecasting include global EPS, regional EPS, and convective-scale EPS (WMO 2012). Currently, the typical global EPS usually forecasts 3–15 days, and its horizontal resolution is about 30–70 km. The regional EPS or the limited-area model EPS focuses on the 1–3-day forecasts and usually with a resolution between 7 and 30 km. Convective-scale EPS can resolve convection (sometimes refers as convection-permitting or convection-allowing) at its grid spacing, usually with a resolution of 1–4 km and a forecast length of 1–3 days. In operation applications, storm-scale EPS is also referred to convective-scale EPS. For such fine resolution, the prediction of convective systems can be much improved. While convective systems evolve rapidly with predictability on a short timescale, the forecast of convective-scale EPS can be quickly affected by atmospheric chaos. Several important weather and climate systems, which closely impact hydrological processes at different scales, will be discussed in this chapter.

4.2 Large-Scale Precipitation

Precipitation occurs when water vapor in the air condenses and falls to the Earth's surface, in liquid or solid phases, such as rainfall, snow, sleet, hail, graupel, freezing rain, and so on. Usually the process associated with vertical precipitation forming is considered as precipitation, while the horizontal condensation, such as frost, dew, and fog, closed to the ground surface are not considered as precipitation.

The mechanism precipitation formation is complex, affected by uncertainties under different weather and environmental scenarios. The basic conditions needed to generate precipitation in the atmosphere include the following: moisture transport in a region (water vapor), strong atmospheric updraft motion to form clouds through the cooling and condensation of air parcel, (unstable atmosphere, i.e., vertical movement), and enough nuclei to form rainfall droplet from cloud droplets (microphysical processes). The first two items are related to synoptic-scale environmental conditions, and the last one is determined by the cloud microphysics.

In general, two main microphysical mechanisms which are needed for cloud droplet growth are dominated (Ahrens 2008). In a cold cloud, while supercool water droplets coexist with ice crystals (most efficient between -10°C and -20°C), the ice crystals would gain mass by vapor deposition at the expense of the liquid droplets that would reduce mass by evaporation, which is called ice-crystal theory (also referred to as the Wegener-Bergeron-Findeisen process or simply the Bergeron process). In a warm cloud, the collision and aggregation/coalescence (merging of two water drops) are the major processes. Most precipitation from the clouds falls as snow, and varies by accretion, melting, and/or evaporation before reaching the ground, to then form as rainfall, snow, sleet, hail, and other types upon reaching the ground. These variations mainly depend on the vertical temperature and velocity profile.

The favorable environment needed to generate precipitation includes the development of synoptic weather systems, such as the strong convergence associated with fronts; the development of convective systems, such as the strong heating in summer; and the presence of orographic forcing, such as the windward uplifting. First, we introduce large-scale precipitation (synoptic scale), such as continuous summer rainfall, winter rainfall, or snowfall. Small-scale (local) precipitation (thunderstorm, hail, etc.) will be covered in the following convective systems. Large-scale precipitation highly affects the national economy and societal development in many ways, such as the direct and indirect impacts on agriculture, industry, defense, and transportation. It also frequently causes great damage of property loss and is a significant threat to human lives.

Many factors can impact the formation of precipitation and its type and distribution, such as atmospheric circulation, topography, and geographical location (such as lake and ocean effects), and the distribution of low-pressure areas (e.g., midlatitude cyclones. The horizontal scale of cyclones is measured by the last closed isobar on the pressure map. Usually, the horizontal scale of cyclones can vary from 1000 to 3000 km. On average, extratropical cyclones are stronger during winter than that during summer and tend to be stronger over the ocean than over land. Usually, the life cycle of a midlatitude cyclone includes four stages, which are the wave cyclone, maturity, occlude (closed cyclone), and dissipation stages. Its life cycle varies for different regions and cyclones. Midlatitude cyclones have a life cycle of 3–7 days. Baroclinic instability is the primary atmospheric condition for the development of a midlatitude cyclone. Cyclones without fronts include tropical cyclones (see Sect. 4.4 in the chapter “Overview of Weather and Climate Systems”), which can turn into hurricanes (or typhoons) as they intensify over tropical oceans. Local cyclones are produced by low-pressure systems due to orographic forcing or surface heating.

Generally, a front (sometimes referred to as a frontal zone) is an interface or transition zone between two air masses of different densities (Ahrens 2008). As such, it is another major weather system that can produce a variety of precipitation types. Since the air density varies with the temperature distribution, a front usually separates air masses of different temperatures, and the horizontal temperature gradient increases in the frontal region. When air climbs along the front, it condenses above the lifting condensation level and generates frontal precipitation. Such type of

precipitation is large scale and lasts for a few days, especially heavy precipitation (rain, snow, hail etc.) associated with cold fronts (the leading edge of a cooler air mass, replacing a warmer air mass at the surface). Similarly, orographic lifting forced by mountains causes ascending airflow, leading to more rainfall over the windward slope and much dryer conditions over the leeward slope. The upward deflection of horizontal larger-scale airflow by the high topography leads to generation of more clouds over the windward slope and increases the rainfall efficiency of condensed water drops in the clouds. On the other hand, the daytime heating of mountain surfaces also results in the air updrafts near the mountain peaks.

Rain is precipitation in the form of liquid water drops. For the classification of rain, different countries or regions adopt different criteria. In observations, the rainfall intensity is characterized by drizzle (small drops, lowest category), light, moderate, heavy, and extremely heavy rainfall, by reaching a given threshold (cm or inch per hour or day). The formation of heavy rainfall requires sufficient water vapor, strong updraft motion, and continuous unstable atmospheric conditions. Snow is a solid form of precipitation and is measured using snow depth or the amount of snow water equivalent. Based on snow intensity and visibility, snow can be categorized as light, moderate, and heavy snow. Usually quantitative precipitation forecasts (QPFs) refer to the water equivalent of all precipitation types (rain, snow, etc.) combined.

QPFs have improved in the past three decades. For example, since 1992 the skill (threat score, TS) of operational QPFs at NOAA (Fig. 6) has gradually increased for the threshold of 1 in./day (25.4 mm/24 h). The skill improvements are due to advancements in regional and global NWP, as well as to the introduction of ensemble forecasts in recent years. Note that the improvements of operational WPC (forecasters' input) versus single deterministic global (GFS) and regional (NAM) model gradually increase with time, especially in recent years. Such improvements indirectly reflects the positive impact of the usage of ensemble weather forecasting by the forecasters, since the application of operational EPS becomes more and more important in routine forecasts and decision-making systems. Interestingly, with the increasing horizontal resolution, the skill of global model outperforms that of regional models at a coarse resolution.

Using six operational global ensemble prediction systems (EPSs) in The Observing System Research and Predictability Experiment Interactive Grand Global Ensemble (TIGGE) data, Su et al. (2014) evaluated QPFs and probabilistic QPFs (PQPFs) of northern hemisphere (NH) summer daily precipitation during 2008–2012. Compared with the Tropical Rainfall Measuring Mission (TRMM) observations, their results show that the QPF and PQPF skill in the NH tropics is less than that in the NH midlatitudes for all six EPSs, suggesting the predictability of precipitation is low in the tropics and has room for improvements in current global EPSs. In addition, the biases in precipitation forecasts for all EPSs indicate that the ensemble post-processing (calibration or bias correction) is necessary to improve the precipitation prediction skill. In general, the European Centre for Medium-Range Weather Forecasts (ECMWF) EPS has the best skill of ensemble precipitation forecast among the six global EPSs. By using a higher model resolution, and

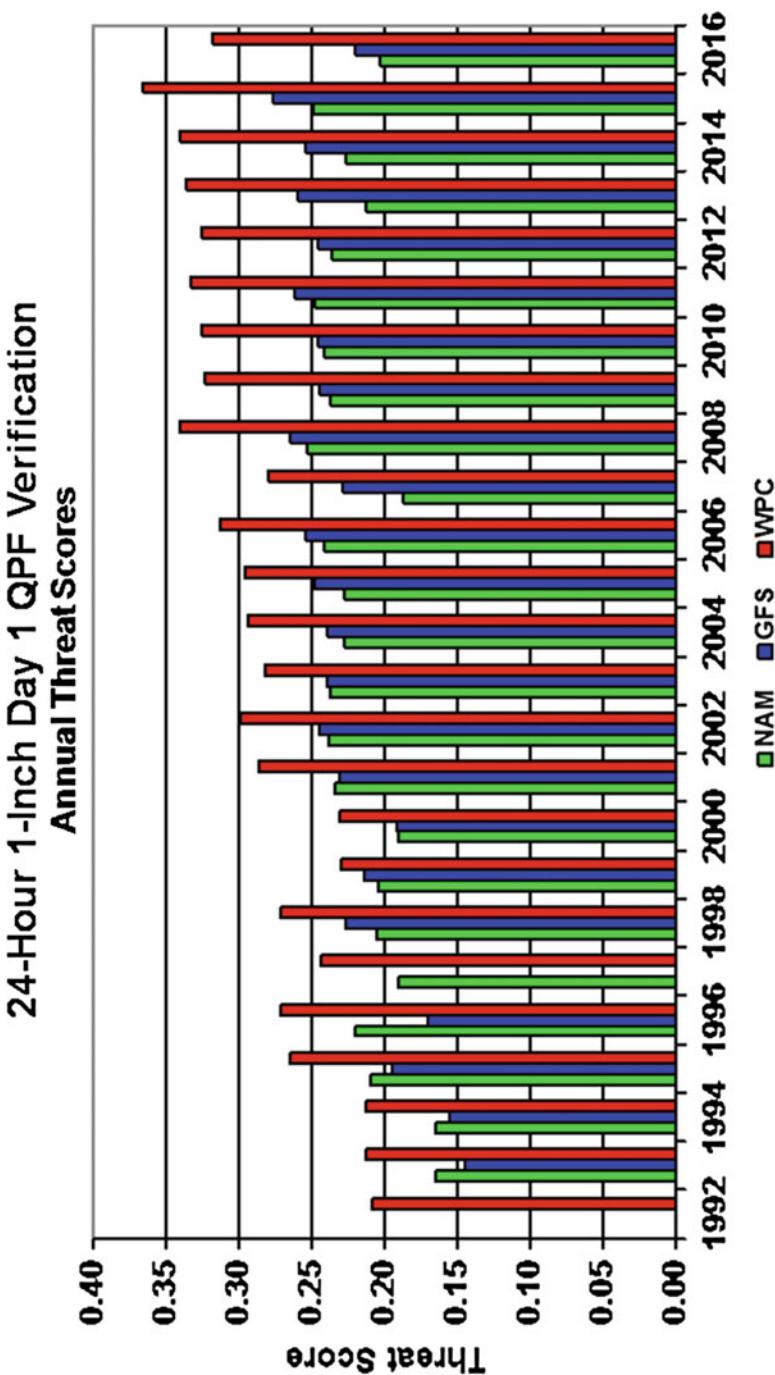


Fig. 6 National Oceanic and Atmospheric Administration (NOAA) quantitative precipitation forecast (QPF) skill during the period from 1993 to 2016 (<http://www.wpc.ncep.noaa.gov/images/hpcvrf/WPCmdsd110yrly.gif>). Regional model forecasts from the North American Model (NAM), global forecasts from the Global Forecast System (GFS), and officially issued forecasts by the NOAA/Weather Prediction Center (WPC)

improved data assimilation, physics, and perturbation methods, the ECMWF EPS has shown large improvements, reducing the seasonal variations in the skill of precipitation forecasts.

4.3 Convective Systems

Convective weather events (e.g., thunderstorm, tornado, squall line, hail, and so on) are associated with convective systems, which can develop violently and cause great amount of damage. Convective systems frequently generate violent weather events, usually accompanied with heavy rain, strong wind, and sometimes lightning and thunder. Typical convective weather events are on local scales and have short temporal durations (a few hours). Historically, there have been severe convective weather events that have lasted for a few days and affected to a large area. For example, on 7 April 2006, over 800 severe weather reports were recorded (including 91 tornado, 215 wind, and 565 hail reports) at the National Weather Service (NWS)/Storm Prediction Center (SPC), and 3 very strong tornadoes caused the loss of 10 lives in the United States (Bright 2007).

As convective systems develop quickly and locally, and the life cycle of convective systems is usually short, the forecasting of convective weather is extremely difficult and challenging. With the improvements in data assimilation methods and NWP mesoscale models (see Sect. 2.2 of the handbook), the forecasting of severe convective weather has been improved. In particular, by means of ensemble forecasting, short-range weather forecasting has made progress, in conjunction with the advanced observation technology (such as Doppler radar and satellite). Several types of convective mesoscale systems (MCSs) often occur together, such as tornado, squall line, hail, and thunderstorm, while, at the same time, each type of MCSs has unique feature. Convective weather events with MCSs have the destructive effects on human life and property and are always of primary focus when deciding observation networks and issuing forecast warnings.

Thunderstorms in general occur at local scales, are produced by a cumulonimbus (a principal cloud type, exceptionally dense and vertically developed) cloud, and often come with lightning and thunder, strong gusts of wind, heavy rain, and sometimes accompanied by hail. Thunderstorms are also electrically active (sometimes called electrical storms) with lightning phenomena. When observing weather phenomena, thunder and lightning are often recorded. The classification scheme for thunderstorms categorizes them as light, medium, or heavy, based on several factors, including the activity of the lightning and thunder, the type and intensity of the precipitation, the strength of the wind, and the variation of clouds and surface temperature. The thunderstorms associated with strong convection and severe weather (such as tornado, hail, etc.) are also called severe thunderstorms. Characterized by the synoptic meteorology and the weather situation, thunderstorms can be classified as an air mass thunderstorm, a frontal thunderstorm, or a squall line thunderstorm.

Single cell thunderstorms are about a dozen kilometers horizontally. Multicell thunderstorms, developed by mesoscale convective complexes (MCCs) or MCSs, usually cover over a few hundred kilometers. Thunderstorms often build to altitudes of 12–15 km in the midlatitudes and to even greater heights in the tropics. A thunderstorm usually has a short life cycle, and most thunderstorms last under 2 h, including the development, maturity, and dissipation phases. A thunderstorm is the result of an atmospheric instability (Yamane and Hayashi 2006), and they usually develop vertically until they reach the tropopause. At that point, they run into a global stable layer, the lower stratosphere, which limits their upward growth. In the early phases, strong convective updrafts are the dominant characteristics of a storm. In their mature phase, precipitation occurs in storms that generate a downdraft, while a strong updraft is the major feature above the downdraft in a cloud. In its dissipation phase, the storm is marked with a strong downdraft in a column of precipitation.

A squall line is a line of thunderstorms that form along a cold front or out ahead of it (Ahrens 2008), which has a larger length-to-width ratio compared with other types of MCSs and has active deep moist convection frequently associated with thunderstorms or cumulonimbus clouds. Squall lines contain multicell thunderstorms and can generate contiguous or multicell precipitation. It is characterized by strong down-drafts of cold air in strong thunderstorms and divergent flow at the surface, causing a dramatic change of wind (sudden gust) and pressure, a decrease of temperature, and the substantial increase of humidity. Compared to other thunderstorms, the squall line has a relatively longer life cycle of about 12–24 h, with a few hundred kilometers of horizontal extent. Hail (sometimes hailstone) is a solid state of precipitation generated from convective clouds (cumulonimbus). Usually hail is in the form of balls or irregular shapes of ice, with a diameter of 5 mm or larger. Favorable thunderstorms for hail formation (sometimes called hailstorm) typically contain large liquid water droplets and cloud drop sizes, as well as strong updrafts and a great vertical extent (to increase the suspension time of droplets, allowing them to grow large in size). Thunderstorms can generate tornadoes that can be seen as a funnel cloud or rotating debris/dust from the ground. A tornado typically is a cyclonically (occasionally anticyclonically) rotating column of air at the microscale (<2 km). It is the most violent vortex compared to all other atmospheric circulation on Earth. The classification of tornado intensity is often based on wind damage according to the Enhanced Fujita Scale. Tornadoes have been observed most commonly in the central plains and the southeastern states of the United States, with an average about 1000 tornadoes per year. Both hailstorms and tornadoes have a short life cycle, from a few minutes to an hour, posing many challenges for weather prediction.

The convective-scale (convection-permitting or convection-allowing) EPSs greatly improve the forecasts of convective weather events by accounting for the large uncertainties of convection initiation and the fast development of convective systems. In the history, for the abovementioned multiple severe weather events on 7 April 2006 in the United States, the first ever 48-h outlook of severe weather (high risk) was issued by NWS/SPC successfully, using short-range ensemble forecasts (Bright 2007). NWS/SPC has developed ensemble guidance for specific applications (such as severe weather, fire weather, winter weather, etc.). For example, besides

regular ensemble products (mean, median, max, min, spread, exceedance probabilities), the combined (or joint) probabilities, derived severe weather parameters (such as super cell and significant tornado parameters), and calibrated probabilities of thunderstorms and severe thunderstorms play an important role in predicting severe weather and providing weather outlook/watch/warning at NWS/SPC. For example, a tornado watch is issued by NWS/SPC if tornadoes are likely to form during the next few hours, while a tornado warning is issued if a tornado is spotted or detected by radar observations (Ahrens 2008).

4.4 Tropical Cyclones

A tropical cyclone is an intense rotating storm system occurring over the tropical oceans, characterized with a low-pressure center and a closed low-level atmospheric cyclonic circulation, strong wind, and heavy rain. Tropical cyclones frequently cause significant damage due to its accompanying torrential rain and storm surge over the tropical oceans and coastal areas. According to the wind scale specified in an international agreement (WMO 2017), the tropical cyclone can be classified for different regions or countries with categories (Fig. 7) delineated by the maximum sustained wind speed near the center of the tropical cyclone. A strong tropical cyclone is usually referred to as a typhoon over the western North Pacific and as hurricane over the North Atlantic and the central/eastern North Pacific.

Since 1970, tropical cyclone track errors have greatly improved for different regions; however, the decrease in intensity errors has been slower compared to the improvement in track forecast error (<http://www.nhc.noaa.gov/verification/verify5.shtml>). For example, in 2010 the track errors for 1–3-day hurricane forecasts reduced by 50% compared to 15 years ago, and the track errors for 4–5-day hurricane forecast reduced by 40% compared to 10 years ago (Cangialosi and Franklin 2011). Since the early 1990s, global ensemble forecasts have gradually become operational worldwide. For example, a vortex relocation technique was applied to NCEP's Global Forecast System (GFS) in 2000 and Global Ensemble Forecast System (GEFS) in 2004 (Liu et al. 2002). The EPS at ECMWF calculates moist singular vector for tropical cyclones with diabatic physics (Rhome 2009). ECMWF began issuing tropical cyclone forecasts in 2004 (ECMWF 2004; www.ecmwf.int). The operational regional EPS for tropical cyclones was established in Japan in 2007 (Yamaguchi and Majumdar 2010), using the singular vector method as done at ECMWF, but its moist singular vector has been applied to the whole tropical area (20°S–30°N), while it is applied only around the tropical cyclone at ECMWF. In China, operational EPS for tropical cyclones was implemented in 2007, with the vortex relocation technique (Qian et al. 2013).

Before the operational EPS of tropical cyclone, a “poor person” (or consensus) ensemble technique was used which took the average of different models or operational forecasts to improve track errors of tropical cyclones (Goerss 2000). And various ensemble post-processing methods for tropical cyclone track have been developed, such as ensemble mean by removing poorly performing members

Comparison of the Tropical Cyclone Classification										
western North Pacific										
North Atlantic, central/eastern North Pacific										
Maximum Sustained Wind Speed near the centre of the tropical cyclone	Hong Kong, China (10-minute average) HKO	China (2-minute average) CMA	Japan (10-minute average) RSMC, Tokyo	United States (1-minute average) JTWC	United States (1-minute average) CPHC, NHC	United States (1-minute average)				
kt	km/h	m/s				United States (1-minute average)				
Tropical Depression (TD)										
< 34	< 63	< 17.1	Tropical Storm (TS)					Tropical Storm		
34 – 47	63 – 87	17.2 – 24.4	Severe Tropical Storm (STS)					Hurricane categories		
48 – 63	88 – 117	24.5 – 32.6	Typhoon (T)					1: 64 – 82 kts		
64 – 80	118 – 149	32.7 – 41.4	Severe Typhoon (ST)					Typhoon 64 – 84 kts		
81 – 99	150 – 184	41.5 – 50.9	Very Strong Typhoon 85 – 104 kts					Typhoon 64 – 129 kts		
			Typhoon 105 – 129 kts					2: 83 – 95 kts		
			Typhoon 130 – 149 kts					3: 96 – 112 kts		
			Violent Typhoon ≥ 150 kts					4: 113 – 136 kts		
≥ 100	≥ 185	≥ 51.0	Super Typhoon (SuperT)					Super Typhoon ≥ 130 kts		
			Super Typhoon ≥ 137 kts					5: ≥ 137 kts		

Fig. 7 Comparison of the tropical cyclone classification. (Courtesy: <http://www.typhooncommittee.org/wp-content/uploads/2015/08/tc-classification1.jpg>).
 (Note: the conversion between kt to km/h and kt to m/s may vary slightly subject to rounding practices and conversion factor decimal places. Acronym: HKO Hong Kong Observatory, CMA China Meteorological Administration, RSMC Regional Specialized Meteorological Centre, Tokyo, JTWC Joint Typhoon Warning Center, CPHC Central Pacific Hurricane Center, Hawaii, NHC National Hurricane Center, Miami)

(Elsberry and Carr 2000) and multivariate linear regression (Williford et al. 2003), that show improved tropical cyclone track forecasts. In particular, the single deterministic forecast has limited skill in forecasting tropical cyclones with an irregular track, but ensemble forecasts may provide useful information for such tropical cyclones (such as Qian et al. 2013). However, as the single model has relatively lower skill for forecasting the intensity of tropical cyclones, ensemble forecasts and post-processing usually show fewer improvements in forecasting the intensity of tropical cyclones, unlike the obvious improvement in tropical cyclone track using ensemble forecasts (Hamill et al. 2012).

4.5 Monsoon

The name for the circulation regime monsoon comes from the Arabic word “mausim,” meaning a season. During early times, people used monsoons to represent the seasonal reversal of surface winds over the Arabian Sea, southern Asia, and the Indian Ocean (Ding et al. 2015). The southwest surface wind prevails during the warm half of the year (summer) and the northeast wind during the cool half of the year (winter). The East Asian monsoon system (Fig. 8) includes the tropical monsoon over the South China Sea, western Pacific region, and the subtropical monsoon over the continental China, Japan region. Influenced by the shift in the prevailing surface winds, precipitation patterns over the monsoon region also shift with seasonal variations. For example, in East Asia, the precipitation and flood season is closely related to the annual development of monsoon circulation. The monsoon is

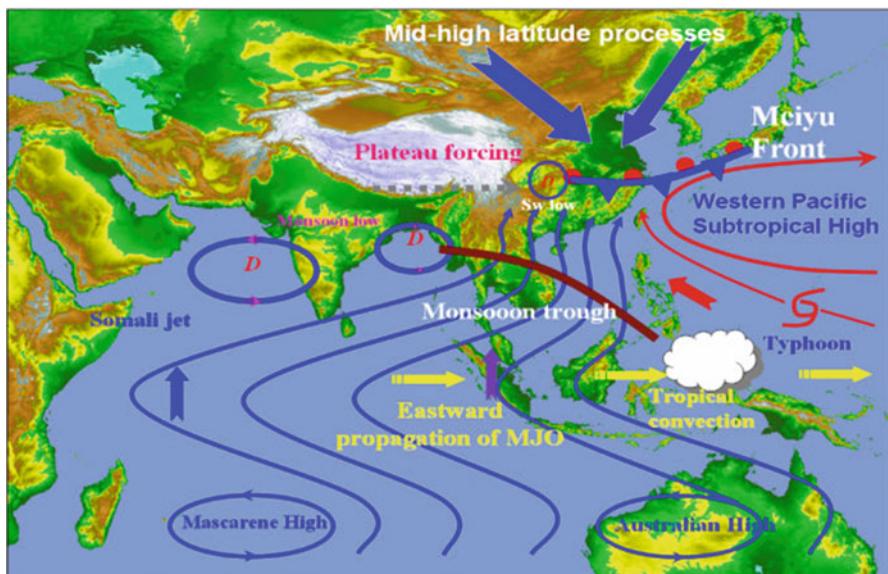


Fig. 8 The East Asian monsoon system. (Courtesy: Ding et al. 2015)

associated with the atmospheric circulation, so when the winds transport more moisture onshore from the oceans, the rainy season can be maintained for the summer months with a high level of humidity. By contrast, a dry winter season begins when the winds blow offshore and thus is accompanied by low humidity from the continents.

The monsoon is not just a regional climate phenomenon; it is also on the planetary scale (Ramage 1971). In 1971, Ramage defined the monsoons in Asia, Africa, and Australia using the shift of surface winds (Wang et al. 2011; Ding et al. 2015). The concept of the global monsoon (Trenberth et al. 2000; Wang et al. 2011) has been recognized in recent years, since the regional monsoon should be studied in isolation. While the difference between local summer and winter precipitation rate exceeds 2.5 mm/day and the rate of local summer/annual precipitation exceeds 55%, the global monsoon precipitation domain is approximately defined (Wang et al. 2011). The regional monsoons (Wang et al. 2011) are the North American monsoon (NAM), North African monsoon (NAF), Indian monsoon (IND), East Asian monsoon (EAS), western North Pacific monsoon (WNP), South American monsoon (SAM), South African monsoon (SAF), and the Australian monsoon (AUS).

A conceptual model of monsoon was described by Webster (1987), based on the large-scale atmospheric circulation. The planetary-scale monsoon is caused by the seasonal oscillation of solar heating in the hemisphere, the land-sea thermal and pressure contrast, the deflected wind due to the rotation of the Earth, and moisture processes through convection (Webster 1987). While net heating plays a more important role on the global monsoon system response, the thermal distribution of land and ocean with topography determines the evolution of the regional monsoons. However, monsoon circulations are associated with complex climate systems and have strong climate variability in terms of their onset, length, and break periods. In recent studies, Ding et al. (2015) pointed out that the monsoon system is a response of the coupled ocean-land-atmosphere-cryosphere system under the annual cycle of solar heating. Monsoons are also affected by global climate change and human activities (Ding et al. 2015). The variability of the summer and winter monsoon cycle poses challenging problems for forecasting monsoons accurately.

4.6 Low-Frequency (MJO, ENSO) Phenomena

Predictability has a strong scale dependence (see, e.g., Toth and Buizza 2018). Some slow-evolving phenomena such as a drought or the anomaly of a climate variable with respect to a preestablished reference can potentially be predicted in a probabilistic form weeks ahead. Considering the limit of “traceable” atmospheric predictability (Toth and Buizza 2018), the spatiotemporal position of smaller- or medium-scale weather events, though, cannot be predicted beyond a few days or up to 2 weeks lead time. Beyond the predictability limit of individual events, the statistics (such as the frequency) of finer-scale weather events, however, may still be predictable (i.e., “climatic predictability,” Toth and Buizza 2018), contingent on the successful prediction of larger-scale phenomena. Example of this type of predictions

includes the likelihood that one or more hurricanes will hit the NE USA next month or a forecast for “higher than normal” precipitation, caused by anomalously frequent storms, in California one season ahead. These types of forecasts are referred to in general as short-term climate forecasts or other various names depending on the time intervals used, for example, monthly, subseasonal, seasonal, interannual, and decadal forecasts. They all serve particular needs and contain useful forecast information at the expense of not detailing the timing and specific location of the target phenomena or anomaly. There are several aspects and levels of refinements at play that the expert takes into account to optimize the information contained in model forecast outputs. First, the “higher than normal” invokes a climate norm and a specific “binning” of the histogram of the event predicted. Defining the norm and choosing the appropriate binning is generally determined by the sample size (i.e., how many events in the past have been observed in the historical data available) and the size of the ensemble forecast. Predicting beyond 2 weeks requires the use of large ensemble forecasts to be able to extract the predictable signal by filtering out the transient anomalies. Predictable signals are usually associated with slow-evolving phenomena.

It is clear then that the success of predictions beyond 2 weeks will depend on whether the models are able to reproduce the most significant phenomena or the evolution of key climate variables. Besides the annual cycle and diurnal variations, which are quasi-periodic signals in some variables, two of the most significant phenomena are the Madden-Julian Oscillation (MJO) and the El Niño-Southern Oscillation (ENSO). In more general terms, these two signals belong to the low-frequency variability. This variability arises from the combination of multiple physical processes including atmosphere-ocean interactions (e.g., surface fluxes and Ekman transport, ocean thermal inertial, and reemergence of anomalies) occurring locally and their exerted remote forcing. Figure 9 shows the spatial and temporal scales of typical phenomena including MJO and ENSO. Typical modes of variability are described in detail in Sect. 2.5.

Forecasts beyond the traditional short-term climate (out to a year approximately) time range are produced to assess climate change impacts. The forecast range typically goes from decadal to centennial time projections. In several respects, this type of forecast differs from the short-term climate forecasts, but as numerical climate models become more sophisticated, the two are becoming much more alike. The models used for climate change predictions tended to enhance the mechanisms driving long-term climate such as atmospheric composition and regarded the initial conditions less relevant than their counterpart. Currently, numerical models used for short-term climate prediction are also used for climate change; however, the approach is usually different. In climate change, the interest is primarily on understanding the sensitivity or measuring the impact of greenhouse gases. For example, doubling the CO₂ concentration might have impact on the global and local climatology of precipitation, temperature, and sea-level rise.

The short-term forecasts of key climate variables are used as input to “downstream models,” such as river route models, and hydrological models, which are important tools for decision-makers for planning, risk management, disaster

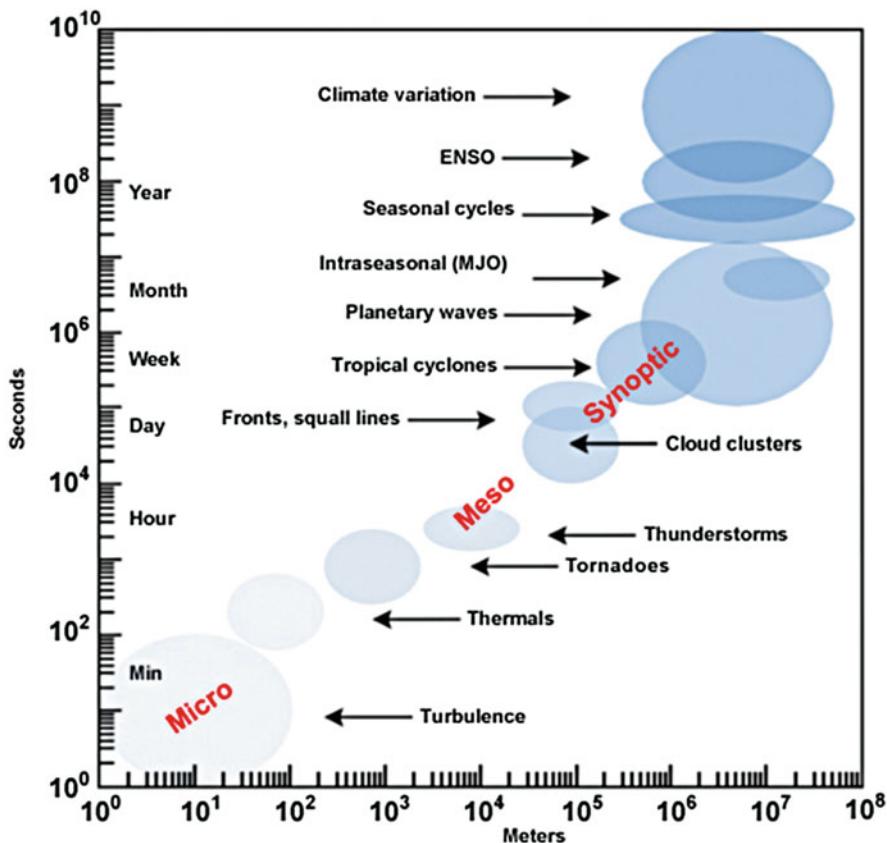


Fig. 9 Typical spatial and temporal scales of phenomena in the climate system. (Credit: The COMET Program). https://www.meted.ucar.edu/about_legal.php

prevention, and mitigation. Models are referred to as “downstream” in the sense that they do not influence the evolution of the climate model forecast. They are important in increasing the realism of model outputs from the climate models by adding a physical model that takes into account the heterogeneity of land and local processes including surface runoff, subsurface flow, and evapotranspiration principally.

5 Coupled Ocean-Atmosphere-Land Systems

5.1 Earth System

The Earth is a natural system in which weather and climate phenomena occur as a result of internal and coupled physical processes spanning over a wide range of temporal and spatial scales. From a macroscale perspective, the evolution of the

Earth system can be attributed to the interaction among four major subsystems: atmosphere, hydrosphere, geosphere, and biosphere. The interactions consist of transforming energy and transporting mass through chemical and physical processes that include the water and carbon cycles. As a physical system, the Earth's processes follow the governing laws of mass and energy conservation and increasing entropy. An impressive discovery in the turn of the twentieth century was that atmospheric masses followed Newton's second law of motion giving rise to modern meteorology and atmospheric sciences. This discovery was followed by many attempts to solve a closed set of coupled differential equations describing the motion of air parcels in the atmosphere or volumes of water in the ocean. These governing equations, also referred to as primitive equations, are nonlinear and are impossible to solve analytically given the many degrees of freedom available in the atmospheric or oceanic system. Some approximations are made to constrain the solution of the full primitive equations. Such approximations include hydrostatic assumption and geostrophic motion. The hydrostatic assumption constrains the vertical motion of the fluid, that is, the ascending and descending motions can be neglected compared to motions in the horizontal. This assumption breaks down in thunderstorms or other phenomena with sizes smaller than 10 km, approximately. The geostrophic motion occurs when there is a balance between the Coriolis force and the pressure gradient force. This balanced flow is common in latitudes away from the equator and for circulations of sizes larger than 103 km. This balanced flow is easily depicted along the contours of surface pressure in global maps. A diagram of the geostrophic flow is shown in Fig. 10. The geostrophic balance breaks down for smaller-scale motions or in cases where the friction of other forces disrupts the balance. The two assumptions together simplify the primitive equations into a quasi-geostrophic set of equations that can be more easily resolved. Quasi-geostrophic models describe the large-scale flows and are still useful for studies that require a large ensemble forecast (Miyoshi et al. 2014), decadal prediction (Kucharski et al. 2006), and data assimilation and predictability studies. These models are quite convenient as they appropriately represent the large-scale structure of the atmosphere and can be run at a relatively fast speed (about 1 year simulation in 12 min) on a one-processor personal computer.

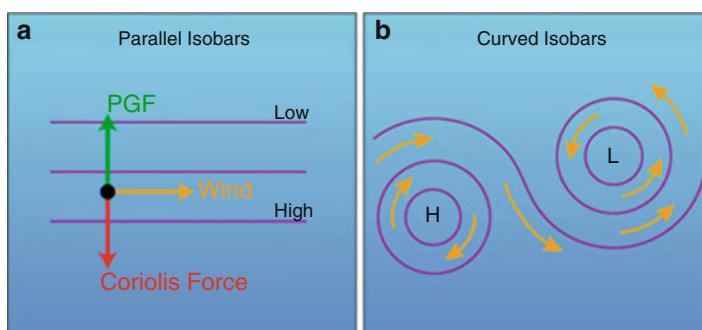


Fig. 10 Wind flow (yellow arrows) in geostrophic balance (Credit: Julius Busecke)

5.2 Global Circulation

With more computer power and algorithmic advances, the primitive equations describing the atmospheric and oceanic circulation were solved through numerical methods. In parallel with this progress, observations gathered during expeditions, experimental campaigns, and through conventional observing systems allowed for a more complete description of the global circulation. For decades, researchers have focused on improving the quantitative simulation of the observed global circulation. Atmospheric general circulation models (AGCMs) and oceanic general circulation models (OGCMs) are the standard models for simulating the global atmosphere and ocean circulations, respectively. There is a diversity of these models in terms of complexity and the way the primitive equations are discretized. All models integrate forward in time in the equations starting from an initial condition, which could be climatology or from a data assimilation scheme. One major difficulty has been to model small-scale processes, such as those occurring near the surface, in what is called the planetary boundary layer. Also, AGCMs do not explicitly simulate convection, which is an important element that drives the large-scale flow. A way to go around this problem is through “parameterization” which consists in replacing a complex process with a simple process that the discrete elements of the AGCM can compute instantly with the state variables. Both AGCM and OGCM models are stand-alone models that require prescription of the other components: AGCMs require sea surface temperatures in order to extract the fluxes that drive the boundary layer, whereas the OGCM needs momentum, thermal, and radiation fluxes to force the surface of the ocean. These approaches are called one-way interaction atmosphere-ocean models because one of the components is prescribed or used to force the other and there is no feedback between them. In the last few decades, fully coupled models have progressed to the point that they are used for seasonal predictions. For a thorough description of these models, the reader is referred to a companion chapter in Sect. 2.

Surface atmosphere (including land-ice-snow-ocean interactions) has been difficult to approach due to the many small-scale processes that govern them. As mentioned at the beginning of this subsection, these processes drive the observed low-frequency variability in the climate. Bulk formulas are used to represent the unknown fluxes from surface in situ observations (e.g., Taylor et al. 1999). Continental surfaces are more complex to model due to the heterogeneity of the surface, which includes land use and the lithosphere, for example. The exchange of water and energy occur so rapidly that in several models they are regarded as part of the atmospheric model. The relevance of this interaction is increasing as some of these exchanges determine long-term processes such as drought and climate change. Storage of water on land (e.g., soil moisture, groundwater, surface water, snow, surface ice) constitutes a memory component within the climate system similar to the heat content in the ocean. Releases of chemical compounds such as CO₂ and vegetation-climate interaction in general strongly constrain the regional climate and influence the global climate.

New sensors on satellites such as MODIS (e.g., Justice et al. 2002) that allow for measurements of vegetation indexes, surface temperature, and soil moisture

estimations complement in situ observations to understand the complex interactions occurring at the surface.

References

- C.D. Ahrens, *Meteorology Today: An Introduction to Weather, Climate, and the Environment*, 9th edn. (Brooks Cole, Belmont, 2008), 549 pp
- P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015). <https://doi.org/10.1038/nature14956>
- G.J. Boer, Predictability as a function of scale. *Atmosphere-Ocean* **41**, 203–215 (2003). <https://doi.org/10.3137/ao.410302>
- G. Boffetta, P. Giuliani, G. Paladin, A. Vulpiani, An extension of the Lyapunov analysis for the predictability problem. *J. Atmos. Sci.* **55**, 3409–3416 (1998)
- D. Bright, Ways of viewing and interpreting ensemble forecasts: applications in severe weather forecasting, in *Preprints, The AMS Short Course on Ensemble Prediction: Conveying Forecast Uncertainty*, American Meteorological Society, San Antonio (2007)
- H.E. Brooks, M. Steven Tracton, David J. Stensrud, Geoffrey DiMego and Zoltan Toth Short-Range Ensemble Forecasting: Report from a Workshop, 25–27 July 1994. *Bulletin of the American Meteorological Society*. **76**(9), 1617–1624 (1995)
- R. Buizza, A. Hollingsworth, F. Lalaurette, A. Ghelli, Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Weather Forecast*. **14**, 168–189 (1999)
- J.P. Cangialosi, J.L. Franklin, 2010 National Hurricane Center forecast verification report. NHC report (2011), 77pp. http://www.nhc.noaa.gov/verification/pdfs/Verification_2010.pdf
- Y.H. Ding, Y.J. Liu, Y.F. Song, J. Zang, From MONEX to the global monsoon: A review of monsoon system research. *Adv. Atmos. Sci.* **32**, 10–31 (2015)
- J. Du, S. Mullen, F. Sanders, Short-range ensemble forecasting of quantitative precipitation. *Mon. Weather Rev.* **125**, 2427–2459 (1997)
- ECMWF, Early medium-range forecasts of tropical cyclones. *ECMWF Newsletter*, P. White, Ed., ECMWF, No. 102 – Winter 2004/05 (2004), pp. 7–14. <https://www.ecmwf.int/en/elibrary/14623-newsletter-no102-winter-2004-05>
- R.L. Elsberry, L.E. Carr III, Consensus of dynamical tropical cyclone track forecasts-error versus spread. *Mon. Weather Rev.* **128**, 4131–4138 (2000)
- E.S. Epstein, Stochastic dynamic prediction. *Tellus* **21**, 739–759 (1969)
- R. Errico, What is an Adjoint model? *Bull. American Meteorol. Soc.* **78**, 2577–2591 (1997)
- R. Errico, T. Vukicevic, K. Raeder, Examination of the accuracy of a tangent linear model. *Tellus* **45A**, 462–497 (1993)
- J. Feng, J. Li, R. Ding, Z. Toth, Comparison of nonlinear local Lyapunov vectors and bred vectors in estimating the spatial distribution of error growth. *J. Atmos. Sci.* **75**, 1073–1087 (2017). (under review)
- T.T Fujita Tornadoes and Downbursts in the Context of Generalized Planetary Scales. *Journal of the Atmospheric Sciences*, **38**(8), 1511–1534 (1981)
- J.S. Goerss, Tropical cyclone track forecasts using an ensemble of dynamical models. *Mon. Weather Rev.* **128**, 1187–1193 (2000)
- T.M. Hamill, S.J. Colucci, Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Weather Rev.* **126**, 711–724 (1998)
- T.M. Hamill, M.J. Brennan, B. Brown, et al., NOAA's future ensemble-based Hurricane forecast products. *Bull. Amer. Meteorol. Soc.* **93**, 209–220 (2012)
- R.N. Hoffman, E. Kalnay, Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**, 100–118 (1983)
- C.O. Justice, J.R.G. Townshend, E.F. Vermote, E. Masuoka, R.E. Wolfe, N. Saleous, D.P. Roy, J.T. Morisette, An overview of MODIS land data processing and product status. *Remote Sens. Environ.* **83**, 3–15 (2002)

- E. Kalnay, *Atmospheric Modeling, Data Assimilation and Predictability* (Cambridge University Press, Cambridge, 2003)., 341pp
- O. Knill, The Lorenz system. Math 118 handout. University of Harvard (2005), 1p. http://www.math.harvard.edu/archive/118r_spring_05/handouts/lorentz.pdf
- F. Kucharski, F. Molteni, A. Bracco, Decadal interactions between the western tropical Pacific and the North Atlantic Oscillation. *Clim. Dyn.* **26**, 79–91 (2006)
- C.E. Leith, Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418 (1974)
- J.M. Lewis, Roots of ensemble forecasting. *Mon. Weather Rev.* **133**, 1865–1885 (2005)
- Q. Liu, T. Marchok, H.L. Pan, M. Bender, S. Lord, *Improvements in Hurricane Initialization and Forecasting at NCEP with Global and Regional (GFDL) Models*. Technical procedures bulletin, vol. 472, NCEP/EMC Technical Report (2002), 7p. <http://www.nws.noaa.gov/om/tpb/472.pdf>
- E.N. Lorenz, Deterministic non-periodic flow. *J. Atmos. Sci.* **20**, 130–141 (1963)
- E.N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.* **26**, 636–646 (1969)
- E.N. Lorenz, *Predictability: does the flap of a butterfly's wings in Brazil set off a Tornado in Texas?* 139th Meeting of the American Association for the Advancement of Science, Boston (1972)
- P. Lynch, The swinging spring: a simple model of atmospheric balance, in *Large-Scale Atmosphere-Ocean Dynamics*, ed. by J. Norbury, I. Roulstone, vol. II (Cambridge University Press, Cambridge, 2002a), pp. 64–108
- P. Lynch, Resonant motions of the three-dimensional elastic pendulum. *Nonlin. Mech.* **37**, 345–367 (2002b)
- L. Magnusson, J.R. Bidlot, S.T. Lang, A. Thorpe, N. Wedi, M. Yamaguchi, Evaluation of medium-range forecasts for hurricane Sandy. *Mon. Weather Rev.* **142**, 1962–1981 (2014)
- P. Markowski, Y. Richardson, *Mesoscale Meteorology in Midlatitudes* (Hoboken, Wiley, 2010), 407pp
- T. Miyoshi, K. Kondo, T. Imamura, The 10,240-member ensemble Kalman filtering with an intermediate AGCM. *Geophys. Res. Lett.* **41**, 5264–5271 (2014)
- F. Molteni, R. Buizza, T.N. Palmer, T. Petroliagis, The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996)
- M. Mu Nonlinear singular vectors and nonlinear singular values. *Science in China (D)* **43**, 375–385 (2000)
- M. Mu, W.S. Duan, B. Wang, Conditional nonlinear optimal perturbation and its applications. *Nonlinear Process. Geophys.* **10**, 493–501 (2003)
- S. Mullen, J. Du, F. Sanders, The dependence of ensemble dispersion on analysis–forecast systems: Implications to short-range ensemble forecasting of precipitation. *Mon. Weather Rev.* **127**, 1674–1686 (1999)
- R. Mureau, F. Molteni, T.N. Palmer, Ensemble prediction using dynamically-conditioned perturbations. *Q. J. R. Meteorol. Soc.* **119**, 299–323 (1993)
- I. Orlanski, A rational subdivision of scales for atmospheric processes. *Bull. Amer. Meteorol. Soc.* **56**, 527–530 (1975)
- T.N. Palmer, F. Molteni, R. Mureau, R. Buizza, P. Chapelet, J. Tribbia, Ensemble prediction, in *Proceedings ECMWF Seminar on Validation of Models over Europe*, vol. 1, (ECMWF, Shinfield Park, 1993), pp. 21–66
- C. Qian, F. Zhang, Y. Duan, Probabilistic evaluation of the prediction and dynamics of super typhoon Megi (2010). *Weather Forecast.* **28**, 1562–1577 (2013)
- C.S. Ramage Monsoon Meteorology. International geophysics series; v. 15. Academic Press, 296pp (1971)
- J. Rhome, Technical summary of the National Hurricane Center track and intensity models. NOAA Mariners Weather Log. **53**(2) (2009). http://www.vos.noaa.gov/MWL/aug_09/techsummary.shtml
- D.J. Stensrud, H.E. Brooks, J. Du, M.S. Tracton, E. Rogers, Using ensembles for short-range forecasting. *Mon. Weather Rev.* **127**, 433–446 (1999)

- I. Stewart, *Does God Play Dice? The Mathematics of Chaos* (Blackwell Publishing, Cambridge, 1989). 393pp
- X. Su, H. Yuan, Y. Zhu, Y. Luo, Y. Wang, Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. *J. Geophys. Res.-Atmos.* **119**, 7292–7310 (2014). <https://doi.org/10.1002/2014JD021733>
- P.K. Taylor, S.A. Josey, E.C. Kent Observations of Ocean Surface Fluxes: Means and Variability [C]//Seminar on Atmosphere-surface interaction. 1999:1-31, Seminar on Atmosphere-surface Interaction, 8–12 September 1997 (1999).
- Z. Toth, R. Buizza, Weather forecasting: What sets the forecast skill horizon? in *The Gap Between Weather and Climate Forecasting: Subseasonal to Seasonal Prediction*, ed. by A. Robinson, F. Vitard (Elsevier, 2018), Elsevier, Cambridge, Massachusetts, USA p. 38. in print
- Z. Toth, E. Kalnay, Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc* **74**, 2317–2330 (1993)
- Z. Toth, E. Kalnay, Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.* **125**, 3297–3319 (1997)
- Z. Toth, E. Kalnay, M.S. Tracton, R. Wobus, J. Irwin, A synoptic evaluation of the NCEP ensemble. *Weather Forecast.* **12**, 140–153 (1997)
- Z. Toth, Y. Zhu, I. Szunyogh, M. Iredell, R. Wobus, Does increased model resolution enhance predictability? in *Preprints, Symposium on Observations, Data Assimilation, and Probabilistic Prediction*, American Meteorological Society, Orlando (2002), CD-ROM, J1.9
- M.S. Tracton, E. Kalnay, Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Weather Forecast.* **8**, 379–400 (1993)
- K.E. Trenberth, D.P. Stepaniak, J.M. Caron, The global monsoon as seen through the divergent atmospheric circulation. *J. Clim.* **13**, 3969–3993 (2000)
- B. Wang, Q.H. Ding, J. Liu, Concept of global monsoon in The Global Monsoon System: Research and Forecast. pp 3–14 (2011)
- P.J. Webster, The elementary monsoon, in *Monsoons*, ed. by J.S. Fein, L. Stephens (Wiley, New York, 1987), pp. 3–32
- C.E. Williford, T.N. Krishnamurti, R.C. Torres, et al., Real-time multimodel superensemble forecasts of Atlantic tropical systems of 1999. *Mon. Weather Rev.* **131**, 1878–1894 (2003)
- WMO, *Guidelines on Ensemble Prediction Systems and Forecasting*. World Meteorological Organization (WMO) Technical report, WMO-no. 1091 (2012), 32pp
- WMO, WMO Tropical Cyclone Program Operational Plan/Manual. World Meteorological Organization (WMO) Geneva, Switzerland. (2017), <http://www.wmo.int/pages/prog/www/tcp/operational-plans.html>. Accessed 28 Oct 2017
- M. Yamaguchi, S.J. Majumdar, Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts. *Mon. Weather Rev.* **138**, 3634–3655 (2010)
- Y. Yamane, T. Hayashi, Evaluation of environmental conditions for the formation of severe local storms across the Indian subcontinent. *Geophys. Res. Lett.* **33**, L17806 (2006). <https://doi.org/10.1029/2006GL026823>



Numerical Weather Prediction Basics: Models, Numerical Methods, and Data Assimilation

Zhaoxia Pu and Eugenia Kalnay

Contents

1	Introduction: Basic Concept and Historical Overview	68
2	NWP Models and Numerical Methods	70
2.1	Basic Equations	70
2.2	Numerical Frameworks of NWP Model	73
2.3	Global and Regional Models	79
2.4	Physical Parameterizations	80
2.5	Land-Surface and Ocean Models and Coupled Numerical Models in NWP	84
3	Data Assimilation	88
3.1	Least Squares Theory	88
3.2	Assimilation Methods	90
4	Recent Developments and Challenges	94
	References	95

Abstract

Numerical weather prediction has become the most important tool for weather forecasting around the world. This chapter provides an overview of the fundamental principles of numerical weather prediction, including the numerical framework of models, numerical methods, physical parameterization, and data assimilation. Historical revolution, the recent development, and future direction are introduced and discussed.

Z. Pu (✉)

Department of Atmospheric Sciences, University of Utah, Salt Lake City, UT, USA
e-mail: Zhaoxia.Pu@utah.edu

E. Kalnay

Department of Atmospheric and Oceanic Science, University of Maryland, College Park, MD, USA
e-mail: ekalnay@atmos.umd.edu

Keywords

Numerical weather prediction · Numerical methods · Physical parameterization · Data assimilation

1 Introduction: Basic Concept and Historical Overview

Along with advances in computer technology, numerical weather prediction (NWP) has become the central component of weather forecasting. For instance, in the United States, daily weather forecasting begins with a supercomputer at the National Oceanic and Atmospheric Administration (NOAA) in Washington, DC. In Europe, the European Centre for Medium-Range Weather Forecasts (ECMWF), the world's largest numerical weather prediction center, provides advanced weather guidance for all member countries of the European Union. Around the world, most countries use NWP as key guidance for their operational weather prediction.

The basic concept of NWP is to solve a set of partial differential equations (PDEs) that govern atmospheric motion and evolution (Kalnay 2003). As will be described in Sect. 2.1.2, this set of PDEs describes basic conservation laws, including the conservation of momentum, mass, energy, and water vapor. In order to predict the atmospheric state in the future, we must integrate this set of equations forward. Therefore, NWP is an initial value problem: given the current atmospheric conditions (initial conditions), we integrate the set of PDEs to obtain future atmospheric states.

This initial value problem was defined early in the 1900s (e.g., Bjerknes 1904). Bjerknes (1904) stated that the ultimate problem in meteorology is weather forecasting (predicting future atmospheric conditions) and outlined an approach for tackling it. According to his approach, two conditions must be satisfied to successfully predict future atmospheric states:

- I. The present atmospheric conditions must be characterized as accurately as possible.
- II. The intrinsic laws, according to which the subsequent states develop out of the preceding ones, must be known.

He outlined a program that was subdivided into three partial problems or components:

1. The observation component
2. The diagnostic or analysis component
3. The prognostic component

Components 1 and 2 are related to the characterization of the present state (condition I), while component 3 is related to condition II.

In today's terminology, Bjerknes's approach would be called deterministic because the forecast is assumed to be completely determined from the present state. In addition, component 1 here refers to the global observing system, although

that was not established until the 1970s. Component 2 would now be named data assimilation, which combines observation information and short-range weather forecasts to form the best possible initial conditions. Component 3 involves solving the PDEs with numerical methods.

After Bjerknes, Richardson (1922) made the first attempt at NWP by hand. He used full primitive equations and a finite difference scheme. He divided the region of interest into cells, like the squares on a chessboard. He read the atmospheric conditions from a weather map using manual interpolation. Even though his methodology was impeccable, he obtained the forecast change in surface pressure at 145 mb in 6 h! The failure of Richardson's forecast set the NWP concept back into the theoretical world for many years. It was not until the 1950s that NWP was attempted again. Charney made the first successful numerical weather forecast with barotropic potential vorticity equations (Charney et al. 1950). Specifically, between the 1920s and 1950s, significant progress was made in the following areas:

- *Dynamic meteorology.* Atmospheric motion includes multiple temporal and spatial scales. Thus, the scale analysis method can be used to simplify the NWP equations based on the scale of motion in which one is interested in making a weather forecast. Based on scale analysis, a Rossby number is defined to validate the geostrophic flow in the midlatitude synoptic atmosphere. The quasi-geostrophic theory was derived to explain the circulations of synoptic flow in the midlatitudes. Conditions for baroclinic instability were also derived. The major benefit of dynamic meteorology to NWP is in solving the simplified equations according to the targeted scale of forecasts instead of having to deal with the full primitive equations (See details in Holton 2004 and Kalnay 2003).
- *Advances in numerical analysis.* Since analytical solutions for NWP equations do not exist, numerical methods must be used in order to archive the numerical integration in discrete grid spaces. The Courant-Friedrichs-Lowy (CFL) criterion sets a bottom-line of requirements between sizes of grid spaces and time steps in order to retain computational stability. Better understanding of nonlinear computational stability also helps in designing schemes that can be used to solve PDEs accurately and efficiently with numerical methods (See details in Kalnay 2003).
- *Atmospheric observations.* The invention of radiosonde made it possible to probe conditions in the upper atmosphere. During World War II, many countries started their radiosonde network. This has been a great help in improving the accuracy of the initial conditions required for NWP.
- *Invention of electronic computers.* The invention of electronic computers enhanced the efficiency of scientific calculation tremendously. Thus, computers help scientists implement NWP in operational practice, and operational NWP centers have been major users of supercomputers.

In 1950, when the first computer forecast was generated with the Electronic Numerical Integrator and Calculator (ENIAC), the first electronic computer in the world, the NWP became operationally practical.

Since the 1950s, continuous and rapid developments have been made in NWP:

- The maturity of global observing systems in the 1970s (<http://www.wmo.int/pages/prog/www/OSY/GOS.html>) and the use of satellite and radar observations
- Advances in data assimilation techniques (see Daley 1991; Kalnay 2003; Evensen 1994, also details in Sect. 3)
- Significant advances in computer technology and the development of numerical methods, especially the development of global spectral models, semi-Lagrangian models, and high-resolution regional models (see Robert 1982; Williamson 2007; Lynch 2008)
- Rapid development in physical parameterizations (see Arakawa 1997, 2004; Stensrud 2007)

In addition, some new developments have taken place in recent years:

- Ensemble forecasting: Instead of a single deterministic forecast, ensemble forecasting has become the mainstream method of operational NWP today (see Kalnay 2003; Bauer et al. 2015).
- Advanced data assimilation methods for satellite and radar data (e.g., Lorenc 1986; Daley 1991; Kalnay 2003; Houtekamer and Zhang 2016).
- Coupled atmosphere-ocean-land models (e.g., Hodur 1997; Chen and Dudhia 2001; Ek et al. 2003; Tolman 2014).

Today, NWP has become a multidisciplinary science in both research and operational environments. Many countries have their own NWP systems for daily operational forecasting, including both global and regional model systems for NWP. In addition, NWP computer models have become useful tools for research and education. The development of community weather and climate models, led by the US institutes, e.g., National Center for Atmospheric Research (NCAR), Penn State University, etc. makes these weather and climate research and forecasting models available to many research institutes and universities around the world.

Along with the rapid development in computer power and computer science during the last 60 years, NWP skill has been steadily improved. Figure 1 shows the forecast skill improvements at ECMWF between 1981 and 2014 (note that other centers follow a similar trend of improvement). It is clear that NWP products are quite reliable within a 5-day range and useful in a 7-day range. The improvement attained since the late 1990s from the advanced use of satellite observations is remarkable, especially in the Southern Hemisphere.

2 NWP Models and Numerical Methods

2.1 Basic Equations

There is a complete set of seven equations with seven unknowns that governs the evolution of the atmosphere: Newton's second law or conservation of momentum (three equations for the three velocity components), the continuity equation or

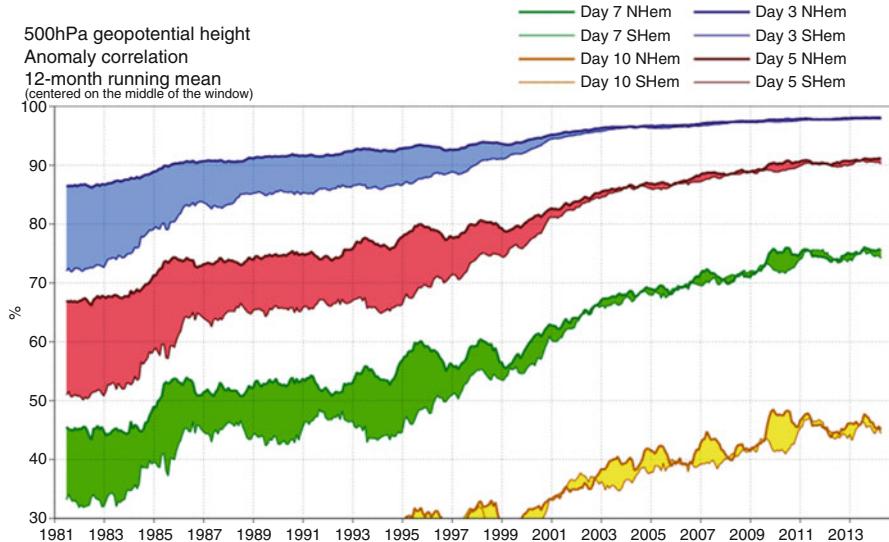


Fig. 1 Time series of the annual running mean of anomaly correlations of 500 hPa geopotential height forecasts evaluated against the operational analyses for the period of January 1981 till present. Values plotted at a particular month are averages over that month, the previous 5 months, and the following 6 months. Forecast lead times of 3, 5, 7, and 10 days are shown, for scores averaged over the northern (bold lines) and southern (thin lines) extratropics. The shading shows differences in scores between the two hemispheres at the forecast ranges indicated. (Adapted and extended from Simmons and Hollingsworth 2002. Verification follows updated WMO/CBS guidelines as specified in the Manual on the GDPFS, Volume 1, Part II, Attachment II.7, Table F, (2010 Edition – Updated in 2012); anomalies are computed with respect to ERA-Interim-based climate (Courtesy of ECMWF web site <http://www.ecmwf.int>; Also see Bauer et al. 2015)

conservation of mass, the equation of state for ideal gases, the first law of thermodynamics or conservation of energy, and a conservation equation for water mass.

In the Cartesian coordinate system, this set of equations can be written as follows:

$$\frac{d\vec{V}}{dt} = -\alpha \vec{\nabla} p - \vec{\nabla} \Phi + \vec{F} - 2\Omega \times \vec{V} \quad (1)$$

$$\frac{\partial \rho}{\partial t} = -\vec{\nabla} \cdot (\rho \vec{V}) \quad (2)$$

$$p\alpha = RT \quad (3)$$

$$Q = C_p \frac{dT}{dt} - \alpha \frac{dp}{dt} \quad (4)$$

$$\frac{\partial \rho q}{\partial t} = -\vec{\nabla} \cdot (\rho \vec{V} q) + \rho(E - C) \quad (5)$$

where $\vec{V} = (u, v, w)$ represents the velocity of air, t is arbitrary time, α is specific volume, ρ is density, p and T are pressure and temperature, Φ is geopotential height, q is the water vapor mixing ratio, Q is heating, E and C represent evaporation and condensation, respectively, R is the gas constant, and \vec{F} is the friction force.

In spherical coordinates, assume that λ and ϕ are longitude and latitude and r is the radius of the Earth:

$$u = \text{zonal}(\text{positive eastward}) = r \cos \phi \frac{d\lambda}{dt}$$

$$v = \text{meridional}(\text{positive northward}) = r \frac{d\phi}{dt} =$$

$$w = \text{vertical}(\text{positive up}) = \frac{dr}{dt}$$

Since $\vec{V} = u \vec{i} + v \vec{j} + w \vec{k}$ and $r = a + z$; $a \gg z$; $r \approx z$; $\frac{\partial}{\partial r} = \frac{\partial}{\partial z}$.

The Eqs. (1), (2), (3), (4), and (5) can be written as:

$$\frac{\partial u}{\partial t} = -u \frac{\partial u}{\partial x} - v \frac{\partial u}{\partial y} - w \frac{\partial u}{\partial z} + \frac{uv \tan \phi}{a} - \frac{uw}{a} - \frac{1}{\rho} \frac{\partial p}{\partial x} - 2\Omega(w \cos \phi - v \sin \phi) + Fr_x \quad (6)$$

$$\frac{\partial v}{\partial t} = -u \frac{\partial v}{\partial x} - v \frac{\partial v}{\partial y} - w \frac{\partial v}{\partial z} - \frac{u^2 \tan \phi}{a} - \frac{uw}{a} - \frac{1}{\rho} \frac{\partial p}{\partial y} - 2\Omega u \sin \phi + Fr_y \quad (7)$$

$$\frac{\partial w}{\partial t} = -u \frac{\partial w}{\partial x} - v \frac{\partial w}{\partial y} - w \frac{\partial w}{\partial z} - \frac{u^2 + v^2}{a} - \frac{1}{\rho} \frac{\partial p}{\partial z} + 2\Omega u \cos \phi - g + Fr_z \quad (8)$$

$$\frac{\partial T}{\partial t} = -u \frac{\partial T}{\partial x} - v \frac{\partial T}{\partial y} + (\gamma - \gamma_d)w + \frac{1}{c_p} \frac{dH}{dt} \quad (9)$$

$$\frac{\partial \rho}{\partial t} = -u \frac{\partial \rho}{\partial x} - v \frac{\partial \rho}{\partial y} - w \frac{\partial \rho}{\partial z} - \rho \left(\frac{\partial u}{\partial x} + \frac{\partial v}{\partial y} + \frac{\partial w}{\partial z} \right) \quad (10)$$

$$\frac{\partial q_v}{\partial t} = -u \frac{\partial q_v}{\partial x} - v \frac{\partial q_v}{\partial y} - w \frac{\partial q_v}{\partial z} + Q_v \quad (11)$$

$$p\alpha = RT \quad (12)$$

where Y and Y_d are the lapse rate and dry adiabatic lapse rate, respectively. Note that the terms in the equations related to the diabatic effects (H), friction (F_r), gain, or losses of water through phase changes (Q_v) must be defined within the model. Commonly, the set of equations above is called the primitive equations. These seven equations with seven unknowns represent a set of universal equations for NWP.

In reality, since the numerical models are built for various purposes that deal with different scales, it is expected that this set of equations will be simplified with some assumptions. For instance, the hydrostatic balance equation will replace the vertical motion equation if the model is designed for dealing with large scales only. The quasi-Boussinesq or anelastic approximation will be used to make the density a constant and also eliminate high-frequency waves in the solution in order to retain computational stability. In addition, the form of the equations can be changed as various coordinate systems (e.g., Cartesian vs. spherical or pressure coordinates), especially vertical coordinate systems (e.g., sigma vs. pressure vertical coordinates), are used.

To these equations we must add appropriate boundary conditions at the bottom and top of the atmosphere, then solve them using an integration process as suggested by Richardson (1922).

$$\begin{aligned}\frac{\partial \varphi}{\partial t} &= F(\varphi, t) \\ \varphi|_{t+\Delta t} &= \varphi|_t + F(\varphi|_t, t)\Delta t\end{aligned}\tag{13}$$

2.2 Numerical Frameworks of NWP Model

2.2.1 Finite Difference Equations (FDEs)

The analytical solution for the aforementioned set of equations (Eqs. 6, 7, 8, 9, 10, 11, and 12) is impossible. The equations must be solved using the discrete form with numerical methods. Therefore, finite difference equations (FDEs) can be used to find approximate solutions of the PDEs.

Using the advection equation as an example:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0\tag{14}$$

we take discrete values for x and t : $x_j = j\Delta x$ and $t_n = n\Delta t$, where Δx is the grid space and Δt is the time step of integration. The solution of the FDE is defined at the discrete points $(x_j, t_n) = (j\Delta x, n\Delta t)$:

$$U_j^n = U(j\Delta x, n\Delta t) = U(x_j, t_n)\tag{15}$$

Here we use a small u to denote the solution of the PDE (continuous) and a capital U to denote the solution of the FDE (discrete).

The FDE that is used to approximate PDE (14) can be written as follows:

$$\frac{U_j^{n+1} - U_j^n}{\Delta t} + c \frac{U_j^n - U_{j-1}^n}{\Delta x} = 0\tag{16}$$

This is called an upstream scheme if we assume $c > 0$. Note that both differences are noncentered with respect to the point $(x_j, t_n) = (j\Delta x, n\Delta t)$.

Since we employ an FDE to approximate a PDE, two fundamental conditions should be satisfied:

- (i) The FDE should be consistent with the PDE.
- (ii) For a given time $t > 0$, the solution of the FDE should converge to that of the PDE as $\Delta x \rightarrow 0$ and $\Delta t \rightarrow 0$.

In order to fulfill these two requirements, the numerical schemes used in the FDE must be as accurate as possible. We say that the FDE is consistent with the PDE if, in the limit $\Delta x \rightarrow 0$, $\Delta t \rightarrow 0$, the FDE coincides with the PDE. This requires that the solutions of the FDE be consistent approximations of the solutions of the PDE. The difference between the PDE and FDE is the discretization error or local (in space and time) truncation error.

In addition, it is very important to keep computational stability during the integration (prediction) process. Commonly, the Courant-Friedrichs-Lowy, or CFL, condition must be satisfied when specifying the time step Δt for a given grid size Δx :

$$0 \leq c \frac{\Delta t}{\Delta x} \leq 1 \quad (17)$$

where $c \frac{\Delta t}{\Delta x}$ is the so-called CFL number and c is a translation velocity. It can commonly be specified by the typical maximum wind speed at the scale of the synoptic/weather event one is dealing with in the model.

This CFL condition, however, is only a necessary condition that ensures that an FDE is computationally stable so that the solution of the FDE at a fixed time $t = n \Delta t$ remains bounded as $\Delta t \rightarrow 0$. Due to the nonlinearity of the NWP equations, computational instability can occur even when CFL conditions are satisfied. Commonly, the time step should be set smaller than what satisfies the CFL conditions. Nevertheless, some implicit numerical schemes allow using large time steps (e.g., the finite volume implicit schemes). In order to achieve accuracy and stable numerical schemes for time and spatial discretization, many advances have been made in computational mathematics, such as semi-Lagrangian schemes, finite volume schemes, and high-order Runge-Kutta schemes (see Robert 1982; Durran 1999; Lin and Rood 1996; Lin 1997).

2.2.2 Spectral Models

In addition to grid space discretization, spectral models describe the present and future states of the atmosphere using the Galerkin approach to perform space discretization with a sum of basis functions $\varphi(x)$:

$$U(x,t) = \sum_{k=1}^K A_k(t) \varphi_k(x) \quad (18)$$

The space derivatives are computed directly from the known $\frac{d\varphi(x)}{dx}$. This procedure leads to a set of ordinary differential equations (ODEs) for the coefficients $A_k(t)$. For instance, considering the one-way advection equation:

$$\frac{\partial u}{\partial t} + c \frac{\partial u}{\partial x} = 0 \quad (19)$$

we use the Fourier transform pairs:

$$\begin{aligned}\xi(u) &= U(k,t) = \int_{-\infty}^{\infty} u(x,t) \exp[-ikx] dx \\ \xi^{-1}(U) &= u(x,t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} U(k,t) \exp[ikx] dk\end{aligned}\quad (20)$$

where ξ stands for the forward Fourier transform operator and ξ^{-1} denotes the inverse Fourier transform operator. From the definitions above, we can show that:

$$\begin{aligned}\xi\left(\frac{\partial u}{\partial t}\right) &= \frac{dU(k,t)}{dt} \\ \xi\left(\frac{\partial u}{\partial x}\right) &= ikU(k,t)\end{aligned}\quad (21)$$

so that:

- The Fourier transform of a time derivative of a function is equal to the time derivative of the Fourier transform of the function.
- The Fourier transform of a space derivative of a function is equal to the Fourier transform of the function itself multiplied by ik .

We can use this and take the Fourier transform of the advection equation as follows:

$$\xi\left(\frac{\partial u}{\partial t}\right) + c \xi\left(\frac{\partial u}{\partial x}\right) = 0 \quad (22)$$

which is the same as:

$$\frac{dU}{dt} + ickU = 0 \quad (23)$$

By using Fourier transforms, we have turned a PDE into an ODE. We can then integrate the ODE forward in time to find the future value of $U(k, t)$ and then take the inverse transform to find $u(x, t)$. This is the essence of spectral methods. We convert the PDEs in real space into ODEs in wave space and then solve them.

The space discretization based on a spectral representation is extremely accurate (the space truncation errors are of “infinite” order), because the space derivatives are computed analytically, not numerically. Given this advantage, spectral models better lend themselves to longer-range forecasts than grid-point models with the same resolution. Thus, many operational global models today are spectral models (e.g., NCEP Global Forecast System). However, local forcing processes (e.g., latent heat release, differential surface heat fluxes) are sometimes discontinuous and can be represented only in physical space. In addition, when a linear combination of waves (e.g., spectral harmonics) is used to represent a large gradient or discontinuity, spurious waves can result (the Gibbs phenomenon). For higher resolutions, spectral models are computationally more demanding than grid-point models. Furthermore, spectral models do not conserve mass or energy with precision. For these reasons, only a few regional, limited-area spectral models have been developed and employed for research and operational prediction. One of the most widely used is the NCEP Regional Spectral Model (Juang and Kanamitsu 1994).

2.2.3 Grid Staggering Methods

Once the continuous PDEs are discrete in the grid mesh, all model variables are defined in the grids. Even in spectral models, since the transformations of spectral space to grids and from grids to spectral space are necessary and commonly used, model variables are defined in the grid space to some extent. The arrangement of model variables on different grid points becomes one of the considerations when designing numerical schemes for an NWP model. Instead of arranging all variables at the same grid point, many numerical models adopt a staggered grid approach.

The staggered grid combines several types of nodal points located in different geometrical positions and looks rather complex. However, the staggered grid allows for a natural and more accurate formulation of several crucial PDEs with finite differences; thus it is widely used in numerical models. Figure 2 shows an example of a staggered grid in the horizontal direction. In the vertical direction, most models have adopted a staggered grid, for instance, with the vertical velocity defined at the boundary of the layers and the prognostic variables in the center of the layer (Fig. 3). A nonstaggered vertical grid, allowing the simple implementation of higher-order differences in the vertical, would also be possible, but it would also have more computational modes present in the solution.

2.2.4 Boundary Conditions

Since numerical models commonly deal with part of the universe; boundary conditions are necessary. For instance, top and bottom boundary conditions should be given in a global atmospheric model. A regional atmospheric model requires lateral boundary conditions in addition to top and boundary conditions.

Upper boundary conditions: The altitude of a model top is usually above features of meteorological interest. Commonly, it is in the stratosphere or above. Since longer timescale processes dominate the stratosphere, climate models use a high upper boundary. Above the upper boundary, the only input of interest is from incoming

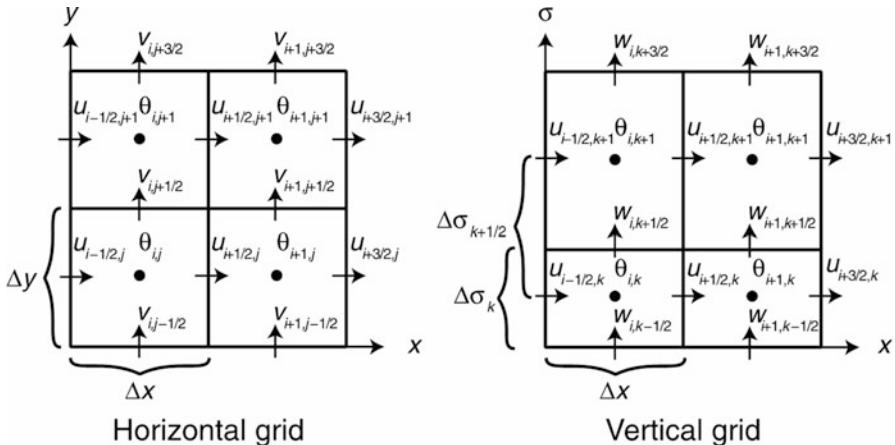


Fig. 2 The Arakawa C staggered grid method (Adapted from Skamarock et al. 2008)

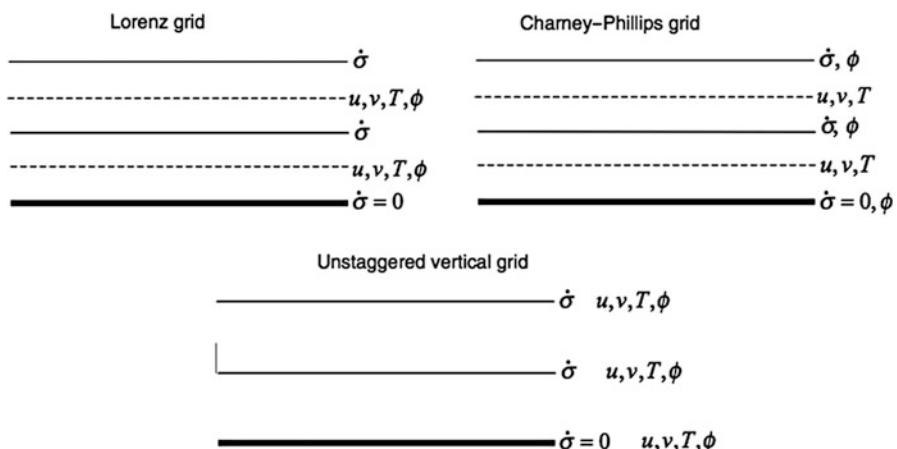


Fig. 3 Staggering in vertical grids (After Arakawa and Konor 1996)

solar radiation, which typically is parameterized. There are many ways to represent the upper boundary. For instance, a rigid lid can cap the model at some specified altitude so that energy reaching this lid is reflected downward. A free-surface method treats the model atmosphere and higher altitude as two distinct, nonmixing fluids and also reflects energy downward. Since the key issue of representing upper boundary conditions is how to handle the transfer of energy by gravity waves upward and out of the domain, an absorption/damping layer is incorporated with both the rigid-lid

and free-surface methods. This damping layer is usually placed right below the top of the model and applies a diffusion/damping operator to selected vertical levels in order to dampen upward-propagating energy. However, it must be relatively thick to mitigate the development of large vertical gradients and wave reflection issues and can dampen to a predefined reference state or to one defined by the model atmosphere. A radiative boundary condition is also used in some models to mimic the effects of wave energy propagating upward and out of the domain at the top of the model. One should usually choose an upper boundary that is sufficiently high to mitigate the issues from upper boundary conditions. See details in Durran (1999), Kalnay (2003), and Warner (2011).

Bottom boundary conditions: The bottom boundary conditions of NWP models are very complicated, as surface characteristics vary significantly. Therefore, the bottom conditions of NWP models are commonly parameterized or represented by a thermal diffusion surface model, land surface and ocean model, or surface drag schemes (as described in the next section).

Lateral boundary conditions (LBCs): The use of regional models for weather prediction has arisen from the desire to reduce model errors through an increase in horizontal resolution that cannot be afforded in a global model. Operational regional models have been embedded or “nested” into coarser-resolution hemispheric or global models since the 1970s. For instance, the current NCEP North American Mesoscale Forecast System (NAM) model is nested inside the NCEP Global Forecast System (GFS) model. The nesting of regional models requires the use of updated lateral boundary conditions obtained from the global model. Commonly, a lateral boundary condition should be satisfied if (a) it transmits incoming waves from the “host” model and provides boundary information without appreciable change in phase or amplitude, and (b) at the outflow boundaries, reflected waves do not reenter the domain of interest with appreciable amplitude.

In practice, boundary conditions are chosen pragmatically and tested numerically to check their appropriateness. Popular choices for lateral boundary conditions include both one-way and two-way nested schemes. In the one-way lateral boundary conditions, the host model, with coarser resolution, provides information about the boundary values to the nested regional model, but it is not affected by the regional model solution. In a two-way interaction in the boundary conditions, i.e., the (presumably more accurate) regional solution, in turn, also affects the global solution.

In addition, to nest a regional model inside a global model, many regional models use the nested domain technique to achieve high-resolution simulations and forecasts, with the high-resolution domain nested inside the coarser regional model domain. In this case, the lateral boundary conditions should also be addressed in either a one-way or two-way interaction. Furthermore, variable resolution models have been developed in recent years. With the use of continuously stretched horizontal coordinates, only the region of interest is solved with high resolution in a variable resolution model. It is evident that with this approach, the equations in regional high-resolution areas do not require special boundary conditions and they do influence the solutions in the regions of coarser resolution so that they can be

considered as two-way interactive nesting (see Kalnay 2003; Warner 2011). Their disadvantage is that the smallest grid size requires the use of short-time steps for the whole domain.

2.3 Global and Regional Models

Both global and regional models are used for NWP. Global models are generally used for guidance in medium-range forecasts (more than 3 days) and for climate simulations. At NCEP, for example, global models are run through 16 days every day. Because the horizontal domain of these global models is the whole Earth, they usually cannot be run at high resolution. However, with advances in computer power, the resolution of global models has increased significantly. For instance, the NCEP Global Forecast System (GFS) and ECMWF medium-range forecast model now run at nearly 16 km (T1297) horizontal resolution, about ten times the horizontal resolution of 20 years ago!

For more detailed forecasts, it is necessary to increase the resolution, and this can be done over only limited regions of interest. Regional models are used for shorter-range forecasts (typically 1–3 days) and are run with a resolution two or more times higher than that of global models. For example, in 1997 the NCEP global model was run with 28 vertical levels, with a horizontal resolution of 100 km for the first week and 200 km for the second week. The regional (Eta) model was run with a horizontal resolution of 29 km and 50 levels. Today, the NCEP regional North American Mesoscale Forecast System (NAM) model runs at a grid spacing of several kilometers (<10 km) with about 100 vertical levels. Because of their higher resolution, regional models have the ability to reproduce smaller-scale phenomena such as fronts, squall lines, and much better orographic forcing than global models. On the other hand, regional models are not “self-contained” because they require lateral boundary conditions at the borders of the horizontal domain. These boundary conditions must be as accurate as possible, because otherwise the interior solution of the regional models quickly deteriorates. Therefore, it is customary to “nest” the regional models within another model with coarser resolution, whose forecast provides the evolving boundary conditions. For this reason, regional models are used only for short-range forecasts. After a certain period, which is proportional to the size of the model, the information contained in the high-resolution initial conditions is “swept away” by the influence of the boundary conditions, and the regional model becomes merely a “magnifying glass” for the coarser model forecast in the regional domain. This can still be useful, for example, in climate simulations performed for long periods (seasons to multiple years), which therefore tend to be run at coarser resolution. A “regional climate model” can provide a more detailed version of the coarse climate simulation in a region of interest. Several other major NWP centers in Europe, including the United Kingdom, France, and Germany and in Japan, Australia, and Canada also have similar global and regional models, whose details can be obtained at their web sites.

2.4 Physical Parameterizations

2.4.1 Basic Principles

The basic equations of an NWP model (e.g., Eqs. 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12) include terms for friction, heating source, and evaporation and condensation processes. Specifically, the momentum equation has the effect of eddy fluxes of momentum; the thermodynamics equation includes radiative heating and cooling, sensible heat fluxes, and condensation and evaporation; and the water vapor equation includes condensation and evaporation, as well as moisture flux. These physical processes in numerical models represent their contribution. Thus, the model should include surface and planetary boundary layer processes, radiative transfer, and cloud microphysics in order to represent their contributions.

Atmospheric motion includes a broad spectrum of temporal and spatial scales. The timescale spans from 1 to 10^6 s and beyond, including the life cycle of a small turbulent air blob to a local storm, synoptic motions, and seasonal to interannual variations. The spatial scale ranges from 1 cm to 10,000 km, including the turbulent microscale, convective scale, mesoscale, and large scale.

Due to the use of numerical discretization methods to solve PDEs, the grid resolution of the atmospheric model is always limited. Therefore, any processes that occur on a scale smaller than the grid space cannot be explicitly represented in the numerical model, even though their contribution cannot be ignored. To give an example, here we apply Reynolds' average to the u component of the motion equation (Eq. 6).

Assume that any variable (e.g., u, v, w, T, p) can be separated into resolvable and unresolvable components, i.e., one can split all dependent variables into mean and turbulent parts, respectively. The mean is defined as an average over a grid cell, as described by Pielke (2002). For example:

$$\begin{aligned} u &= \bar{u} + u', \\ T &= \bar{T} + T', \text{ and} \\ p &= \bar{p} + p'. \end{aligned}$$

These expressions are substituted into Eqs. (6), (7), (8), (9), (10), (11), and (12); produce the expansions such the following one for Eq. (6):

$$\begin{aligned} u \frac{\partial u}{\partial x} &= (\bar{u} + u') \frac{\partial}{\partial x} (\bar{u} + u') = \bar{u} \frac{\partial \bar{u}}{\partial x} + \bar{u} \frac{\partial u'}{\partial x} + u' \frac{\partial \bar{u}}{\partial x} + u' \frac{\partial u'}{\partial x}. \\ \overline{u \frac{\partial u}{\partial x}} &= \overline{\bar{u} \frac{\partial \bar{u}}{\partial x}} + \overline{\bar{u} \frac{\partial u'}{\partial x}} + \overline{u' \frac{\partial \bar{u}}{\partial x}} + \overline{u' \frac{\partial u'}{\partial x}}. \end{aligned}$$

Since

$$\begin{aligned} \overline{a'} &= 0, \\ \bar{a} = \bar{a} &\text{ and } \overline{\bar{a}b} = \overline{\bar{a}\bar{b}} = \bar{a}\bar{b}, \text{ and} \\ \overline{\bar{a}b'} &= \overline{\bar{a}\bar{b}'} = \bar{a}\bar{b}' = 0. \end{aligned}$$

Therefore,

$$\overline{u \frac{\partial u}{\partial x}} = \bar{u} \frac{\partial \bar{u}}{\partial x} + \bar{u} \overbrace{\frac{\partial \bar{u}'}{\partial x}}^0 + \overbrace{\bar{u}' \frac{\partial \bar{u}'}{\partial x}}^0 + \overline{u' \frac{\partial u'}{\partial x}} = \bar{u} \frac{\partial \bar{u}}{\partial x} + \overline{u' \frac{\partial u'}{\partial x}}.$$

Thus,

$$\frac{\partial u}{\partial t} = -u \frac{\partial u}{\partial x} - v \frac{\partial u}{\partial y} - w \frac{\partial u}{\partial z} - \frac{1}{\rho} \frac{\partial p}{\partial x} + fv + \frac{1}{\rho} \left(\frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{yx}}{\partial y} + \frac{\partial \tau_{zx}}{\partial z} \right).$$

where

$$\begin{aligned} \tau_{zx} &= \mu \frac{\partial u}{\partial z}, \\ \frac{\partial \bar{u}}{\partial t} &= -\bar{u} \frac{\partial \bar{u}}{\partial x} - \bar{v} \frac{\partial \bar{u}}{\partial y} - \bar{w} \frac{\partial \bar{u}}{\partial z} - \frac{1}{\bar{\rho}} \frac{\partial \bar{p}}{\partial x} + f\bar{v} - \overline{u' \frac{\partial u'}{\partial x}} - \overline{v' \frac{\partial u'}{\partial y}} - \overline{w' \frac{\partial u'}{\partial z}} + \frac{1}{\bar{\rho}} \left(\frac{\partial \bar{\tau}_{xx}}{\partial x} + \frac{\partial \bar{\tau}_{yx}}{\partial y} + \frac{\partial \bar{\tau}_{zx}}{\partial z} \right). \\ \frac{\partial u'}{\partial x} + \frac{\partial v'}{\partial y} + \frac{\partial w'}{\partial z} &= 0. \\ \frac{\partial \bar{u}}{\partial t} &= -\bar{u} \frac{\partial \bar{u}}{\partial x} - \bar{v} \frac{\partial \bar{u}}{\partial y} - \bar{w} \frac{\partial \bar{u}}{\partial z} - \frac{1}{\bar{\rho}} \frac{\partial \bar{p}}{\partial x} + f\bar{v} - \overline{\bar{u}' u'} - \overline{\bar{u}' v'} - \overline{\bar{u}' w'} + \frac{1}{\bar{\rho}} \left(\frac{\partial \bar{\tau}_{xx}}{\partial x} + \frac{\partial \bar{\tau}_{yx}}{\partial y} + \frac{\partial \bar{\tau}_{zx}}{\partial z} \right). \\ T_{xx} &= -\bar{\rho} \bar{u}' u', \\ T_{yx} &= -\bar{\rho} \bar{u}' v', \\ T_{zx} &= -\bar{\rho} \bar{u}' w'. \end{aligned}$$

$$\begin{aligned} \frac{\partial \bar{u}}{\partial t} &= -\bar{u} \frac{\partial \bar{u}}{\partial x} - \bar{v} \frac{\partial \bar{u}}{\partial y} - \bar{w} \frac{\partial \bar{u}}{\partial z} - \frac{1}{\bar{\rho}} \frac{\partial \bar{p}}{\partial x} + f\bar{v} \\ &\quad + \frac{1}{\bar{\rho}} \left(\frac{\partial}{\partial x} (\tau_{xx} + T_{xx}) + \frac{\partial}{\partial y} (\tau_{yx} + T_{yx}) + \frac{\partial}{\partial z} (\tau_{zx} + T_{zx}) \right). \end{aligned} \quad (24)$$

In these equations, the first component of the five terms can be explicitly represented by model grid values. The second component of the three terms inside the parentheses cannot be explicitly resolved at model grid points, but their contributions cannot be ignored, because these subgrid-scale processes depend on and in turn affect the large-scale fields and processes that are explicitly resolved by numerical models. Therefore, parameterization schemes are then necessary in order to properly describe the impact of these subgrid-scale mechanisms on the large-scale flow of the atmosphere. In other words, the ensemble effect of the subgrid-scale processes has to be formulated in terms of the resolved grid-scale variables. Furthermore, forecast weather parameters, such as 2-m temperature, precipitation, and cloud cover, are computed by the physical parameterization of the model.

Overall, an NWP model consists of two major parts, as shown in Fig. 4: The “dynamics of the model” indicates schematically the resolved processes and the

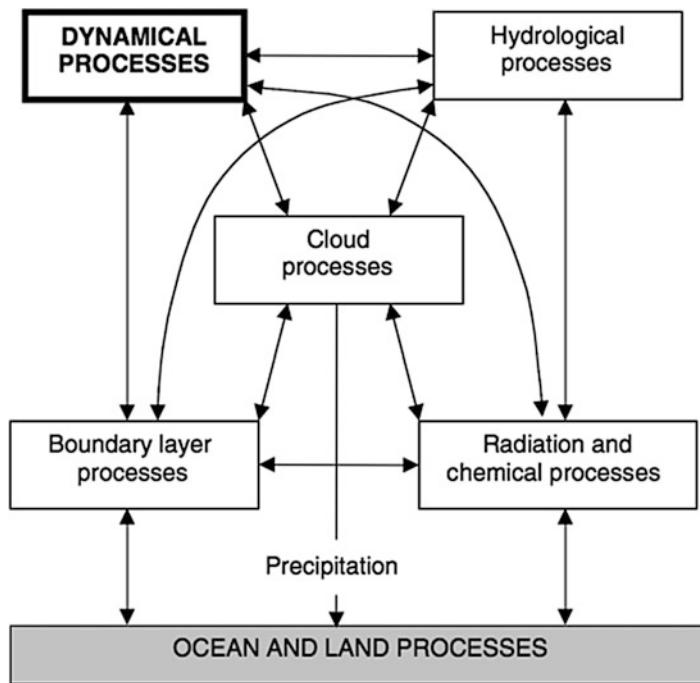


Fig. 4 Physical processes in the atmosphere and their interactions. The dynamical processes for resolvable scales, in bold, are explicitly computed by the model “dynamics.” The other subgrid-scale processes are parameterized in terms of the resolved-scale fields (Adapted from Arakawa 1997, 2004)

“model physics,” the processes that must be parameterized. Since all physical processes interact not only with the dynamics of the model but also with each other, an NWP model is numerically complicated and computationally expensive.

2.4.2 Overview of Physical Parameterizations

Physical parameterization schemes in a numerical model should be designed to (1) represent the physical processes that interact with the dynamics; and (2) explicitly calculate the contributions from the subgrid-scale processes parameterized as a function of the large-scale, resolved scales. The common parameterization schemes included in a numerical model, taking as an example those included in the current ECMWF global model, can be briefly be described as follows.

Radiation and chemical processes: The radiation scheme performs computations of shortwave and long-wave radiative fluxes using the predicted values of temperature, humidity, cloud, and monthly mean climatologies for aerosols and the main trace gases (CO_2 , O_3 , CH_4 , N_2O , etc.). The radiation parameterization describes the radiative transfer processes. Cloud-radiation interactions are usually taken into account. Since finding the solution of radiative transfer equations to obtain the fluxes is computationally very expensive, depending on the model configuration, full

radiation calculations are commonly performed on a reduced (coarser) radiation grid and/or at a reduced time frequency. The results are then interpolated back to the original grid. (See Liou 1980; Stephens 1984 for details about the radiative processes.)

Convection: The moist convection scheme represents deep (including congestus), shallow, and midlevel (elevated moist layers) convection. The distinction between deep and shallow convection is made in the convection scheme. Moist convection also resolves the entrainment process and diurnal variation of the convection. The effects of updrafts and downdrafts are also simulated (See Arakawa 2004).

Cloud microphysics and precipitation: Cloud microphysics encompasses all cloud processes that occur on the scales of the cloud droplets and the hydrometeors, including cloud droplets, raindrops, ice crystals, snow flake, rimed ice particles, graupel particles, and hail stones, rather than on the scale of the cloud itself. Microphysical parameterizations aim to represent, as thoroughly as possible, the processes described in the microphysical processes, including condensation, accretion, evaporation, ice and snow aggregation, accretion by frozen particles, vapor deposition, melting, and freezing.

In a large-scale model, clouds and large-scale precipitation are parameterized with a number of prognostic equations for cloud liquid, cloud ice, rain and snow water content, and a subgrid fractional cloud cover. The cloud scheme represents the sources and sinks of clouds and precipitation due to the major generation and destruction processes, including cloud formation by detrainment from cumulus convection, condensation, deposition, evaporation, and collection, melting, and freezing (see Houze 1993; Straka 2009).

In high-resolution models, especially regional models at the cloud-permitting scale, cloud microphysical processes are explicitly represented by the microphysics of the liquid, ice, and vapor with detailed configurations and phase changes. In models at the cloud-permitting scale, since the clouds are explicitly resolved, cumulus convection schemes can be eliminated.

Soil/surface: The surface parameterization scheme represents the surface fluxes of energy and water and the corresponding subsurface quantities. The scheme should describe different subgrid surface types for vegetation, bare soil, snow, and open water. The surface energy balance equation is also included. Soil layers are represented as well as snow mass and density. The evaporative fluxes consider the fractional contributions from snow cover, wet and dry vegetation, and bare soil. An interception layer collects water from precipitation and dewfall, and infiltration and runoff should be represented, depending on soil texture and subgrid orography. A carbon cycle may be included, and land-atmosphere exchanges of carbon dioxide may be parameterized to respond to diurnal and synoptic variations in the water and energy cycles. The soil/surface parameterization can be classified as a simple bulk scheme or as a full land-surface model.

Turbulent diffusion and planetary boundary layer scheme: The turbulent diffusion scheme represents the vertical exchange of heat, momentum, and moisture through subgrid-scale turbulence. Vertical turbulent transport is treated differently in the surface layer than it is above. For instance, in the surface layer of the ECMWF

global model as in 2013, the turbulence fluxes are computed using a first-order K-diffusion closure based on the Monin-Obukhov (MO) similarity theory. Above the surface layer, a K-diffusion turbulence closure is used everywhere, except for unstable boundary layers where an eddy-diffusivity mass-flux (EDMF) framework is applied, to represent the nonlocal boundary-layer eddy fluxes. The scheme is written in moist conserved variables (liquid static energy and total water) and predicts total water variance. A total water distribution function is used to convert from the moist conserved variables to the prognostic cloud variables (liquid/ice water content and cloud fraction) but only for the treatment of stratocumulus. Convective clouds are treated separately by the shallow convection scheme. (See Stull 1988 for details about the boundary layer and turbulence.)

Orographic drag: Limited by their resolution, NWP models cannot fully resolve the orographic features of the terrain. The effects of unresolved orography on the atmospheric flow are parameterized as a sink of momentum (drag). The turbulent diffusion scheme includes a parameterization in the lower atmosphere to represent the turbulent orographic drag induced by small-scale (<5 km) orography. In addition, in stably stratified flow, the orographic drag parameterization represents the effects of low-level blocking due to unresolved orography (blocked flow drag) and the absorption and/or reflection of vertically propagating gravity waves (gravity wave drag) on the momentum budget.

Non-orographic gravity wave drag: The non-orographic gravity wave drag parameterization accounts for the effects of unresolved non-orographic gravity waves. These waves are generated in nature by processes like deep convection, frontal disturbances, and shear zones. Propagating upward from the troposphere, the waves break in the middle atmosphere, comprising the stratosphere and mesosphere, where they deposit momentum and exert a strong drag on the flow (see Teixeira 2014).

2.5 Land-Surface and Ocean Models and Coupled Numerical Models in NWP

2.5.1 Land-Surface Models

The representation of soil, vegetation, snow, mountains, and water bodies is an integral part of the NWP system. Land can affect the weather, the magnitude of the weather effects, and the evolution of human activities. The effects of land-surface state anomalies can persist for several days, thus increasing the importance of correct initial conditions and model evolution. A refined representation of land-surface processes and their accurate initialization hold potential for further improvement of weather prediction up to the monthly range, as indicated in predictability studies.

Land-surface models (LSMs) are used to represent and parameterize land-surface processes. LSMs are important because they provide the necessary lower boundary conditions for NWP and climate models (Ek et al. 2003). They also calculate radiation flux, heat flux, and moisture flux to NWP or climate models. Such fluxes are frequently dominant driving mechanisms for mesoscale circulations. Furthermore,

land-surface processes can influence near-surface forecasting such as 2-m temperature, 10-m wind speed, boundary layer structures, and precipitation forecasting.

The energy and water budgets at the land-surface control the temperature and moisture content of the substrate and vegetation, which interact with the atmosphere. The energy conservation equation can be written for a unit mass or unit area of the surface that is experiencing energy gain or loss:

$$R_n + G + LE + H = 0 \quad (25)$$

where R_n is the net radiation at the surface, namely, the sum of downward longwave radiation, downward shortwave radiation, upward shortwave radiation (reflected by the surface, controlled by the albedo of the surface), and upward longwave radiation (surface emission). G is ground heat flux. It can also be interpreted as minus the rate of heat storage beneath the surface. E is the rate of evaporation, LE represents latent heat flux, and H is sensible heat flux. Thus, the LSM calculates sensible and latent fluxes using parameters in surface and canopy layers.

The soil temperature transfer equation is part of the LSM. For instance, a Fourier law of diffusion can be used to govern the soil heat and moisture transfer:

$$(\rho C)_s \frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \left[\lambda_T \frac{\partial T}{\partial z} \right] \quad (26)$$

where $(\rho C)_s$ is the volumetric soil heat capacity ($\text{J m}^{-3} \text{K}^{-1}$). It is a function of soil texture and soil moisture. T is soil temperature, z is the vertical coordinate (distance from the surface), t is time (s), and λ_T is the thermal conductivity.

The soil heat capacity can be estimated as a weighted sum of the heat capacity of its phases. Then

$$(\rho C)_s = (1 - \theta_s)(\rho C)_m + \theta_m(\rho C)_w$$

The subscripts m and w refer to the soil matrix and water, respectively.

The soil water movement obeys “Richard’s” equation:

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left[\lambda \frac{\partial \theta}{\partial z} \right] - \frac{\partial \gamma}{\partial z} - S_\theta \quad (27)$$

λ and γ are functions of soil texture and soil moisture. S_θ represents sources (rainfall) and sinks (evaporation).

The surface water budget equation expresses water conservation as:

$$DS = P - R - E \quad (28)$$

where DS is the change in soil water content, P is precipitation, R is runoff, and E is evaporation. Specifically, evaporation is a function of soil moisture and vegetation type, rooting depth/density, green vegetation cover, etc. Specifically:

$$E = E_{\text{dir}} + E_t + E_c + E_{\text{snow}} \quad (29)$$

E is the total surface evaporation from combined soil/vegetation; E_{dir} is direct evaporation from soil; E_t is transpiration through the plant canopy; E_c is evaporation from canopy-intercepted rainfall; and E_{snow} is sublimation from the snowpack.

With comprehensive representation of the surface energy and water balance and heating and water transport, the inclusion of LSMs in an NWP model has been shown to be beneficial for numerical prediction of near-surface atmospheric conditions as well as quantitative precipitation forecasting (e.g., Chen and Dudhia 2001).

There are many LSMs that have been developed by different centers for various purposes. For instance, the current mesoscale community Weather Research and Forecasting (WRF) model (Skamarock et al. 2008) has six options of land-surface models. The one developed by multiple US government agencies (include NCEP and the Air Force) and universities (e.g., Oregon State University) is the so-called NOAH land-surface model (Ek et al. 2003), which is also used in the NCEP NAM regional model.

2.5.2 Ocean Models

The marine component of the Earth has an important influence on the atmosphere on a range of timescales. A fully coupled model of the marine system may include surface waves, ocean, and sea ice.

It has long been known that waves affect the marine boundary layer of the atmosphere by modifying the surface roughness. Most climate models are coupled to an ocean model. Some NWP models (depending on their applications) are also coupled with ocean models. Ensemble and seasonal forecast systems use a coupled atmosphere-ocean model, which includes a simulation of the general circulation of the ocean and any associated coupled feedback processes.

Ocean wave modeling: Ocean wave modeling is used to predict the genesis and evolution of ocean waves and their associated energy. Many operational centers have developed ocean wave models in either stand-alone or coupled modes. For instance, ECMWF has developed the wave model (WAM), which is coupled to the atmospheric model or runs as a stand-alone model in the limited-area wave (LAW) configuration. Since 1998 ECMWF has been running a coupled forecasting system in which the atmospheric component of the Integrated Forecasting System (IFS) communicates with the WAM through exchange of the Charnock parameter, which determines the roughness of the sea surface (Janssen 2004). At NOAA NCEP, a wave model, WAVEWATCH III® (e.g., Tolman 2014), is developed. WAVEWATCH III solves the random phase spectral action density balance equation for wavenumber-direction spectra. The implicit assumption of this equation is that properties of medium (water depth and current) as well as the wave field itself vary on time and space scales that are much larger than the variation scales of a single wave (see details in Tolman 2014).

Ocean modeling: In addition to the ocean wave model, ocean models have also been developed to compute the time evolution of sea surface elevation, currents, salinity, and temperature. Similar to atmospheric models, ocean models are designed

to be either hydrostatic or non-hydrostatic. A baroclinic, primitive-equation ocean model contains conservation equations for mass (continuity) and momentum. Thermodynamics are also used to describe the salinity and temperature. Many numerical schemes and parameterization methods for ocean models are similar to those used in atmospheric models, although there have been many new developments in recent years. Initialization is necessary for ocean models. 3D variational data assimilation methods are used for many existing ocean models. Ocean models have been used in the operational ensemble prediction system and the seasonal forecast system. For instance, ECMWF currently uses an ocean model with 1° of resolution, initialized with the 3D variational assimilation system. Since 2013, ensemble forecasts have been coupled with the atmosphere-wave-ocean model from the start of the forecast. This is important, as it allows capture of the two-way feedback between the atmosphere and sea surface temperatures; for example, when a tropical cyclone moves slowly, it can cool the sea surface. NCEP is implementing a new Local Ensemble Transform Kalman Filter (LETKF) and a gain hybrid that combines the 3D-VAR GODAS and the LETKF, following Penny (2014).

Sea ice modeling: Sea ice is an important component of the Earth's system; it is highly reflective, altering the amount of solar radiation that is absorbed; it changes the salinity of the ocean where it forms and melts; and it acts as a barrier to the exchange of heat and momentum fluxes between the atmosphere and ocean. Current operational weather forecast systems do not commonly predict sea ice dynamically, however. In coupled forecast systems, sea ice modeling is coupled with ocean modeling to represent the dynamic and thermodynamic evolution of sea ice, mainly for seasonal and interannual prediction.

2.5.3 Coupled Numerical Models in NWP

With the rapid development of land and ocean models, many NWP models have become coupled numerical models. For instance, it has been proven that the coupling of land-surface models with atmospheric models can significantly improve quantitative precipitation forecasting (QPF). In addition, in order to accurately predict tropical cyclone formation, intensification, and dissipation, coupled ocean and atmospheric models are found to be necessary. In regional models that emphasize hurricane intensity forecasts (such as hurricane WRF or HWRF), sea spray has also been taken into account in the parameterizations. Coupled atmosphere, land, and ocean models are necessary for medium-range weather forecasting and extended-range forecasting (beyond 10 days).

Land-atmosphere coupling can be achieved by land-surface models, with input from near-surface and atmospheric conditions and characteristics of the Earth's surface to the land-surface model and output to the atmospheric model by providing the water and energy fluxes.

In coupled ocean-atmosphere modeling systems, there are up to three models in use: an atmospheric model, an ocean model, and a wave model. However, owing to computational costs, timescales of interest, and the intended application, many systems today couple two of these models, either an ocean-atmosphere-coupled modeling system or a wave-atmosphere-coupled modeling system. The coupling

takes place at the air-sea interface. For example, when all three models are used, the atmospheric model provides the surface stress to the wave model, which uses this information to derive the two-dimensional wave energy spectrum. The wave model provides the wave-induced roughness length to the atmospheric model for use in calculating surface fluxes, which also requires the SST provided by the ocean model. The wave-induced stress from the wave model along with the surface fluxes and radiation from the atmospheric model is used by the ocean model to derive the SST.

While synoptic observational data are generally sufficient to start an atmospheric model, observations are sparse below the surface of the ocean. This leads to the ocean model being run for months or years prior to the start time of any simulation or forecast in order to develop a representative three-dimensional ocean state. During the assimilation period, the ocean model is forced by surface wind stresses provided by global analyses or an atmospheric model, observed sea surface height anomalies derived from satellite data, and the observed SSTs. Wave models generally do not need to be initialized prior to the start of the coupled model simulation unless wave characteristics during the first day are important.

3 Data Assimilation

As mentioned in the beginning of the chapter, to make a forecast, we need to know the current state of the atmosphere and the Earth's surface (land and oceans). Modern numerical weather prediction makes extensive use of terrestrial and satellite observations, along with conventional observations (e.g., surface observations, radiosondes from weather stations, ships, buoys, and other components). These observations provide atmospheric, ocean, and land-surface information. Satellites now provide most data, although more observations are still important.

The weather forecasts produced by operational centers use *data assimilation* to estimate initial conditions for the forecast model from observations. The quality of forecasts depends on how well we use information received in real time from the global observing system, which consists of numerous satellite instruments, weather stations, ships, buoys, and other components. The purpose of data assimilation is to determine a best possible atmospheric state using observations and short-range forecasts. Data assimilation is typically a sequential time-stepping procedure, in which a previous model forecast is compared with newly received observations, the model state is then updated to reflect the observations, a new forecast is initiated, and so on. The update step in this process is usually referred to as the *analysis*; the short model forecast used to produce the analysis is called the *background*.

3.1 Least Squares Theory

The best estimate of the state of the atmosphere (analysis) is obtained, as indicated by Talagrand (1997), from combining prior information about the atmosphere (background or first guess) with observations, but in order to combine them

optimally, we also need *statistical information* about the errors in these “pieces of information.” A classic example of determining the best estimate of the true value of a scalar (e.g., the true temperature T_t) given two independent observations (or pieces of information), T_1 and T_2 , serves as an introduction to statistical estimation:

$$T_1 = T_t + \varepsilon_1; \quad T_2 = T_t + \varepsilon_2 \quad (30)$$

The observations have errors ε_i that we don’t know. Let $E(\)$ represent the *expected value*, i.e., the average that one would obtain if making many similar measurements. We assume that the instruments that measure T_1 and T_2 are unbiased: $E(T_1 - T_t) = E(T_2 - T_t) = 0$ or, equivalently,

$$E(\varepsilon_1) = E(\varepsilon_2) = 0 \quad (31)$$

and that we know the variances of the observational errors:

$$E\varepsilon_1^2 = \sigma_1^2 \text{ and } E\varepsilon_2^2 = \sigma_2^2 \quad (32)$$

We also assume that the errors of the two measurements are uncorrelated:

$$E(\varepsilon_1 \varepsilon_2) = 0 \quad (33)$$

Equations (31), (32), and (33) represent the statistical information we need about the actual observations. We try to estimate T_t from a linear combination of the two observations, since they represent all the information that we have about the true value of T :

$$T_a = a_1 T_1 + a_2 T_2 \quad (34)$$

The “analysis” T_a should be unbiased:

$$E(T_a) = E(T_t) \quad (35)$$

which implies

$$a_1 + a_2 = 1 \quad (36)$$

T_a will be the *best estimate* of T_t if the coefficients are chosen to minimize the mean squared error of T_a :

$$\sigma_a^2 = E[(T_a - T_t)^2] = E\left[\left(a_1(T_1 - T_t)^2 + a_2(T_2 - T_t)^2\right)\right] \quad (37)$$

subject to the constraint (37). Substituting $a_2 = 1 - a_1$, the minimization of σ_a^2 with respect to a_1 gives:

$$a_1 = \frac{\frac{1}{\sigma_1^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}}, \quad a_2 = \frac{\frac{1}{\sigma_2^2}}{\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}} \quad (38)$$

or

$$a_1 = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2}, \quad a_2 = \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} \quad (39)$$

i.e., the weights of the observations are proportional to the “precision” or accuracy of the measurements (defined as the inverse of the variances of the observational errors). Moreover, substituting the coefficients (39) in (37), we obtain a relationship between the analysis variance and the observational variances:

$$\frac{1}{\sigma_a^2} = \frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2} \quad (40)$$

i.e., if the coefficients are optimal, and the statistics of the errors are exact, then the “precision” of the analysis (defined as the inverse of the variance) is the sum of the precisions of the measurements. More importantly, Eq. (40) also indicates that the error variance of the “analysis” (σ_a^2) is smaller than the error variance of either the “background” or “observations.”

According to the least squares theory, one could achieve an analysis out of two uncertain pieces of information (background and observations) and make the analysis more accurate than either one of them alone could.

3.2 Assimilation Methods

In practice, the analysis x^a is obtained by adding the innovations to the background (model forecast or first guess) with weights W that are determined based on the estimated statistical error covariances of the forecast and the observations:

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{W}[\mathbf{y}^o - \mathbf{H}(\mathbf{x}^b)] \quad (41)$$

The innovations are here defined as the difference between the observations and the model “guess” observations, namely, $\mathbf{y}^o - \mathbf{H}(\mathbf{x}^b)$. Specifically, the background (model forecast) is interpolated to the observation location, and if they are of different type, they are converted from model variables to observed variables \mathbf{y}^o (such as satellite radiances or radar reflectivities). The first guess of the observations is therefore $\mathbf{H}(\mathbf{x}^b)$, where \mathbf{H} is the observation operator that performs the necessary interpolation and transformation from model variables to observation space.

Different analysis schemes are based on (41) but differ in the approach taken to combine the background and observations to produce the analysis. Earlier methods

such as the successive correction method were of a form similar to (41), with weights determined empirically. The weights are a function of the distance between the observation and the grid point, and the analysis is iterated several times. In optimal interpolation, the matrix of weights \mathbf{W} is determined from the minimization of the analysis errors at each grid point. Even modern advanced data assimilation methods can be interpreted in a similar way, as introduced below. See details in Kalnay (2003).

3.2.1 A Three-Dimensional Variational (3D-VAR) Data Assimilation Method

In the 3D-VAR approach, one defines a cost function proportional to the square of the distance between the analysis and both the background and the observations. The cost function is minimized directly to obtain the analysis. Lorenc (1986) showed that optimal interpolation (OI) and the 3D-VAR approaches are equivalent if the cost function is defined as:

$$J = 1/2 \left\{ [\mathbf{y}^o - H(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y}^o - H(\mathbf{x})] + 1/2 (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) \right\} \quad (42)$$

The cost function J in (42) measures the distance of a field x to the observations (the first term in the cost function) and the distance to the first guess or background \mathbf{x}^b (the second term in the cost function). The distances are scaled by the observation error covariance \mathbf{R} and by the background error covariance \mathbf{B} , respectively. The minimum of the cost function is obtained for $x = x^a$, which is defined as the “analysis.” The analysis obtained in (41) and (42) is the same if the weight matrix in (41) is given by:

$$\mathbf{W} = \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R}^{-1})^{-1} \quad (43)$$

In 3D-VAR, the minimization of (42) is performed directly, allowing for additional flexibility and a simultaneous global use of the data.

3.2.2 A Four-Dimensional Variational (4D-VAR) Data Assimilation Method

4D-VAR is an important extension of 3D-VAR that allows for observations distributed within a time interval (t_0, t_n) (e.g., Courtier and Talagrand 1987; Derber 1989; Bouttier and Rabier 1997). The cost function includes a term measuring the distance to the background *at the beginning of the interval* and a summation over time of the cost function for each observational increment computed with respect to the model integrated to the time of the observation:

$$\begin{aligned} \mathbf{J}[\mathbf{x}(t_0)] &= \frac{1}{2} [\mathbf{x}(t_0) - \mathbf{x}^b(t_0)]^T \mathbf{B}_0^{-1} [\mathbf{x}(t_0) - \mathbf{x}^b(t_0)] \\ &+ \frac{1}{2} \sum_{i=0}^N [\mathbf{H}(\mathbf{x}_i) - \mathbf{x}_i^0]^T \mathbf{R}_i^{-1} [\mathbf{H}(\mathbf{x}_i) - \mathbf{y}_i^o] \end{aligned} \quad (44)$$

The control variable (the variable with respect to which the cost function is minimized) is the *initial state* of the model with the time interval $\mathbf{x}(t_0)$, whereas the analysis at the end of the interval is given by the *model integration* from the solution $\mathbf{x}(t_n) = M_0$ [$\mathbf{x}(t_0)$]. Thus, the model is used as a *strong constraint*, i.e., the analysis solution has to satisfy the model equations. In other words, 4D-VAR seeks an initial condition such that the forecast best fits the observations within the assimilation interval. The fact that the 4D-VAR method assumes a perfect model is a disadvantage since, for example, it will give the same credence to older observations at the beginning of the interval as to newer observations at the end of the interval. Derber (1989) suggested a method of correcting for a constant model error (a constant shape within the assimilation interval).

In order to minimize the cost function, the gradient of the cost function with respect to the background and observation components can be given by:

$$\frac{\partial J_b}{\partial \mathbf{x}(t_0)} = \mathbf{B}_0^{-1} [\mathbf{x}(t_0) - \mathbf{x}^b(t_0)] \quad (45)$$

$$\left[\frac{\partial J_o}{\partial \mathbf{x}(t_0)} \right] = \sum_{i=0}^N \mathbf{L}(t_i, t_0)^T \mathbf{H}_i^T \mathbf{R}_i^{-1} [H(\mathbf{x}_i) - \mathbf{y}_i^o] \quad (46)$$

Equation (46) shows that the 4D-VAR minimization requires the computation of the gradient, i.e., computing the increments $[H(\mathbf{x}_i) - \mathbf{y}_i^o]$ at the observation times t_i during a forward integration, multiplying them by $\mathbf{H}_i^T \mathbf{R}_i^{-1}$, and integrating these weighted increments back to the initial time using the adjoint model. Since parts of the backward adjoint integration are common to several time intervals, the summation in (46) can be arranged more conveniently. Assume, for example, that the interval of assimilation is from 00 to 12 Z and that there are observations every 3 h. We compute during the forward integration the weighted negative observation increments $\bar{\mathbf{d}}_i = \mathbf{H}_i^T \mathbf{R}_i^{-1} [H(\mathbf{x}_i) - \mathbf{y}^o] = \mathbf{H}_i^T \mathbf{R}_i^{-1} \mathbf{d}_i$. The adjoint model $\mathbf{L}^T(t_i, t_{i-1}) = \mathbf{L}^T$ applied on a vector “advances” it from t_i to t_{i-1} . Then we can write (46) as:

$$\frac{\partial J_o}{\partial \mathbf{x}_o} = \bar{\mathbf{d}}_o + \mathbf{L}_0^T \{ \bar{\mathbf{d}}_1 + \mathbf{L}_1^T [\bar{\mathbf{d}}_2 + \mathbf{L}_2^T (\bar{\mathbf{d}}_3 + \mathbf{L}_3^T \bar{\mathbf{d}}_4)] \} \quad (47)$$

From (45) plus (46) or (47), we obtain the gradient of the cost function, and the minimization algorithm modifies appropriately the control variable $\mathbf{x}(t_0)$. After this change, a new forward integration and new observational increments are computed, and the process is repeated.

The most important advantage of 4D-VAR is that if we assume that (a) the model is perfect and (b) the a priori error covariance at the initial time \mathbf{B}_0 is correct, *it can be shown that the 4D-VAR analysis at the final time is identical to that of the extended Kalman filter* (Lorenc 1986; Daley 1991). This means that *implicitly 4D-VAR is able to evolve the forecast error covariance from \mathbf{B}_0 to the final time*. Unfortunately, this implicit covariance is not available at the end of the cycle and neither is the new

analysis error covariance. In other words, 4D-VAR is able to find the best linear unbiased estimation but not its error covariance, except in an approximation form.

Meanwhile, in 3D-VAR, the background term is defined statistically and is static with time. Therefore, neither 3D-VAR nor 4D-VAR can represent the follow-dependent background (model forecast) error covariance in the data assimilation.

3.2.3 Ensemble Kalman Filter

In a stochastic ensemble Kalman filter, an ensemble of K data assimilation cycles is carried out simultaneously (Evensen 1994). All the cycles assimilate the same real observations, but different sets of *random perturbations have to be added to the observations* assimilated in each member of the ensemble data assimilations. Another type of ensemble Kalman filter is known as deterministic or square-root filter (Tippett et al. 2003; Ott et al. 2004). Here there is a single data assimilation, and the new analysis perturbations are obtained from the background (forecast) ensemble perturbations using a square-root algorithm.

After completing the ensemble of analyses at time t_{i-1} , and the K forecasts $\mathbf{x}^f(t_i) = M^k [\mathbf{x}^a(t_{i-1})]$, one can obtain an estimate of the forecast error covariance from the K forecasts $\mathbf{x}_k^f(t_i)$, as:

$$\mathbf{P}^f \approx \frac{1}{K-1} \sum_{k=1}^K \left(\mathbf{x}_k^f - \bar{\mathbf{x}}^f \right) \left(\mathbf{x}_k^f - \bar{\mathbf{x}}^f \right)^T \quad (48)$$

where the overbar represents the ensemble average. This tends to underestimate the variance of the forecast errors due to nonlinearities. Thus, an inflation of covariance is usually implemented with the ensemble Kalman filters.

3.2.4 Hybrid 3D-VAR/4D-VAR and Ensemble Kalman Filter

Hamill and Snyder (2000) also suggested a hybrid between 3D-VAR and ensemble Kalman filtering, where the forecast error covariance is obtained from a linear combination of the (constant) 3D-VAR covariance $\mathbf{B}_{3D\text{-Var}}$:

$$P_l^{f(\text{hybrid})} = (1 - \alpha) P_l^f + \alpha \mathbf{B}_{3D\text{-Var}} \quad (49)$$

where α is a tunable parameter that varies from 0 for pure 3D-VAR and pure ensemble Kalman filtering from (49) to (1). Since the ensemble Kalman filtering covariance is estimated from only a limited sample of ensemble members, its rank is $K-1$, much smaller than the number of degrees of freedom of the model, so that it is rank deficient. The combination with 3D-VAR, computed from many estimated forecast errors (using, e.g., the method of Parrish and Derber 1992), may ameliorate this sampling problem and “fill out” the error covariance. In the experiments of Hamill and Snyder (2000), the best results were obtained for low values of α , between 0.1 and 0.4, indicating a good impact of the use of the ensemble-evolved forecast error covariance. They found that 25–50 ensemble members were enough to provide the benefit of ensemble Kalman filtering (but this may be different when

using a more complex model than the quasi-geostrophic model used here). Recently, both hybrid 3D-VAR and 4D-VAR and ensemble Kalman filter scheme (refer to 3dEnVar and 4dEnVar, respectively) have been implemented and run at NCEP (Wang et al. 2013; Kleist and Ide 2015) with major improvements compared with the previous 3D-VAR analysis scheme, namely the Gridpoint Statistical Interpolation (GSI) system. A different hybrid (Penny 2014) combines the Kalman gain rather than the covariances and has been tested with very good results by ECMWF (Hamrud et al. 2015).

The ensemble Kalman filtering approach has several advantages: (a) K is of the order of 10–100, so that the computational cost (compared with OI or 3D-VAR) is increased by a factor of 10–100. Although this increased cost may seem large, it is small compared to extended Kalman filtering, which requires a cost increase on the order of the number of degrees of freedom of the model. (b) Ensemble Kalman filtering does not require the development of a linear and adjoint model. (c) It does not require the linearization of the evolution of the forecast error covariance. (d) It can provide excellent initial perturbations for ensemble forecasting. Ensemble Kalman filtering appears at the present time to be one of the most promising approaches for the future (Houtekamer and Zhang 2016).

4 Recent Developments and Challenges

Along with an increase in computer power, advances in science, and demands from various applications, NWP has not only become a major forecasting tool but is also active in research and application (Bauer et al. 2015). In recent years, notable developments have been made in several areas with challenges at the same time.

First, as of today (2016), convection-permitting and cloud-resolving scale modeling have become practically feasible, along with the successful usage of large eddy simulation in developing subgrid-scale parameterizations for these models. Many national hydrometeorology centers are now running regional models in the 2–5 km grid size range and will be increasing resolution at a steady rate such that several centers may be around 1 km in next a few years. Physical parameterizations face a challenge to deal with so-called gray zones, in which the explicit model dynamics is almost capable of resolving features that were parameterized at the coarser scale. Nevertheless, one might anticipate that with increasing resolution, the need of parameterization would be gradually reduced. For radiation and cloud processes and land-surface models, this is a matter of moving current schemes toward fully explicit models. For convection, the situation is more complicated because large tropical convective clouds or organized convection occurs even at currently resolved scale (~ 15 km), while embedded small-scale convective plumes may not be resolved even at 1 km and will still require parameterization (Hong and Dudhia 2012; Bauer et al. 2015). In addition, more physical and chemical processes will be added into the NWP models. Additional physical processes will be needed to represent the coupling of atmosphere with ocean, land-surface, and sea ice models. Thus, studies on developing physical parameterization schemes, stochastically physical

parameterizations (e.g., Palmer and Williams 2008), and super-parameterizations (e.g., Khairoutdinov et al. 2005) will remain active areas.

Second, using more of the existing and new observations, and advances in data assimilation, poses more science challenges for NWP. How to better utilize the available conventional surface observations and satellite and radar observations still remain some challenges. Developing more feasible data assimilation in the “big data” era is still needed (Miyoshi et al. 2016). In addition, NWP is also limited by insufficient observational data. Beyond the maintenance of the backbone satellite and ground-based observing systems, measurements of vertical profiles of temperature, moisture, clouds, and near-surface weather, fundamental observations are missing. Moreover, coupled data assimilation will become critical for the initialization of the future coupled models (Brunet et al. 2010). The assimilation will need to include atmospheric composition (aerosols, trace gases) as well as ocean, land surface, and sea ice.

In light of the challenges in both physical parameterization and data assimilation, it is anticipated that the ensemble forecasting will remain as the mainstream of the future developments in NWP.

References

- A. Arakawa, C. S. Konor, Vertical differencing of the primitive equations based on the Charney–Phillips grid in hybrid $\sigma - p$ vertical co-ordinates. *Mon. Wea. Rev.* **124**, 511–528 (1996)
- A. Arakawa, Adjustment mechanisms in atmospheric motions. *J. Meteorol. Soc. Jpn.* **75**, 155–179 (1997)
- A. Arakawa, The cumulus parameterization problem: past, present, and future. *J. Clim.* **17**, 2493–2525 (2004)
- P. Bauer, A. Thorpe, G. Brunet, The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015)
- V. Bjerknes, Das Problem der Wettervorhersage betrachtet vom Standpunkt der Mechanik und Physik. *Meteorol. Z.* **21**, 1–7 (1904)
- F. Bouttier, F. Rabier, The operational implementation of 4D-Var. *ECMWF Newslett.* **78**, 2–5 (1997)
- G. Brunet et al., Collaboration of the weather and climate communities to advance subseasonal-to-seasonal prediction. *Bull. Am. Meteorol. Soc.* **91**, 1397–1406 (2010)
- J.G. Charney, R. Fjørtoft, J.V. Neumann, Numerical integration of the barotropic vorticity equation. *Tellus* **2**, 237–254 (1950)
- F. Chen, J. Dudhia, Coupling an advanced land-surface/hydrology model with the Penn State/NCAR MM5 modeling system. Part I: model description and implementation. *Mon. Weather Rev.* **129**, 569–585 (2001)
- P. Courtier, O. Talagrand, Variational assimilation of meteorological observations with the adjoint vorticity equations, Part II, numerical results. *Quart. J. Roy. Meteor. Soc.* **113**, 1329–1347 (1987)
- J. Derber, A variational continuous assimilation technique. *Mon. Weather Rev.* **117**, 2437–2446 (1989)
- R. Daley, *Atmospheric Data Analysis* (Cambridge University Press, Cambridge, 1991)
- D.R. Durran, *Numerical Methods for Wave Equations in Geophysical Fluid Dynamics* (Springer, New York, 1999)
- M.B. Ek, K.E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, J.D. Tarpley, Implementation of Noah land surface model advances in the National Centers for Environmental Prediction operational mesoscale Eta model. *J. Geophys. Res.* **22**, 8851 (2003)

- G. Evensen, Sequential data assimilation with a nonlinear quasigeostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**, 10143–10162 (1994)
- T.M. Hamill, C. Snyder, A hybrid ensemble Kalman filter-3D variational analysis scheme. *Mon. Weather Rev.* **128**, 2905–2919 (2000)
- M. Hamrud, M. Bonavita, L. Isaksen, EnKF and hybrid gain ensemble data assimilation. Part I: EnKF implementation. *Mon. Weather Rev.* **143**, 4847–4864 (2015)
- S.Y. Hong, Dudhia, Next-generation numerical weather prediction: bridging parameterization, explicit clouds, and large eddies. *Bull. Am. Meteorol. Soc.* **93**, ES6–ES9 (2012)
- R.M. Hodur, The naval research laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). *Mon. Weather Rev.* **125**, 1414–1430 (1997)
- J. Holton, An introduction to dynamic meteorology. Fourth edition. (Elsevier Academic Press, 2004)
- P.L. Houtekamer, F. Zhang, Review of the ensemble Kalman filter for atmospheric data assimilation. *Mon. Weather Rev.* **144**, 4489–4452 (2016)
- R.A. Houze Jr., *Cloud Dynamics* (Academic, London, 1993)
- P.A.E.M. Janssen, *The Interaction of Ocean Waves and Wind* (Cambridge University Press, Cambridge, UK, 2004)
- H.M. Juang, M. Kanamitsu, The NMC nested regional spectral model. *Mon. Weather Rev.* **122**, 3–26 (1994)
- E. Kalnay, *Atmospheric Modeling, Data Assimilation, and Predictability* (Cambridge University Press, 2003)
- M.F. Khairoutdinov, D.A. Randall, C. DeMott, Simulations of the atmospheric general circulation using a cloud-resolving model as a super-parameterization of physical processes. *J. Atmos. Sci.* **62**, 2136–2154 (2005)
- D.T. Kleist, K. Ide, An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS. Part II: 4DEnVar and hybrid variants. *Mon. Weather Rev.* **143**, 452–470 (2015)
- S.-J. Lin, A finite-volume integration method for computing pressure gradient forces in general vertical coordinates. *Q. J. R. Meteorol. Soc.* **13**, 1749–1762 (1997)
- S.J. Lin, R.B. Rood, Multidimensional flux-form semi-Lagrangian transport scheme. *Mon. Weather Rev.* **124**, 2046–2070 (1996)
- K.-N. Liou, *An Introduction to Atmospheric Radiation* (Academic, London, 1980)
- A.C. Lorenc, Analysis methods for numerical weather prediction. *Q. J. R. Meteorol. Soc.* **112**, 1177–1194 (1986)
- P. Lynch, The origins of computer weather prediction and climate modeling. *J. Comput. Phys.* **227**, 3431–3444 (2008)
- T. Miyoshi et al., “Big Data Assimilation” revolutionizing severe weather prediction. *Bull. Am. Meteorol. Soc.* **97**, 1347–1354 (2016)
- E. Ott et al., A local ensemble Kalman filter for atmospheric data assimilation. *Tellus* **56A**, 415–428 (2004)
- D.F. Parrish, J.C. Derber, The National Meteorological Center's spectral statistical interpolation analysis system. *Mon. Weather Rev.* **120**, 1747–1763 (1992)
- S.G. Penny, The hybrid local ensemble transform Kalman filter. *Mon. Weather Rev.* **142**, 2139–2149 (2014)
- T.N. Palmer, P.D. Williams, Introduction: stochastic physics and climate modelling. *Phil. Trans. R. Soc. A* **366**, 2421–2427 (2008)
- R.A. Pielke Sr., *Mesoscale meteorological modelling*. Second edition (Academic Press, 2002)
- L.F. Richardson, *Weather Prediction by Numerical Process* (Cambridge University Press, Cambridge, UK, 1922)
- A.J. Robert, A semi-Lagrangian and semi-implicit numerical integration scheme for the primitive meteorological equations. *J. Meteorol. Soc. Jpn.* **60**, 319–324 (1982)
- A.J. Simmons, A. Hollingsworth, Some aspects of the improvement in skill of numerical weather prediction. *Q. J. R. Meteorol. Soc.* **128**, 647–677 (2002)
- W.C. Skamarock, J.B. Klemp, J. Dudhia, D.O. Gill, M. Barker, K.G. Duda, X.Y. Huang, W. Wang, J.G. Powers, A description of the advanced research WRF version 3. NCAR Tech. Note, NCAR/TN-475+STR, 113 pp. (2008)

- D.J. Stensrud, *Parameterization Schemes: Keys to Understanding Numerical Weather Prediction Models* (Cambridge University Press, Cambridge, UK, 2007)
- G.L. Stephens, The parameterization of radiation for numerical weather prediction and climate models. *Mon. Weather Rev.* **112**, 826–867 (1984)
- J.M. Straka, *Cloud and Precipitation Microphysics: Principles and Parameterization* (Cambridge University Press, Cambridge, UK, 2009)
- R.B. Stull, *An Introduction to Boundary Layer Meteorology* (Kluwer Academic Publishers, Dordrecht, 1988)
- O. Talagrand, Assimilation of observations, an introduction. *J. Met. Soc. Jpn. Spec. Issue* **75**(1B), 191–209 (1997)
- M. Teixeira, The physics of orographic gravity wave drag. *Front. Phys.* **2**, 43 (2014)
- M. Tiedtke, The general problem of parameterization. ECMWF Lecture Note (1984), <http://www.ecmwf.int/en/learning/education-material/introductory-lectures-nwp>
- M.K. Tippett, J.L. Anderson, C.H. Bishop, T.M. Hamill, J.S. Whitaker, Ensemble square-root filters. *Mon. Weather Rev.* **131**, 1485–1490 (2003)
- H.L. Tolman, User manual and system documentation of WAVEWATCH III version 4.18. NOAA/NWS/NCEP/MMAB Technical Note 316, 194 pp. (2014)
- X. Wang, D. Parrish, D. Kleist, J. Whitaker, GSI 3DVarbased ensemble-variational hybrid data assimilation for NCEP global forecast system: single-resolution experiments. *Mon. Weather Rev.* **141**, 4098–4117 (2013)
- T. Warner, *Numerical Weather and Climate Prediction* (Cambridge Press, Cambridge, UK, 2011)
- D.L. Williamson, The evolution of dynamical cores for global atmospheric models. *J. Meteorol. Soc. Jpn. B* **85**, 241–269 (2007)



Ensemble Methods for Meteorological Predictions

Jun Du, Judith Berner, Roberto Buizza, Martin Charron,
Peter Houtekamer, Dingchen Hou, Isidora Jankov, Mu Mu,
Xuguang Wang, Mozheng Wei, and Huiling Yuan

Contents

1	Introduction	101
2	Methods to Address Uncertainty in Initial Conditions	103
3	Methods to Address Uncertainty in Model	118
4	Virtual Ensembles	134
5	Ensemble Size	139
6	Ending Remarks	141
	References	142

J. Du (✉) · D. Hou

Environmental Modeling Center/National Centers for Environmental Prediction (NCEP), NOAA, College Park, MD, USA

e-mail: jun.du@noaa.gov; dingchen.hou@noaa.gov

J. Berner

National Centers for Atmospheric Research, Boulder, CO, USA

e-mail: judithberner@gmail.com

R. Buizza

European Centre for Medium Range Weather Forecasts, Reading, UK

e-mail: roberto.buizza@ecmwf.int

M. Charron · P. Houtekamer

Canadian Meteorological Center, Environmental Canada, Montreal, Canada

e-mail: Martin.Charron@ec.gc.ca; Peter.Houtekamer@ec.gc.ca

I. Jankov

Cooperative Institute for Research in the Atmosphere, Earth System Research Lab (ESRL)/NOAA, Boulder, CO, USA

e-mail: isidora.jankov@noaa.gov

M. Mu

Institute of Atmospheric Sciences, Fudan University, Shanghai, China

e-mail: mumu@fudan.edu.cn

Abstract

Since the atmospheric system is a nonlinear chaotic system, its numerical prediction is bound by a predictability limit due to imperfect initial conditions and models. Ensemble forecasting is a dynamical approach to quantify the predictability of weather, climate, and water forecasts. This chapter introduces various methods to create an ensemble of forecasts based on three aspects: perturbing initial conditions (IC), perturbing a model, and building a virtual ensemble. For generating IC perturbations, methods include (1) random, (2) time-lagged, (3) bred vector, (4) ensemble transform (ET), (5) singular vector (SV), (6) conditional nonlinear optimal perturbation (CNOP), (7) ensemble transform Kalman filter (ETKF), (8) ensemble Kalman filter (EnKF), and (9) perturbations in boundaries including land surface and topography. For generating model perturbations, methods include (1) multi-model and multi-physics, (2) stochastically perturbed parametrization tendency (SPPT), (3) stochastically kinetic energy backscatter (SKEB), (4) convection triggering, (5) stochastic boundary-layer humidity (SHUM), (6) stochastic total tendency perturbation (STTP), and (7) vorticity confinement. A method to create a spatially correlated random pattern (mask) needed by SPPT, SKEB, etc. is introduced based on the Markov process; a factor separation method is introduced to estimate the relative impact of various physics schemes and their interactions. A method of perturbing a dynamic core to create an ensemble is also mentioned. Quantitative forecast uncertainty information and ensemble products can also be generated from “virtual ensembles” based on existing deterministic forecasts through at least five different approaches including (1) time-lagged, (2) poor-man’s, (3) hybrid, (4) neighborhood, and (5) analog ensembles. Generally speaking, the selection of perturbation methods in constructing an EPS is more important for smaller-scale and shorter-range forecasts and less critical for larger-scale and longer-range forecasts. Finally, the frequently asked question about the trade-off between ensemble size and model resolution is discussed. By introducing these methods, we hope to help readers who are interested in ensemble forecasting but not familiar with these approaches to build their own EPS or produce ensemble products as well as for students to learn the subject of ensemble forecasting.

X. Wang

School of Meteorology, The University of Oklahoma, Norman, OK, USA

e-mail: xuguang.wang@ou.edu

M. Wei

Oceanography Division, Navy Research Laboratory, Stennis Space Center, MS, USA

e-mail: mzw800@gmail.com

H. Yuan

School of Atmospheric Sciences and Key Laboratory of Mesoscale Severe Weather, Ministry of Education, Nanjing University, Nanjing, China

e-mail: yuanhl@nju.edu.cn

Keywords

Ensemble forecasting · Initial condition perturbation · Boundary perturbation · Model physics and dynamic core perturbations · Virtual ensembles · Ensemble size

1 Introduction

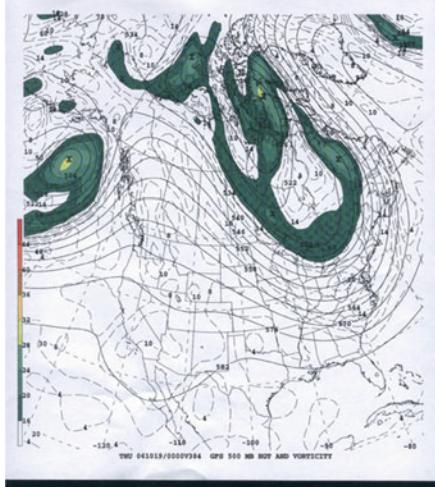
The value of science is in predicting the future. The process of numerical weather prediction (NWP) has three basic components: data collection (observation), assimilation of observed data into initial conditions used by a numerical weather prediction model, and model integration to project an initial state into the future. Intrinsic errors are introduced at each of these three steps: for example, instrumental and human error in collecting and interpreting observations, imperfect methods for data retrieval and assimilation, imperfect model physics and numerical methods. Furthermore, there are inevitably inconsistencies in the adaptation of model forecasts to real-world applications, whether by variations in post-processing methods or by a human diversity in interpretation. All these errors are intrinsic, largely unavoidable, and perhaps even unknown to us in real-world operations.

Due to its nonlinear nature, a numerical prediction model of weather, climate, or water is chaotic, i.e., a tiny difference in the initial state can be amplified into significantly larger differences in a future state (Lorenz 1963, 1965, 1993; Thompson 1957). These differences could be as large as those between two randomly picked fields from climatology, in which case lose all value. Therefore, any prediction of weather, climate, or water events has uncertainty and limits to predictability. For example, Fig. 1 shows two NCEP (National Centers for Environmental Prediction) operational GFS (global forecast system) model medium-range (16-day) forecasts, initialized only 6 h apart in time. These two numerical forecasts predicted two very different large-scale flow patterns at the 500-hPa level: one places a strong trough over the east coast, while the other places it over the Western United States (more than 3000 km apart). Similarly impactful discrepancies are not uncommon in real-time operations in all time scales, including the short range, especially during major high-impact weather events (Wang et al. 2011). For further reading about the predictability of weather and climate, readers are referred to Palmer and Hagedorn (2006).

Therefore, quantifiable information about uncertainty and predictability is an important aspect of a weather forecast. Besides the prediction of an event itself, the uncertainty and predictability associated with the prediction also need to be estimated (“a prediction of predictability”). As Socrates once taught us *a wise man is he who knows he knows not*. We need to humbly admit that unless forecast uncertainty is quantified, a forecast is incomplete. Ensemble forecasting is a dynamical and flow-dependent approach of quantifying this forecast uncertainty (errors of the day) and provides a basis to communicate forecast confidence to end users who can then be better prepared. It has become an increasingly important aspect of a forecast

Two consecutive NCEP operational Global Forecasting System (GFS) 16-day 500mb HGT/VORT forecasts (with only 6hr-hour apart in initial conditions)!

(A) 00z, Oct. 3 – 00z, Oct. 19, 2006



(B) 06z, Oct. 3 – 06z, Oct. 19, 2006

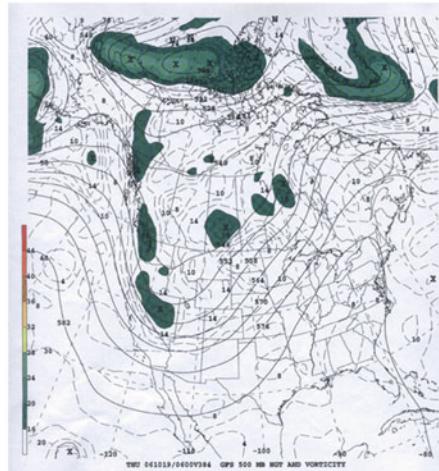


Fig. 1 Two consecutive NCEP global forecast system (GFS) 16-day 500-hPa HGT/VORT forecasts that are 6 h apart in initiation time

(National Research Council 2006; Wang et al. 2011). Information about uncertainty can be used in many ways. For example, it can be integrated into the decision-making process based on economic values (Richardson 2000; Du and Deng 2010).

Lewis (2005) discusses the roots of ensemble forecasting. After Lorenz (1963, 1965) discovered the chaotic nature of atmospheric behavior in the 1960s, some pioneering scientists started to seriously consider stochastic approaches to predicting weather and climate (Epstein 1969; Leith 1974). Since we do not exactly know a single ground truth but have many equally plausible initial conditions (IC) or physics options, a scientific and complete description of IC and model physics is best done with a probabilistic distribution and in stochastic fashion within a reasonable range of error (uncertainty). As a result, there might be a number of possible realizations for each forecast derived from a highly nonlinear chaotic numerical model (Lorenz 1993). In other words, *a complete forecast* of a particular point value should be not as a mere single deterministic value but as a probabilistic distribution with forecast uncertainty or confidence explicitly expressed. This is the basic concept of *ensemble forecasting* which can be schematically described by Fig. 2. The primary mission of an ensemble prediction system (EPS) is, therefore, to encompass truth by an ensemble of solutions with a good spread-skill relationship (i.e., ensemble spread quantifiably reflects the error of the ensemble mean forecast, and the derived probabilistic forecasts are statistically reliable, Du et al. 2014). Besides this primary goal, an important by-product of an EPS is to improve the performance and reduce the uncertainty of a deterministic forecast through various approaches such as

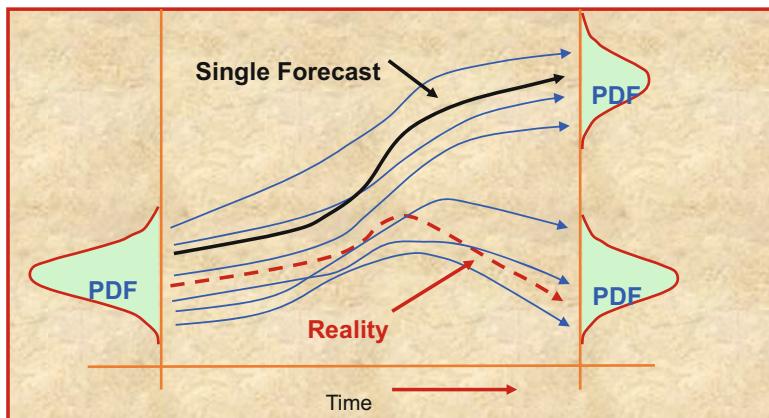


Fig. 2 A schematic concept of “ensemble prediction”

ensemble averaging, ensemble-based data assimilation, and targeted observations (Du et al. 2014). By the above definition of its primary mission, ensemble forecasting is most valuable when there is a large forecast uncertainty, and it has minimal value when the weather is quiescent and highly predictable relative to a single deterministic forecast. By the way, ensemble forecasting mainly deals with random error but not systematic error in a forecast; a good quality model and initial conditions are the basis for the success of an EPS (Wang et al. 2018).

In the early days, an EPS such as the Monte Carlo approach proposed by Leith (1974) mainly addressed uncertainty in the IC, but this definition has now been greatly expanded by addressing uncertainties in all components of a modeling system including atmospheric initial states, model physics and numerical methods, lower boundary forcing such as land or sea surface states, lateral boundary conditions (LBC, for regional model), and other coupling mechanisms like air-sea interaction. Some methodologies used to construct an EPS or produce an ensemble of forecasts are described in this chapter, grouped into three categories: IC uncertainty, model uncertainty, and virtual ensembles.

2 Methods to Address Uncertainty in Initial Conditions

Some IC perturbation methods are introduced in this section. Generally speaking, IC methods have evolved through three generations: the first generation is a Monte Carlo type, i.e., a random perturbation; the second generation is related to dynamical unstable modes such as breeding vector, SV, and CNOP; and the third generation is directly connected with analysis or observation errors such as ETKF, EnKF, etc.

1. *Random perturbation.* A perturbation is randomly generated based on some kind of error statistics (usually a normal distribution) representing the average error range in an analysis. These error statistics such as mean and standard deviation

can be derived from the differences between two commonly used operational analyses such as the NCEP and ECMWF global analyses over a long time period (Errico and Baumhefner 1987; Mullen and Baumhefner 1994). Therefore, these error statistics are not purely random but normally exhibit spatial and temporal patterns which are consistent with our meteorological knowledge: e.g., the error is larger over less-observed regions such as the oceans than over densely observed areas such as the Continental United States, larger in active weather regions such as the two storm tracks over the Pacific and Atlantic oceans, and larger when the natural variability is larger: e.g., larger in the middle and high latitudes than in the tropics, larger in winter than in summer, and so on. Although this type of perturbation represents the average error of analysis well, it does not truly reflect “errors of the day,” and more critically it lacks a spatially correlated dynamical structure coupled with daily weather systems. As a result, the perturbation growth with forecast time is not ideal, especially in the short range, which often leads to less spread among ensemble members. The random perturbation method was used in early ensemble studies such as Mullen and Du (1994) and Du et al. (1997). The multi-analysis (without any perturbations added) approach can be viewed as another type of random perturbation method, such as was used in the early prototype NCEP SREF (Stensrud et al. 1999) and the northwest US regional ensemble system at the University of Washington (Grimit and Mass 2002). Since the number of available analyses is always limited, a purely multi-analysis-based ensemble normally has a small ensemble size. Note that an EnKF-based data assimilation-produced multiple analysis is not regarded as an ad hoc multi-analysis approach but is regarded as a perturbing-observation method which will be discussed separately as method 7 in this section. Through EnKF, the limitation of small ensemble size for a multi-analysis-based EPS can be overcome.

2. *Scaled time-lagged perturbation (STL pert).* This method directly applies the forecast errors from a few of the most recent past cycles (e.g., 12, 9, 6, and 3 h ago) as IC perturbations for the current cycle’s ensemble run, e.g., using 06z, 09z, 12z, and 15z cycles’ 12-, 9-, 6-, and 3-h forecast errors as IC perturbations for the 18z cycle of the ensemble run. In this approach the magnitude of the perturbations obviously depends on a forecast’s age since forecast error normally increases with lead time. To have a similar magnitude in all perturbations derived from different-aged forecasts, these past forecast errors are first scaled to the same magnitude and then either added to or subtracted from the current cycle’s control analysis (as a pair of perturbed members) to create multiple perturbed ICs to initialize an EPS (Ebisusaki and Kalnay 1991; Kalnay 2003), as illustrated by Eq. (1):

$$\text{TL pert} = \text{scaling factor} \times (\text{past forecast} - \text{current analysis}) \quad (1)$$

The practice of “pairing” (adding and subtracting) ensures that the perturbations are centered on the control IC in addition to doubling the ensemble membership.

Time-lagged IC perturbation has dynamically growing structures associated with developing weather systems, which is beneficial as the ensemble spread grows with forecast lead time. Another advantage of this method is that it can carry over information from past initial conditions. A limitation is that it cannot create an ensemble with a very large membership since the number of “usable” old forecasts is limited. An example of this approach is used in the 1998 Storm and Mesoscale Ensemble Experiment (SAMEX, Hou et al. 2001).

3. Bred growing mode (BGM). The BGM method is also called *breeding* or *bred vector* (BV). It evolved from the scaled time-lagged method (Eq. 1). It differs from STL in that it uses the difference between a pair of past (say, 6 or 12 h ago) concurrent forecasts valid at the current model initial time rather than a past forecast error to calculate IC perturbations (Eq. 2). The difference is then scaled and added to or subtracted from the current cycle’s control analysis (Toth and Kalnay 1993, 1997). In this way, one can overcome the membership limitation in the scaled time-lagged method to create as many members as desired as long as one has enough initial perturbation seeds in a cold start. Since there is no bred vector available yet in a cold start run, another substitute perturbation (such as random, time-lagged, or perturbations borrowed from other available EPS) is needed to start an initial ensemble run.

$$\mathbf{BV} = \text{scaling factor} \times (\text{past forecast 1} - \text{past forecast 2}) \quad (2)$$

Kalnay (2003) proved that BV is a nonlinear extension of a Lyapunov vector with fast growing dynamical structure. Experience (e.g., the NCEP SREF, Du et al. 2004) shows that a bred vector becomes mature in structure and leads to a healthy spread growth after being cycled for about 2–3 days after its cold start. Toth and Kalnay pointed out that the spatial structure of a mature bred vector is not sensitive to the scaling period (But our experiences with both the NCEP SREF and GEFS show that the quality of an IC perturbation is actually strongly dependent on scaling period, e.g., a 12-h forecast-based BV perturbation worked much better than the 6-h forecast-based one, as the latter has too much small-scale noisy structure in space.) and the norm selected and that the bred vector well reflects the analysis error introduced during a data assimilation cycle. If the pair of past concurrent forecasts used in the BV calculation have different physics schemes in model integration (e.g., two different convective schemes), the resulting BV perturbation will automatically contain physics uncertainty information. By taking advantage of this, the NCEP SREF (Du and Tracton 2001) was particularly designed to use a pair of members with different physics to calculate its BV. The advantage of using two members with different physics has also been examined by Chen et al. (2005). On other hand, one might argue that BV is a “looking-backward” method because the difference between two past forecasts really reflects the fast growing modes that occurred in the past but is not guaranteed to grow fast in the future, although Kalnay argues that BV has the ability to indicate “the future,” especially the coming of major events or a

regime transition (Kalnay 2007). Another weakness is that bred perturbations from different ensemble members are not orthogonal but are correlated to each other, resulting in less independent information contained in an ensemble (Wang and Bishop 2003; Martin et al. 2007). As a result, the magnitude of ensemble spread growth is closely related to the initial size of a bred vector. There are various ways proposed to increase the orthogonality of a bred vector. One of them is the ensemble transform with rescaling (ETR) method (see method 4). Another effort to improve the classical breeding method is geometric breeding, which controls the spatial correlation of bred vectors among members, making them less correlated to each other (Martin et al. 2007). Geometric breeding shows a better spread-skill relationship than classical breeding.

Given its effectiveness, simplicity, and small computational cost, the breeding method is popular and widely used in many numerical weather prediction centers around the world, including NCEP for both its global and regional EPSs (Tracton and Kalnay 1993; Du and Tracton 2001).

4. *Ensemble transform with rescaling (ETR)*. To orthogonize bred vectors, the ETR technique is used to make bred vectors more orthogonal to each other by applying a simplex transformation matrix to transform forecast-based perturbations to analysis perturbations (Wei et al. 2008). The ETR method is based on an improved version of the ensemble transfer (ET) technique originally developed for a target observation study (Bishop and Toth 1999). The technique is described below. Let

$$\mathbf{Z}^f = \frac{1}{\sqrt{k-1}} [\mathbf{z}_1^f, \mathbf{z}_2^f, \dots, \mathbf{z}_k^f], \mathbf{Z}^a = \frac{1}{\sqrt{k-1}} [\mathbf{z}_1^a, \mathbf{z}_2^a, \dots, \mathbf{z}_k^a], \quad (3)$$

where the n -dimensional state vectors $\mathbf{z}_i^f = \mathbf{x}_i^f - \mathbf{x}^f$ and $\mathbf{z}_i^a = \mathbf{x}_i^a - \mathbf{x}^a$ ($i = 1, 2, \dots, k$) are k ensemble forecast and analysis perturbations for all the model variables, respectively. \mathbf{x}^f is the mean of k ensemble forecasts from the forecast model and \mathbf{x}^a is the analysis from the data assimilation (DA) system. In the ensemble representation, the $n \times n$ forecast and analysis covariance matrices are approximated, respectively, as

$$\mathbf{P}^f = \mathbf{Z}^f \mathbf{Z}^{fT} \quad \text{and} \quad \mathbf{P}^a = \mathbf{Z}^a \mathbf{Z}^{aT}, \quad (4)$$

where the superscript T indicates the matrix transpose. For a given set of forecast perturbations \mathbf{Z}^f at time t , the analysis perturbations \mathbf{Z}^a are obtained through an ensemble transformation \mathbf{T} such that

$$\mathbf{Z}^a = \mathbf{Z}^f \mathbf{T}. \quad (5)$$

In the ETR method, the best possible initial analysis error variance from the DA is used to restrain and construct the above transformation matrix. If \mathbf{P}^a is the diagonal matrix with diagonal values being the analysis error variances obtained from the

operational DA system, the transformation matrix \mathbf{T} can be constructed as follows. For any ensemble forecast system, the forecast perturbations \mathbf{Z}^f can be constructed as Eq. (1). One can solve the following eigenvalue problem

$$\mathbf{Z}^{f^T} \mathbf{P}^{a^{-1}} \mathbf{Z}^f = \mathbf{C} \boldsymbol{\Gamma} \mathbf{C}^{-1}, \quad (6)$$

where \mathbf{C} contains the column orthonormal eigenvectors (\mathbf{c}_i) of $\mathbf{Z}^{f^T} \mathbf{P}^{a^{-1}} \mathbf{Z}^f$ or equivalently the singular vectors of $\mathbf{P}^{a^{-1/2}} \mathbf{Z}^f$ and $\boldsymbol{\Gamma}$ is a diagonal matrix containing the associated eigenvalues (λ_i) with the magnitude in decreasing order, that is, $\mathbf{C} = [\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_k]$, $\mathbf{C}^T \mathbf{C} = \mathbf{I}$ and $\boldsymbol{\Gamma} = \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_k)$. The new analysis perturbations can be constructed through transformation:

$$\mathbf{Z}^a = \mathbf{Z}^f \mathbf{C} \boldsymbol{\Gamma}^{-1/2} \mathbf{C}^T \quad (7)$$

To make sure the initial spread distribution is similar to the analysis error variance, a final rescaling step is carried out in the ETR. This regional rescaling process is similar to that used in the breeding method, i.e., the individual initial perturbations from (7) are rescaled by the analysis error variance using

$$\mathbf{y}_m^a(i,j,l) = \alpha(i,j,l) \mathbf{z}_m^a(i,j,l), \quad (8)$$

where i, j, l are indices for the horizontal and vertical directions in grid point space and $m = 1, 2, \dots, k$ is the index for the ensemble member. α is the rescaling factor derived from the analysis error variance (\mathbf{P}^a) and the grid point values of analysis perturbations. It is defined as the ratio of the square root of kinetic energy from \mathbf{P}^a and the square root of the kinetic energy of analysis perturbations at each grid point. The reader is referred to Wei et al. (2008) for more mathematical details.

ETR-based perturbations are centered around a control analysis and span a subspace that has a maximum number of degrees of freedom. Thus, the variance is maintained in as many directions as possible within the ensemble subspace. Theoretically the covariance constructed from the newly generated initial perturbations is approximately consistent with the analysis covariance from the DA, if the number of ensemble members is large. Wei et al. (2008) have also showed that the orthogonality of the initial perturbations will increase as the number of ensemble members increases. If the number of ensemble members approaches infinity, then the transformed perturbations will be orthogonal under the inverse of the analysis error variance norm. In addition, the initial perturbations generated from the ETR are flow-dependent if the analysis variance is derived from the DA system at every cycle.

ETR belongs to the third generation of initial perturbation techniques, which generate initial perturbations that are consistent with the DA system (Wei et al. 2008). Other such methods include EnKF (ensemble Kalman filter, Evensen 1994; Houtekamer et al. 1996; see method 7), EAKF (ensemble adjustment Kalman filter, Anderson 2001), and ETKF (ensemble transform Kalman filter, Bishop et al. 2001; see method 6). ETR is computationally efficient and was implemented in the NCEP

global ensemble system in 2006 to enhance the breeding method (Wei et al. 2006, 2008). Other similar ensemble forecast systems based on ET, for meteorological and ocean forecasts, have been developed at the NRL's Marine Meteorological Division and Oceanography Division and implemented at the Navy's Fleet Numerical Meteorology and Oceanography Center (FNMOC) and at the US Naval Oceanographic Office (NAVOCEANO), respectively (McLay et al. 2007, 2010; Wei et al. 2014).

5. *Singular vector (SV)*. Singular vectors (SVs) are the perturbations with the fastest growth during a finite time interval. They have been used in the ECMWF ensembles to simulate the effect of initial errors projecting along these directions, since these are the components that would grow fastest and have the largest impact on the forecast quality (Buizza and Palmer 1995; Molteni et al. 1996). At ECMWF, growth is measured by a metric based on a total energy norm. The SVs are computed by solving an eigenvalue problem defined by an operator that is a combination of the tangent forward and adjoint model versions integrated over a time period named the optimization time interval. The advantage of using singular vectors is that if the forecast error evolves linearly and the proper initial norm is used, the resulting ensemble captures the largest amount of forecast error variance at optimization time (Ehrendorfer and Tribbia 1997). Farrell (1982), while studying the growth of perturbations in baroclinic flows, noted that the long-time asymptotic optimization is dominated by discrete exponentially growing normal modes. However, other physically realistic perturbations are possible, which amplify more over a given finite time interval than the most unstable normal mode. Subsequently, Farrell (1988, 1989) showed that perturbations with the fastest growth over a finite time interval could be identified by solving the eigenvalue problem of the product of the tangent forward and adjoint model propagators, supporting earlier conclusions by Lorenz (1965), who pointed out that perturbation growth in realistic models is related to the eigenvalues of the operator product. After Farrell and Lorenz, calculations of perturbations growing over finite time intervals have been performed, for example, by Borges and Hartmann (1992) using a barotropic model and by Molteni and Palmer (1993) using a barotropic and a three-level quasi-geostrophic model at spectral triangular truncation T21. Buizza et al. (1993) first identified singular vectors in a primitive equation model with a large number of degrees of freedom.

The singular vectors are computed by solving an eigenvalue problem defined by the tangent forward and adjoint model equations. Consider the nonlinear model equations:

$$\frac{\partial \chi}{\partial t} = A(\chi, t) \quad (9)$$

The time evolution of a small perturbation x around a time evolving trajectory $\chi(t)$ can be described, in a first approximation, by the linearized model equations:

$$\frac{\partial x}{\partial t} = A_l \cdot x \quad (10)$$

where:

$$A_l = \left. \frac{\delta A(x)}{\delta x} \right|_{\chi(t)} \quad (11)$$

is the tangent operator computed at the trajectory point $\chi(t)$. The equations are clearly valid only for a finite time interval, up to the time when the nonlinear terms can be neglected. Let $L(t,0)$ denote the tangent forward propagator of the linear model equation. In other words, this is the operator that would evolve the initial state $x(0)$ to the forecast time t :

$$x(t) = L(t,0) \cdot x(0) \quad (12)$$

Let E denote a total energy metric, so that:

$$\|x(t)\|^2 = \langle x(t), E \cdot x(t) \rangle \quad (13)$$

is the squared total energy norm of the perturbation $x(t)$, where $\langle \dots, \dots \rangle$ denotes the Euclidean product. Using the definition of the tangent forward operator, the total energy of the perturbation $x(t)$ can be computed as:

$$\|x(t)\|^2 = \langle L \cdot x(0), E \cdot L \cdot x(0) \rangle \quad (14)$$

Let L^* denote the adjoint of the operator L with respect to the Euclidean norm, so that:

$$\langle L \cdot x(0), E \cdot L \cdot x(0) \rangle = \langle x(0), L^* \cdot E \cdot L \cdot x(0) \rangle \quad (15)$$

Then the total energy norm can be computed as:

$$\|x(t)\|^2 = \langle x(0), L^* \cdot E \cdot L \cdot x(0) \rangle \quad (16)$$

The singular vectors are the phase-space directions with the maximum ratio between the final-time and the initial-time norms:

$$\frac{\langle x(0), L^* \cdot E \cdot L \cdot x(0) \rangle}{\langle x(0), E \cdot x(0) \rangle} \quad (17)$$

They are computed by solving an eigenvalue problem defined by the product of the tangent forward and adjoint operators and the total energy metric. In the ECMWF ensemble, the norm is the dry total energy metric, the optimization time interval is 48 h, and the singular vectors are computed at T42L91 resolution. The

reader is referred to Buizza and Palmer (1995), Molteni et al. (1996), and Palmer et al. (2007) for more details.

6. *Conditional nonlinear optimal perturbation (CNOP).* Singular vector approach has been used to explore optimal growth of initial uncertainties in numerical weather forecast and climate prediction. The SVs are a group of orthogonal initial perturbations that possess the largest growth rate in different but mutually orthogonal subspaces of initial perturbations in linearized models. The SV approach was first introduced to meteorology by Lorenz (1965) and established on the basis that the evolution of initial perturbations can be described approximately by the tangent linear model (TLM) of a nonlinear model. The leading SV (LSV), i.e., the SV of the largest growth rate in the TLM, is often used to represent the optimal initial error that has the largest growth rate at prediction time. However, the LSV has difficulties in describing the nonlinear optimal growth of initial perturbations of the finite amplitude due to its linear approximation and then fails to reveal the initial errors that cause the largest prediction errors in predictability studies of weather and climate.

Considering the limitations of LSV, Mu et al. (2003) proposed the approach of conditional nonlinear optimal perturbation (CNOP) to study the nonlinear optimal growth of initial errors. The CNOP represents the initial perturbation that satisfies a certain physical constraint and possesses the largest nonlinear evolution at prediction time (Mu et al. 2003). For a chosen norm $\|\cdot\|$, an initial perturbation $\mathbf{u}_{0\delta}$ is called the CNOP if and only if

$$J(\mathbf{u}_{0\delta}) = \max_{\|\mathbf{u}_0\| \leq \delta} \|\mathbf{M}_{t_0, t}(\mathbf{U}_0 + \mathbf{u}_0) - \mathbf{M}_{t_0, t}(\mathbf{U}_0)\|, \quad (18)$$

where $M_{t_0, t}$ is the propagator of a nonlinear model from an initial time t_0 to a future time t , \mathbf{U}_0 is the initial value of the reference state $U(t)$ [i.e., $U(t) = M_t(\mathbf{U}_0)$] to be predicted, \mathbf{u}_0 is superimposed on $U(t)$ and represents an initial perturbation, and the inequality $\|\mathbf{u}_0\| \leq \delta$ is the constraint of the initial perturbation amplitudes defined by a chosen measurement $\|\cdot\|$. To solve the CNOP, one can transform Eq. (18) into a minimization problem by considering its negative and then calculate the minimization problem by some minimization solvers such as the spectral projected gradient 2 (SPG2; Birgin et al. 2000), sequential quadratic programming (SQP; Powell 1982), limited memory Broyden-Fletcher-Goldfarb-Shanno method (L-BFGS; Liu and Nocedal 1989), or other intelligence *algorithms*.

The CNOP is a natural generalization of LSV in nonlinear regime and defined by directly using a nonlinear model. When the bound of initial constraint is sufficiently small, the CNOP can be approximated by the LSV; when the initial constraint is large, the LSV's approximation to the CNOP does not hold (see the review of Duan and Mu (2009)). In this case, the CNOP represents the initial perturbation that has the largest nonlinear evolution at prediction time and is superior to the LSV in identifying the optimal initial perturbation in nonlinear model (Duan and Mu 2009).

The CNOP has been successfully used to reveal the optimal initial errors and determine the optimal observing locations in the El Nino-Southern Oscillation, tropical cyclone, Indian Ocean Dipole, Kuroshio large meander, and Northern Atlantic Ocean forecasting (see the review of Mu et al. (2015) and Dai et al. (2016)). In the studies of ensemble forecast, Mu and Jiang (2008) focused on the limitation of linear theory of SVs and replaced the LSV with CNOP while keeping other SVs unchanged (hereafter referred to as CNOP+SV approach) to obtain the initial perturbations of ensemble forecasts, showing much higher skill of the CNOP+SV approach against the SV approach (also see Jiang and Mu 2009). This suggests that it is useful for improving ensemble forecast skill to consider nonlinearity in yielding initial perturbations. Therefore, to fully consider nonlinearities in ensemble initial perturbations, Duan and Huo (2016) extended to calculate orthogonal CNOPs. The orthogonal CNOPs are a group of nonlinear optimal initial perturbations denoted as 1st-CNOP, 2nd-CNOP, 3rd-CNOP, ..., n th-CNOP. The j th-CNOP represents the nonlinear optimal initial perturbation in the subspace Ω_j that is orthogonal to 1st-CNOP, 2nd-CNOP, 3rd-CNOP, ..., $j-1$ th-CNOP. The j th-CNOP can be obtained by Eq. (18) but with the constraint condition being $u_{0j} \in \Omega_j$, where Ω_j is as in Eq. (19).

$$\Omega_j = \left\{ \begin{array}{l} \{ u_{0j} \in \mathbb{R}^n \mid \|u_{0j}\| \leq \delta \}, j = 1 \\ \{ u_{0j} \in \mathbb{R}^n \mid \|u_{0j}\| \leq \delta, u_{0j} \perp \Omega_k, k = 1, \dots, j-1 \}, j > 1 \end{array} \right., \quad (19)$$

where u_{0j} is the initial perturbation in the subspace Ω_j . Clearly, the 1st-CNOP possesses the largest nonlinear evolution in the whole space of initial perturbations, and the j th-CNOP possesses the largest nonlinear evolution in the subspace Ω_j orthogonal to the leading $j-1$ CNOPs. These CNOPs are orthogonal and can be ranked as 1st-CNOP $>$ 2nd-CNOP $>$ 3rd-CNOP $>$... $>$ n th-CNOP according to the magnitudes of their evolutions at the final time t .

Duan and Huo (2016) adopted the orthogonal CNOPs to conduct ensemble forecast experiments by using the Lorenz 96 model (Lorenz 1996). They further increased the ensemble forecast skill generated by the CNOP+SVs because the orthogonal CNOPs fully considered the nonlinearity in ensemble initial perturbations. Furthermore, Huo (2016) applied the orthogonal CNOPs to much realistic weather model MM5 (fifth-generation Pennsylvania State University National Center for Atmospheric Research Mesoscale Model; Dudhia 1993) and conducted the ensemble forecast experiments for typhoon track. They also showed great usefulness of the orthogonal CNOPs in achieving ensemble forecasts of higher skill. The CNOPs could be another useful approach to generating initial perturbations of ensemble forecasting.

7. Ensemble transform Kalman filter (ETKF). The ETKF method, since its inception, has gone through a series of theoretical developments and has been widely applied to various applications including targeting observations (Bishop et al. 2001; Majumdar et al. 2002), ensemble prediction (Wang and Bishop 2003;

Wang et al. (2004); Bowler et al. (2008); Hacker et al. (2011), and hybrid ETKF-variational data assimilation (Wang et al. 2007, 2008a, b, 2009, 2011). The ETKF algorithm used for ensemble generation is briefly summarized here, but more details are documented in Bishop et al. (2001), Wang and Bishop (2003), Wang et al. (2004), and Wang et al. (2007).

In the ETKF ensemble generation scheme, the forecast ensemble perturbations are updated by the ETKF to produce the analysis ensemble perturbations. The ETKF transforms the matrix of forecast perturbations \mathbf{X}^b into a matrix of analysis perturbations \mathbf{X}^a , whose columns contain K analysis perturbations, \mathbf{x}'^a_k , $k = 1 \dots K$. The transformation happens through the postmultiplication by the matrix \mathbf{T} , that is,

$$\mathbf{X}^a = \mathbf{X}^b \mathbf{T} \quad (20)$$

The transformation matrix is chosen to ensure that the analysis error covariance formed from the outer product of the transformed perturbations will be precisely equal to the true analysis error covariance, assuming that the covariance of the forecast ensemble denotes the true forecast error covariance, all errors are normally distributed, and \mathbf{H} is linear. As shown in Bishop et al. (2001), Wang and Bishop (2003), and Wang et al. (2004), a precise spherical simplex solution of \mathbf{T} is:

$$\mathbf{T} = \mathbf{C}(\mathbf{I} + \mathbf{\Gamma})^{-1/2} \mathbf{C}^T \quad (21)$$

where \mathbf{C} contains the eigenvectors and $\mathbf{\Gamma}$ the eigenvalues of the $K \times K$ matrix $(\mathbf{X}^b)^T \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{X}^b$ and \mathbf{I} is the identity matrix. For the ensemble size K of 100 or less, the computation is relatively inexpensive.

Research has shown that when K is significantly smaller than the rank r of the true forecast error covariance, this formula systematically underestimates the analysis error variance. In Wang and Bishop (2003), the ensemble of analyzed deviations \mathbf{X}^a was inflated by large factors to compensate for the ETKF's underestimate of the analysis error variance (Wang and Bishop 2003). An alternative formulation of the ETKF was proposed by Wang et al. (2007) that significantly ameliorated this bias by accounting for (i) the fact that the sample covariance of K forecast trials systematically overestimates the true error variance within the ensemble subspace when $K \ll r$ and (ii) the expected difference in angle subtended between ensemble-based eigenvectors and true eigenvectors. Based on these arguments, the ETKF transformation matrix \mathbf{T} becomes:

$$T = C(\rho\Gamma + I)^{-1/2} C^T \quad (22)$$

where the scalar factor ρ is the fraction of the forecast error variance projected onto the ensemble subspace. It is estimated by:

$$\rho = \frac{\overline{\left(R^{-\frac{1}{2}}y - \tilde{H}\bar{x}^b \right)^T E E^T \left(R^{-\frac{1}{2}}y - \tilde{H}\bar{x}^b \right)} - (K-1)}{\overline{\left(R^{-\frac{1}{2}}y - \tilde{H}\bar{x}^b \right)^T \left(R^{-\frac{1}{2}}y - \tilde{H}\bar{x}^b \right)} - p} \quad (23)$$

where ρ is the number of observations, $\tilde{\mathbf{H}}$ is the normalized observation operator $\tilde{\mathbf{H}} = \mathbf{R}^{-1/2}\mathbf{H}$, and the columns of \mathbf{E} contain the eigenvectors of the ensemble covariance in normalized observation space. As shown in Eq. (12) of Bishop et al. (2001):

$$\bar{E} = \tilde{H}\bar{x}^b C\Gamma^{-1/2} / \sqrt{K-1} \quad (24)$$

The overbar represents the average over some independent samples. Note that the computational efficiency of the ETKF is realized by solving the transformation matrix in ensemble perturbation subspace (Bishop et al. 2001; Wang and Bishop 2003).

To further ameliorate the underestimation of the analysis error variance, an inflation factor is applied to increase the ensemble covariance. For example, the maximum-likelihood inflation method is applied to the analysis perturbations in the study of Wang and Bishop (2003). The idea is to multiply the initial perturbations obtained at time t_i by an estimated inflation factor Π_i , that is:

$$X_i^a = X_i^f T_i \boldsymbol{\Pi}_i \quad (25)$$

The purpose of this is to ensure that at time t_{i+1} the background ensemble forecast variance is consistent with the ensemble-mean background-error variance over global observation sites. Specifically, we define $\tilde{\mathbf{d}}_i$ as the innovation vector at t_i , normalized by the square root of the observation error covariance matrix, that is, $\tilde{\mathbf{d}}_i = \mathbf{R}^{-1/2}(\mathbf{y}_i - \mathbf{Hx}^b_i)$, where \mathbf{y}_i is the observation vector at t_i and \mathbf{Hx}^b_i is the ensemble mean background forecast valid at time t_i mapped into observation space by the observation operator \mathbf{H} . Given that the inflation factor at t_{i-1} was Π_{i-1} , the inflation factor for the transformed perturbation at t_i is obtained by first checking if $\tilde{\mathbf{d}}_i^T \tilde{\mathbf{d}}_i$ is equal to $\text{Tr}(\tilde{\mathbf{H}}\mathbf{P}_i^e \tilde{\mathbf{H}}^T + \mathbf{I})$, where Tr denotes the trace. If it is not, we need to introduce a parameter c_i so that:

$$\tilde{d}_i^T \tilde{d}_i = \text{Tr} \left(\tilde{\mathbf{H}} c_i \mathbf{P}_i^e \tilde{\mathbf{H}}^T + \mathbf{I} \right) \quad (26)$$

Then the inflation factor Π_i is defined as:

$$\Pi_i = \Pi_{i-1} \sqrt{c_i} \quad (27)$$

This rescaling of the initial perturbations attempts to correct the spread of the set of forecast ensemble perturbations at time t_{i+1} by using the rescaling factor which would have produced a proper forecast ensemble spread at t_i if it had been applied to the transformed perturbations at t_{i-1} . Accordingly:

$$c_i = \frac{\tilde{d}_i^T \tilde{d}_i - p}{\text{Tr}(\tilde{H} P_i^e \tilde{H}^T)} \quad (28)$$

where p is the number of observations. Π_i , therefore, is a product of these c parameters from the first forecast at time t_1 to that at time t_i , that is:

$$\Pi_i = \sqrt{c_1 c_2 \cdots c_i} \quad (29)$$

Implicitly we assume $\tilde{\mathbf{d}}^T \tilde{\mathbf{d}}_i = \text{Tr}(\tilde{\mathbf{d}} \tilde{\mathbf{d}}_i^T)$, which requires the number of independent elements in the innovation vector $\tilde{\mathbf{d}}_i$ to be large. The real-time global observational network meets this assumption well (Dee 1995). For regional applications, because the number of observations in our experiment is rather limited, we replace $\tilde{\mathbf{d}}^T \tilde{\mathbf{d}}_i$ by using the average of the squared innovation vectors from 2 weeks prior to time t_i , denoted as $\overline{\tilde{\mathbf{d}}^T \tilde{\mathbf{d}}}_{\text{prior } t_i}$. Thus it becomes:

$$c_i = \frac{\overline{\tilde{d}^T \tilde{d}_{\text{prior } t_i}} - p}{\text{Tr}(\tilde{H} P_i^e \tilde{H}^T)} \quad (30)$$

8. *Ensemble Kalman filter (EnKF).* The EnKF can be viewed as a method to perturb observations. It performs a Monte Carlo simulation of errors as they evolve in a data assimilation cycle (Evensen 1994; Houtekamer and Mitchell 1998). In the Monte Carlo method, each member of an ensemble samples the uncertainty in the inputs (e.g., observations) of the system. For each set of perturbed inputs, the system is subsequently run to provide an output that reflects the uncertainty in the input. If the input ensemble is representative of the input uncertainty, then the output ensemble will be representative of the output uncertainty.

In the EnKF, a data assimilation cycle is performed using Eqs. 31–33. In Eq. 31, an N-member ensemble $\hat{f}_i(t)$ of prior estimates of the state of the atmosphere at time t is combined with the new observations o available at this time. The forward operator H goes from model space to observation space and can be nonlinear. To sample uncertainty in Eq. 31, one can use an ensemble oi of observation sets that have been randomly perturbed with respect to the actual observation set o . Similarly, when in Eq. 33 the forecast model M is used to integrate the analyses $ai(t)$ forward in time to obtain an ensemble of prior estimates $\hat{f}_i(t+1)$ at the next analysis time $t+1$, an ensemble of perturbation fields qi can be added to reflect the uncertainty in the numerical model. Note that a Kalman gain matrix K is used to give an appropriate weight to the observations o , which have error covariance R , and the prior estimate f , which has error covariance Pf as in Eq. 32.

$$ai(t) = \hat{f}_i(t) + K(oi - H\hat{f}_i(t)), i = 1, \dots, N \quad (31)$$

$$K = P_f H T (H P_f H T + R) - 1 \quad (32)$$

$$f_i(t+1) = M(a_i(t)) + q_i, \quad i = 1, \dots, N \quad (33)$$

The EnKF is conceptually similar to the random perturbation method, but it is probably easier to specify the uncertainties in the observations o than to specify the uncertainties in the analysis a . In the data assimilation cycle, the perturbations grow and evolve with the dynamics of the flow as represented with the model M . They will thus reflect the errors of the day as in the breeding method. In the data assimilation step (Eqs. 31 and 32), the covariances P_f estimated from the available ensemble of trial fields f_i are used to determine an optimal weight of the information in the trial fields and the new l observations. Thus, the ensemble-based knowledge of the flow-dependent errors in the trial field is used to optimally spread the information of observations into space. Here, the main difference with the Kalman filter (Kalman 1960) is that the covariance matrix P_f for the error in the high-dimensional trial fields is estimated using an ensemble of typically $O(100)$ members as in Eq. 34:

$$P_f = \frac{1}{N-1} \sum_{i=1}^N (f_i - \bar{f})(f_i - \bar{f})^T \quad (34)$$

The rank $N - 1$ of the ensemble-based covariance matrix is usually augmented using a covariance localization technique (Hamill et al. 2001), and practical implementations will avoid having to store the full matrix P_f .

The data assimilation step (Eq. 31) will reduce uncertainty, as estimated with the evolving ensemble, where observations are available. This is similar to the ETR method except that the ensemble covariances, not just the variances in a specific space, are fully consistent with the statistical properties of the observational network.

The ensemble of analyses that are available from the EnKF can be used to provide the ensemble of initial conditions for an ensemble prediction system. In the ensemble prediction system, the same Monte Carlo principles can be applied to sample important uncertainty in the forecast model and in boundary conditions such as the soil moisture field. The coherent treatment of all known sources of error is an attractive property of the EnKF. It is increasingly being used for theoretical studies as well as for operational applications. At Environment Canada an EnKF has been in operational use for global atmospheric data assimilation since 2006 (Houtekamer et al. 2014). It provides initial conditions for the global and regional ensemble prediction systems. At NCEP, an EnKF has been used to support the global deterministic assimilation system since 2012, initially as a hybrid 3DEnVar (Kleist and Ide 2015a; Hamill et al. 2011; Wang et al. 2013), which was then upgraded to hybrid 4DEnVar in 2016 (Kleist and Ide 2015b). The 20 out of 80 EnKF-based analyses are also used to initialize the NCEP global ensemble. At ECMWF, an EnKF is now being developed in research.

9. Perturbations in boundary conditions including land surface and topography.

Besides perturbations to the model's interior initial states, the lower, upper, and

lateral (for limited area models) boundary conditions need to be perturbed too. The lower boundary forcing is introduced by land and water surface layer parameters such as sea surface temperature, heat and moisture flux, ice and snow cover, soil properties (moisture, temperature, and type), surface albedo, roughness, and greenness. For example, soil moisture uncertainty has a significant impact on convective precipitation in the warm season (Sutton et al. 2006; Aligo et al. 2007; Du et al. 2007) and surface temperature (Du et al. 2007). The sensitivity of 2-m temperature to initial soil moisture shows a strong diurnal variation related to solar radiation (much stronger during daytime than nighttime) and geographically preferred regions (Du et al. 2007). Du et al. (2007) also reported that the impact of soil moisture perturbation on an ensemble forecast depends on the perturbation's spatial structure and magnitude, e.g., uniformly wet or dry in space and larger magnitude perturbations produce a larger ensemble spread than spatially uncorrelated random wet or dry and smaller magnitude perturbations. The uncertainties in an upper boundary forcing from space (such as solar activity) and the way an upper boundary (model top) been treated in a model haven't been accounted for in current EPSs. This needs to be studied quantitatively.

The difference between model and real-world topography contributes to the forecast error especially for precipitation, temperature, and wind. A terrain perturbation scheme has been first incorporated into an EPS by Li et al. (2017). The terrain ensemble in their study is constructed by using different combinations of two terrain smoothing schemes and three terrain interpolation schemes within the WRF model to produce six ensemble members. They tested the terrain-perturbing scheme and compared with other initial condition and physics perturbation ensembles using the extremely heavy rain event that occurred on July 21, 2012 in Beijing. They found that perturbing model terrain could produce a spatial pattern of ensemble spread similar to those produced by either initial condition or physics perturbation ensembles, although the magnitude of spread is impactfully large but much smaller in the terrain-perturbing ensemble compared to those in the initial condition and physics perturbation ensembles (Fig. 3). However, that ensemble spread and probabilistic forecasts were improved by incorporating terrain uncertainty into initial condition perturbation ensemble, while the ensemble mean of precipitation forecasts remains similar (Fig. 3). Therefore, perturbing topography alone will not produce large enough ensemble spread. Instead terrain perturbation should be combined with initial condition and physics perturbation methods in constructing an ensemble. By the way, in this same research, the authors decomposed the ensemble spread into different spatial scales and concluded that the selection of perturbation methods in constructing an EPS is more important for small-scale (<448 km) and very short-range forecasts (<12 h) than for larger-scale and longer-range forecasts. In other words, for larger synoptic scale and longer-range (over a day) forecasts, the selection of perturbation methods is less critical, and the impacts from either IC perturbations or physics perturbations or other types of perturbations are often

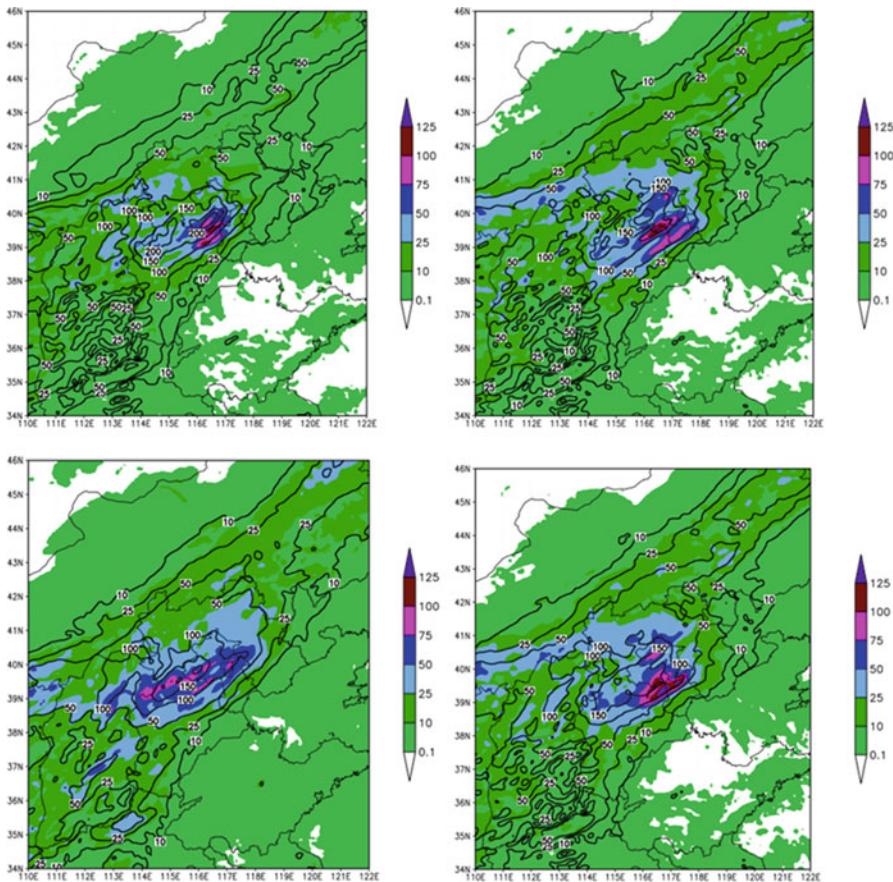


Fig. 3 Ensemble mean forecasts (contour) and spreads (color) of the 24-h accumulated precipitation during the 24–48-h WRF model integration period for the four perturbation schemes (upper left, terrain perturbation; upper right, IC perturbation; lower left, physics perturbation; and lower right, a combination of terrain and IC perturbations. Unit: mm)

similar to each other. This can be used as a general guideline to design an effective EPS.

For a regional EPS with a small domain, lateral boundary conditions (LBCs) could play a dominant role in defining the ensemble spread of atmospheric state variables (less so for precipitation) (Du and Tracton 1999; Warner et al. 1997). Therefore, LBCs should also be perturbed to ensure diverse ensemble solutions. At the same time, a sufficiently large model domain is recommended for a regional EPS. Currently, using different members from a global EPS as LBCs for different members in a regional EPS is a common practice, such as in the NCEP SREF (Du et al. 2004). Nutter et al. (2004a, b) suggested an approach to compensate for the ensemble spread loss due to LBC.

3 Methods to Address Uncertainty in Model

Model errors stem from two sources: (a) mathematical simplification and limitation of numerical calculations both spatially and temporally, such as finite grids including a limited domain for regional models and discrete time steps, and (b) imperfect treatment of physical processes such as parameterization. As a consequence, any numerical forecast is bound to develop errors due to differences between the model tendencies and the real-world tendencies even if the ICs were perfect. Model errors could behave systematically or randomly. A properly designed, reliable EPS should aim at representing the random errors in the tendencies in order to estimate reliable uncertainties. Although much less is understood about the behavior of model uncertainty compared to that of IC uncertainty, the following three approaches are currently applied to address model-related uncertainties: multi-model, multi-physics, and stochastic physics. Prior to these methods being used to directly address model-related uncertainty, inflation of the IC perturbation size was often used to count indirectly for model uncertainty.

1. *Multi-model and multi-physics (MM-MP).* MM-MPs are ad hoc but effective approaches. Multi-model can represent uncertainties in both the physics and numerical schemes (dynamics), while multi-physics represents uncertainties only in the physics schemes. The advantages of these methods are relatively straightforward construction and effectiveness in capturing forecast uncertainties with a large ensemble spread. Potential bias cancelation in ensemble averaging is another benefit from the MM-MP methods, due to the different biases in different members which often result in a much improved ensemble mean forecast (Duan et al. 2012). In contrast, an IC perturbation-based or stochastic physics-based EPS can reduce random error but not systematic error in its ensemble averaging, due to similar biases in all the members. On other hand, the disadvantages of these methods might include the following: they simulate the uncertainty in model formulation but not in the subgrid processes; it is hard to achieve the statistical equal likeliness of members' performance due to differences in members' quality; members are usually grouped by models or by physics schemes, which results in spurious ensemble spread caused by individual model or scheme biases; and it is costly for a single NWP center to maintain and develop multiple model or physics schemes. Based on favorable research results (Mullen et al. 1999; Tracton et al. 1998), NCEP implemented a MM-MP-based short-range ensemble forecast (SREF) system in operations (Du and Tracton 2001; Du et al. 2003). The multi-model approach is widely accepted and used nowadays, e.g., the THORPEX Interactive Grand Global Ensemble (TIGGE) and North American Ensemble Forecasting System (NAEFS). Note that an MM ensemble is really meant as an ensemble of ensembles on many occasions.

Slightly altering the values of the physics parameters is one approach to stochastic physics. It can be viewed as multiple slightly varied versions of the same physics or

model. It is easier to configure: one normally doesn't need to modify a physics scheme itself but just set different values in a model-controlling namelist file for the different members. The challenging part is that one needs to determine a reasonable range for a parameter to vary, which needs a lot of testing and experience. If the variation is too big, it could push an atmospheric flow into an unrealistic regime. More sophisticated stochastic physics approaches are described in detail below.

2. *Stochastically perturbed parameterization tendency (SPPT) scheme.* Schematically, each ensemble forecast is defined by the time integration of model equations:

$$x_j(t) = x_j(0) + \int_0^t F_j d\tau \quad (35)$$

$$F_j = A_j + P_j + (r_j P_j + b_j) \quad (36)$$

where the *j-th* suffix identifies the *j-th* ensemble member, $x_j(t)$ is the model state vector at time t , F_j is the full model tendency, and $x_j(0)$ is the *j-th* member initial condition. The total tendency F_j includes an adiabatic term A_j ; a diabatic term P_j , which depends on the physical parameterization scheme; a multiplicative ($r_j P_j$); and an additive (b_j) model error term. At the ECMWF, these model error terms are defined by two schemes:

- The multiplicative term is defined by the SPPT scheme (Buizza et al. 1999; Palmer et al. 2009).
- The additive term is defined by the stochastic kinetic energy backscatter (SKEB) scheme (Shutts 2004, 2005; Berner et al. 2009; Palmer et al. 2009).

These two schemes have been designed to both have a physical basis and to be as simple and effective as possible. Since the two schemes represent different possible sources of model error, they are both used, albeit in slightly different configurations, in the ECMWF ensembles (see Shutts et al. 2011 for more details on the settings of the operational schemes). The SPPT is described in this subsection and the SKEB will be described in the next subsection.

The rationale behind the SPPT scheme is that

There are certainly good grounds for believing that there is a significant source of random error associated with the parameterized physical processes ... The sort of random error in parameterized forcing will be coherent between the different parameterisation models, and will have a certain coherence on the space and time scales associated, for example, with organized convection. Moreover, the scheme assumes that the larger the parameterized tendencies, the larger the random error component will be. The notion of coherence between modules allows the stochastic perturbation to be based on the total tendency of all parameterized process. (Buizza et al. 1999)

The current SPPT scheme at the ECMWF is a revision of the original stochastic diabatic tendency scheme of Buizza et al. (1999) and perturbs the parameterized tendency of physical processes with multiplicative noise. It is based on the notion that, especially with increasing numerical resolution, the equilibrium assumption no longer holds and the subgrid-scale state should be sampled rather than represented by the equilibrium mean. Consequently, SPPT multiplies the accumulated physical tendencies at each grid point and time step with a random pattern that has spatial and temporal correlations. This concept can be schematically described by Eq. (37) for any model state variable X at forecast time t :

$$X_t = X_{t-1} + F + S = X_{t-1} + F + \mu^* r^* F, \quad (37)$$

where S is an added stochastic term which is a product of a random number $\mu^* r$ and tendency F . μ is a vertical weight which decays with height from 1.0 at the surface to 0.0 in the stratosphere (100–50 hPa). r is a random number ranging from -1.0 to 1.0 representing horizontal weights and has a pattern in space and time (e.g., a red noise process with a temporal time scale of 6 h and an e-folding spatial scale of 500 km). SPPT uses same random pattern generator as SKEB but with a different normalization. The stochastic pattern evolves in spectral space as

$$f_n^m(t + \Delta t) = \alpha f_n^m(t) + g_n \varepsilon(t) \quad (38)$$

where all variables are as defined above. The temporal correlations are given by the decorrelation time τ defining $\alpha = \exp(-\Delta t/\tau)$. The noise amplitudes are given as

$$g_n = F_0 \exp(-Ln(n+1)/2), \quad (39)$$

$$F_0 = \left(\frac{\sigma^2(1-\alpha^2)}{2 \sum_{n=1}^N (2n+1) \exp(-Ln(n+1))} \right)^{1/2}, \quad (40)$$

where L is a horizontal length scale defining the spatial correlations and σ^2 is the perturbation variance at each grid point. The normalization constant F_0 is chosen so that the variance at any grid point, σ^2 , is given by the total variance in spectral space (Weaver and Courtier 2001). The resulting stochastic pattern follows at each grid point a Gaussian with mean zero and the variance σ .

Using the NCEP global EPS, Whitaker et al. (2013, NCEP seminar) compared the SPPT with the control ensemble (no stochastic physics) for wind forecasts over the globe at day 5. They found that the overall impact on ensemble spread is small with a slight spread increase, mostly in the tropics, and ensemble mean forecast error remained unchanged. The SPPT method has been included in the NCEP global forecast system for EnKF-3DVAR hybrid data assimilation. Recently, SPPT has also been used experimentally to model diffusion processes (Qiao et al. 2017).

3. Stochastic kinetic energy backscatter (SKEB) scheme. The rationale behind the SKEB scheme is that due to its finite approximation, while numerical models

simulate the energy cascade from the resolved to unresolved scales, they do not represent the upscale energy transfer from scales smaller than the model grid onto the scales resolved by the model. SKEB estimates the downscale energy transfer and simulates the energy backscatter from the unresolved to resolved scales. Therefore, the scheme aims to represent model uncertainty arising from unresolved subgrid-scale processes by introducing random perturbations to the stream function and, depending on the implementation, the potential temperature tendencies. Originally developed in the context of large eddy simulations (Mason and Thompson 1992) and applied to models of intermediate complexity (Frederiksen and Davies 1997), it was adapted by Shutts (2004, 2005) to NWP. SKEB assumes that a small fraction of the model dissipated energy interacts with the resolved-scale flow and acts as a systematic forcing. Its impact on weather and climate forecasts is reported, e.g., in Berner et al. (2008, 2009, 2011, 2012), Bowler et al. (2008, 2009), Li et al. (2008), Palmer et al. (2009), Doblas-Reyes et al. (2009), Charron et al. (2010), Hacker et al. (2011), Tennant et al. (2011), and Weisheimer et al. (2011). Below is a technical description of the method.

Let $f(\phi, \lambda, t)$ be a 2D stochastic pattern expressed in a triangularly truncated spherical harmonics expansion:

$$f(\phi, \lambda, t) = \sum_{m=-N}^N \sum_{n=|m|}^N f_n^m(t) P_n^{|m|}(\cos \phi) \exp^{im\lambda}, \quad (41)$$

where λ and ϕ denote longitude and latitude in physical space and t time. In spherical harmonics space, m and n denote the zonal and total wavenumbers and N is the truncation wavenumber of the numerical model and is the associated Legendre function P_m^n of degree n and order m . The spherical harmonics $Y_m^n = P_m^n e^{im\lambda}$ form an orthogonal set of basis functions on the sphere. If the f_n^m are nonvanishing for at least one $n < N$ and do not follow a white-noise spectrum, the pattern perturbations will be spatially correlated in physical space.

Since the physical processes mimicked by this forcing have finite correlation times, temporal correlations are introduced by evolving each spectral coefficient as a first-order autoregressive (AR1) process:

$$f_n^m(t + \Delta t) = \alpha f_n^m(t) + \sqrt{(\alpha - 1)} g_n \varepsilon(t), \quad (42)$$

where α is the linear autoregressive parameter determining the temporal decorrelation time, g_n the wavenumber-dependent noise amplitude, and ε is a Gaussian white-noise process with mean zero and variance η . The noise amplitude g_n is chosen to have power-law behavior, $g_n = b n^p$, and to determine the variance spectrum of the forcing.

The pattern $f(\phi, \lambda, t)$ is interpreted as the stream function tendency forcing. In cases of perturbed potential temperature, a second perturbation pattern is created analogously, but with a different power-law behavior and a potentially different

temporal correlation. The adaption to 2D-double periodic domains as used in regional models is straightforward and described, e.g., in Berner et al. (2011).

The behavior of this scheme is determined by the following parameters: the exponent of the power law, p ; the wavenumber perturbation range, n_1-n_2 ; and the amplitude of forcing energy, which determines the normalization constant b .

In the ECMWF implementation of SKEB, the stream function pattern is subsequently weighted with the normalized total instantaneous dissipation rate from numerical dissipation, deep convection, and gravity and mountain wave drag (Shutts 2005; Berner et al. 2009). A simplified version of SKEB is available to the public as part of the Weather Research and Forecasting (WRF) Model with Advanced Research WRF dynamics solver (Skamarock et al. 2008). This simplification no longer injects random perturbations proportional to the estimated total dissipation rate, but assumes a spatially and temporally constant dissipation, resulting in a state-independent stochastic forcing (Berner et al. 2011). Instead it relies on the underlying model dynamics to determine which perturbations will grow and which ones will be damped.

The SKEB method is also compared to a control ensemble (no stochastic physics) using the NCEP global EPS for wind forecasts over the globe at day 5 in Whitaker et al. (2013, NCEP seminar). They found that the SKEB increased ensemble spread in mid-latitude jets where numerical dissipation is active but was less effective in the tropics due to convective dissipation not included in the scheme, although it did add spread in the lower tropics. The impact on ensemble mean forecast is neutral. The SKEB method has been included in the NCEP global forecast system for EnKF-3DVAR hybrid data assimilation.

4. *Stochastic trigger of convection.* Since the SPPT only modulates the existing physics tendency inside a parameterized physics scheme, it can only change convection intensity and cannot trigger new convection or eliminate present convection. To alter the area of convection, two methods have been used. One is the stochastic trigger of convection (STC, Li et al. 2015) which directly adds a stochastic term to the convection trigger function:

$$T_{lcl} + (1 + r) * \Delta T > T_{env} \quad (43)$$

where T_{lcl} is the air parcel temperature at LCL (lifted condensation level), ΔT is the temperature change of the parcel, r a random number or pattern, and T_{env} is the environmental temperature. When the resulting parcel temperature exceeds its environmental temperature, free convection is triggered.

5. *Stochastic boundary-layer humidity (SHUM).* Another method to trigger a convection is an indirect method called stochastic boundary-layer humidity (SHUM, Whitaker et al. 2013, NCEP seminar). The rationale behind SHUM is that triggers in convection schemes are very sensitive to boundary-layer humidity. The specific humidity q in the boundary layer is stochastically perturbed at each time step as follows:

$$qp = (1 + r^* \mu)q \quad (44)$$

where q and qp are the original and perturbed specific humidity, vertical weight μ decays exponentially from the surface, and random pattern r has the same horizontal/temporal scales as SPPT with a small amplitude of ~ 0.001 . By comparing SHUM to a control ensemble (no stochastic physics) using the NCEP global EPS for wind forecasts over the globe at day 5, Whitaker et al. (2013, NCEP seminar) found that the SHUM notably improved spread-error consistency in the tropics, especially the upper tropics, e.g., the maximum forecast error near the tropopause was reproduced in the spread, but had little or no effect in the winter hemisphere poleward of 30° latitude. As a result, the ensemble mean forecast error was reduced in the tropics, especially the upper tropics, and the summer hemisphere. The SHUM method has been included in the NCEP global forecast system for EnKF-3DVAR hybrid data assimilation.

6. *Stochastic total tendency perturbation (STTP) scheme.* In the previous SPPT scheme, individual physics schemes are perturbed separately. Therefore, one needs to perturb all model physics schemes in order to sample the full physics uncertainties from all known sources. However, even if all physics schemes are perturbed, uncertainties related to model numerical structures and any unknown sources are still being missed. To overcome this weakness, the model total tendency, instead of partial tendencies in individual physics, is perturbed. This scheme is called the stochastic total tendency perturbation (STTP) scheme and was implemented in the NCEP global ensemble forecast system (GEFS) in February 2010 (Hou et al. 2006, 2008). Although the general principle (adding a stochastic term to the tendency of state variables) is the same as the SPPT, the STTP is technically different from SPPT in two ways, i.e., using the perturbation tendency instead of full variable tendency and using a series of orthogonal weights to combine the perturbation tendencies from all members. The ensemble perturbation tendencies are first randomly combined following certain rules to form stochastic total tendency perturbations, which are then scaled to the appropriate size and used as stochastic forcing terms in the model equations. This method is based on the hypothesis that differences in the tendencies among ensemble perturbations provide a representative sample of the random total model errors associated with the formulation of the dynamic and physical processes, truncation, and parameterizations. The detailed scheme is described below.

With subscript i identifying one of the N ensemble members, $i = 1, 2, \dots, N$ (0 is the control forecast) and t the time of the integration, the conventional model equations for an ensemble forecast system running with only initial perturbations can be written as:

$$\frac{\partial X_i}{\partial t} = T(X_i, t) \quad (45)$$

where T is the total tendency, including dynamical and physical processes, calculated at the grid scale with parameterization of subgrid-scale effects.

Considering the uncertainty in the model formulation and numerical approximation, a stochastic forcing term S_i should be added to each member, i.e.,

$$\frac{\partial X_i}{\partial t} = T(X_i; t) + S_i(t) \quad (46)$$

The formulation of stochastic forcing with the SPPT approach generally relates S to the tendency increment due to a particular component of tendency T , i.e., representing the random error associated with a particular physical process. However, in the STTP S comes from the total tendency T . As the perturbations in ICs lead to a reasonably (though not perfectly) representative sample of the possible model states, one can assume that the conventional tendencies in the individual ensemble members collectively provide a representative sample of the unknown true value of the total tendency. By comparing the total tendency in each ensemble member against the control forecast, N tendency perturbations, i.e.,

$$P_i(t) = T_i(t) - T_0(t) \text{ for } i = 1, 2, \dots, N \quad (47)$$

can be identified, and they form a representative sample of the differences between the true tendency and that formulated in the model Eq. (45). Therefore, these tendency perturbations can be used as the basis in formulating the stochastic forcing S .

As in the SPPT approach, random numbers are introduced to address the uncertainty in the total tendency. Although each single P_i , if chosen randomly, can be a valid candidate, a random combination of all N tendency perturbations would be a better choice, in hopes that more directions in the phase space would be explored for the ensemble perturbations $X_i - X_0$ to grow faster and the ensemble spread to be increased. Symbolically, we have:

$$S_i(t) \propto \sum_{j=1}^N w_{i,j}(t) P_j(t) \text{ for } i = 1, 2, \dots, N \quad (48)$$

where the coefficients $w_{i,j}$ are random weights assigned for each P . The stochastic forcing for each ensemble member i corresponds to a different set of random weights $w_{i,j}$, $j = 1, \dots, N$.

Using matrix notation and omitting the time t , the relation (48) can be rewritten as:

$$S_{NM} \sim W_{NN} P_{NM} \quad (49)$$

where the subscripts indicate the matrix dimensions and M is the number of grid points. Note that this formula is the same as the ensemble transform (ET) used in the generation of initial perturbations, if P_{NM} are the perturbations of model states.

To determine the combination or weighting matrix W , one needs to consider the requirements for the stochastic forcing S . First, as (46) is to be applied to all model state variables with the same set of weights, the stochastic forcing S should be in approximate balance as are the tendency perturbations P . Second, the S vectors should be orthogonal to each other. Since the P vectors form an approximately orthogonal set, the orthogonality in S can be achieved if the W matrix is orthonormal, i.e., the w vectors are normalized and orthogonal to each other. Therefore, the problem is to specify a random but orthonormal matrix W as a function of time. The temporal variation of the W matrix is represented by random rotations from one application to the next or mathematically as:

$$W_{NN}(t) = W_{NN}(t-1)R_{NN}(t) \quad (50)$$

where R is a random matrix only slightly different from the identity matrix I , representing a random and slight rotation of the $N w$ vectors in an N -dimensional space. The rotation at a particular time, $R(t)$, can be viewed as the combination of a steady rotation, which is represented by a random but temporally invariant matrix R^0 , and a random rotation R^1 , which changes at every application of the scheme, i.e.,

$$R_{NN}(t) = R^0_{NN}R^1_{NN}(t-1) \quad (51)$$

James Purser of NCEP developed the methodology and software to generate a random orthonormal matrix and a random rotation matrix. Both procedures start with filling an $N \times N$ matrix A with independent random numbers from a Gaussian distribution. The orthonormalization is then realized by applying the Gram-Schmidt procedure (Golub and Van Loan 1996) to A . The rotation matrices R^0 and R^1 are generated by applying the same procedure to $(I + \alpha(A - A^T))$, where A^T is the transpose of A and α the “degree” of rotation. These algorithms are used to generate the temporally varying weighting matrix W via the following procedures: (1) initializing W by generating a random orthonormal matrix $W(t=0)$; (2) specifying the fractional numbers α_0 to prescribe the “degree” of rotation in the steady rotation and generate R^0 ; (3) specifying another fractional number α_1 for the degree of random rotation to find R^1 ; (4) generating a random slight rotation matrix for each time that the stochastic perturbation scheme is applied, using the same α_1 but a different seed; and using (50) and (51) to update the W matrix.

The temporal evolution of the weighting matrix W can be viewed as N vectors in the N -dimensional space, changing their directions slightly with random vibrations (R^1) imposed on a steady rotation (R^0). Similarly, the evolution of each scalar weighting factor $w_{i,j}$ is seen as random increments (corresponding to R^1) superimposed on a smooth trend in the form of a periodic function of time (corresponding to R^0) with the level of noise (due to the random increment) and the period controlled by α_1 and α_0 , respectively. α_1 and α_0 are the only two parameters required to specify $W(t)$. While a higher value of α_1 defines noisier curves, a larger α_0 corresponds to shorter periods. Figure 4 depicts some examples of these curves in a 10-member ($N = 10$) ensemble system, showing the curves for $i = 10$ and $j = 1, 2, \dots, 10$, i.e.,

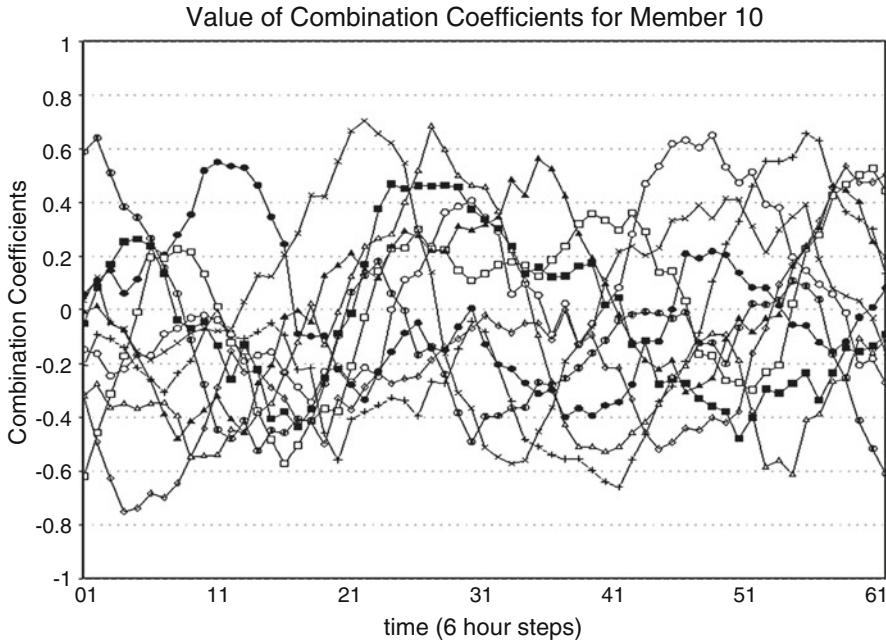


Fig. 4 Examples of combination coefficients $w_{i,j}$ as a function of forecast lead time, in a 10-member ($N = 10$) ensemble system. Shown are ten curves for $j = 1, 2, \dots, 10$ (each of the ten originating members) with fixed $i = 10$ (for the member 10) and $\alpha_0 = \alpha_1 = 0.05$

the temporal variation of weighting factors that determine the stochastic forcing for ensemble member 10. In this particular case with $\alpha_0 = \alpha_1 = 0.05$ and 6-h intervals between applications, it can be seen from Fig. 3 that the period of the trend is about 6 days and the curves look fairly noisy. For reference, $\alpha_1 = 0.005$ defines smoother curves, while $\alpha_0 = 0.005$ corresponds to a much longer period (>10 days).

In principle, the stochastic perturbation scheme can be applied at every time step of the model integration. However, a less frequent application is preferable, to reduce the computational cost in practice. For this purpose, a finite difference version of Eq. (46) is used in an operational implementation. With a specified time interval of the application designated as Δt , the stochastic scheme can be implemented by integrating Eq. (45) instead of (44) from $t-\Delta t$ to t and modifying the model state variables (X) by using:

$$X'_i = X_i + \gamma(t) \sum_{j=1}^N w_{i,j}(t) \left\{ \left[(X_j)_t - (X_j)_{t-\Delta t} \right] - \left[(X_0)_t - (X_0)_{t-\Delta t} \right] \right\} \quad (52)$$

for $i = 1, 2, \dots, N$ at $t = \Delta t$, $t = 2\Delta\tau, \dots$. $\gamma(t)$ is a scaling factor that varies with time but is uniform across all ensemble members. Its values depend on the choice of time interval Δt , and its temporal variation is related to that of the size of the ensemble

perturbations. The scaling factor is empirically determined for a fixed Δt and factorized as a global rescaling factor γ_0 and a regional rescaling factor γ_1 , i.e.,

$$\gamma = \gamma_1(\varphi, d)\gamma_0(t) \quad (53)$$

The global rescaling factor is a function of forecast lead time only and expressed as:

$$\gamma_0(t) = \pm \left[p_2 + (p_1 - p_2) \left\{ 1.0 - \frac{1.0}{1.0 + e^{-p_3(t-p_4)}} \right\} \right] \quad (54)$$

where p_1 , p_2 , p_3 , and p_4 are empirical parameters, e.g., $p_1 = 0.100$, $p_2 = 0.01$, $p_3 = 0.11$, and $p_4 = 252$ h were used in the February 2010 GEFS implementation, and $p_1 = 0.105$, $p_2 = 0.03$, $p_3 = 0.12$, and $p_4 = 252$ h were used in the January 2012 GEFS implementation. Experiments suggest that these empirical values generally work well, although fine tuning and optimization could be done. The regional rescaling factor, in its current form, changes with latitude and the day of a year:

$$\gamma_1(\varphi, d) = 1.0 + 0.2 \sin(\varphi) \cos \frac{2\pi d}{364} \quad (55)$$

Equation (55) indicates that the perturbation size in the winter hemisphere is larger than that in the summer hemisphere. As shown as an example in Fig. 4, the stochastic forcing vectors have structures of random noise, and the size, represented by a vector norm similar to total energy, shows a flow-dependent global distribution with the largest amplitudes associated with the mid-latitude jets in both hemispheres.

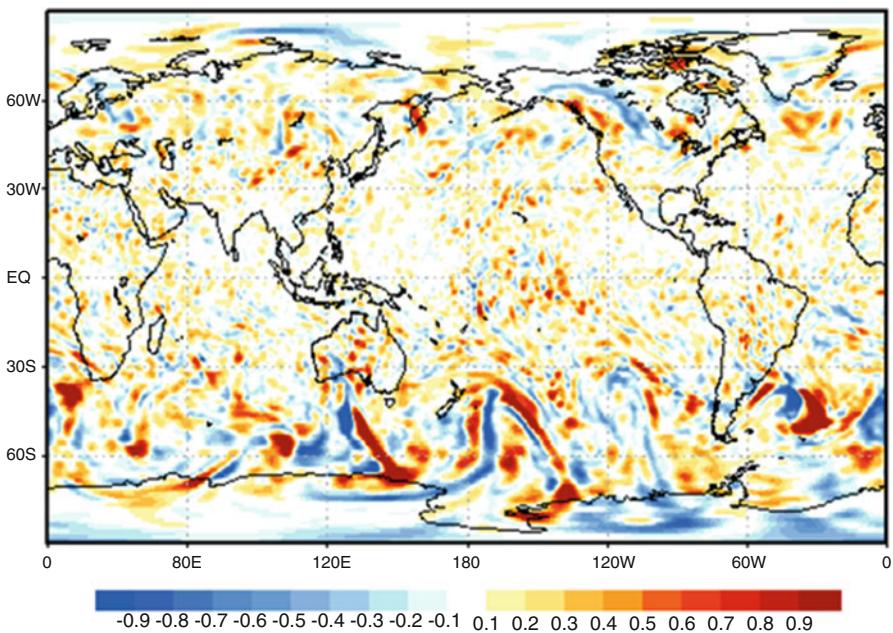
The quantity defined by the summation in Eq. (52) is referred as a stochastic perturbation (SP) applied to the i -th ensemble member, and it can be rewritten as:

$$SP_i = \sum_{j=1}^N w_{i,j}(t) \left\{ \left[(X_j)_t - (X_0)_t \right] - \left[(X_j)_{t-\Delta t} - (X_0)_{t-\Delta t} \right] \right\} \text{ for } i = 1, 2, \dots, N. \quad (56)$$

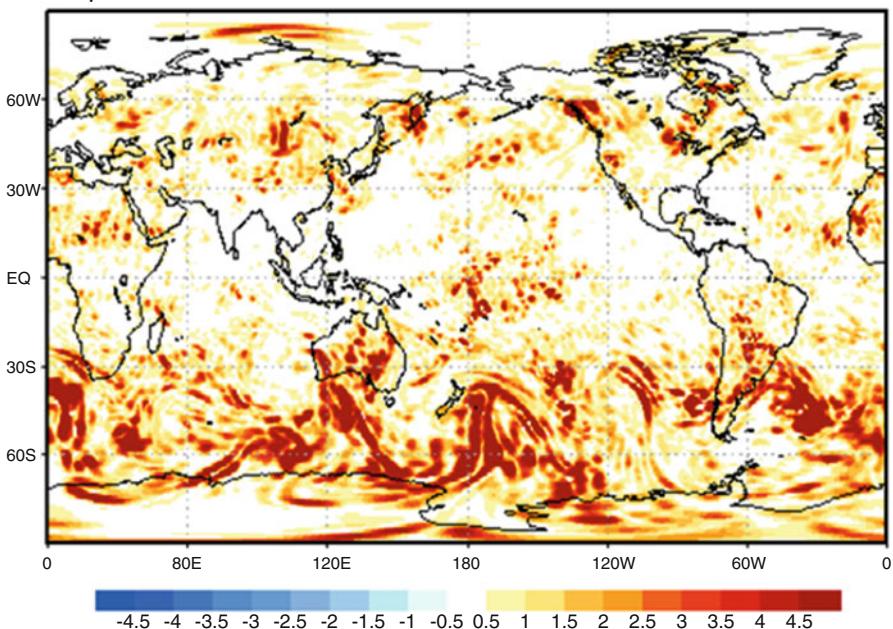
As the weighting matrix W is orthonormal, the size of each SP is determined by the changes in all ensemble perturbations during the past Δt time interval (the quantity in the curly brackets). The SPs generated by (56) are for all prognostic variables of the model state and they are in approximate balance. An example of SPs is shown in Fig. 5.

Its implementation is straightforward, with a periodic stopping of the N otherwise independent integrations (45), modifying each model state with information from all N model states using (52) and repeating the procedure every Δt hours (6 h is currently used at NCEP). This requires all N sets of model states at time level t and $t-\Delta t$ to be available simultaneously and can be easily realized if the N ensemble members concurrently run within a single executable. The values of the required parameters Δt , α_0 , α_1 , p_1 , p_2 , p_3 , and p_4 can be manipulated in a namelist.

Temp pert., Memb. 20, LEVEL 13, 120hr fcst 2008082500Z



pert. TE. Memb, 20, LEVEL 13, 120hr fcst 2008082500Z

**Fig. 5** (continued)

Using the NCEP GEFS, Whitaker et al. (2013, NCEP seminar) also compared the STTP to a control ensemble (no stochastic physics) for wind forecasts over the globe at day 5. They found that the STTP increased spread mainly in the winter hemisphere, less so in the summer hemisphere, and very little in the tropics and had little impact on the ensemble mean forecast, which is consistent with the design of this method.

7. *Vorticity confinement.* This method was originally proposed by Steinhoff and Underhill (1994) and Shutts and Allen (2007) and was first tested in an atmospheric general circulation model by Sanchez et al. (2013). They found that resolution-dependent biases grow when model resolution decreases as a result of the lack of transient eddy kinetic energy (TEKE). This might inhibit the development of mid-latitude variability phenomena such as synoptic cyclones and blocking events. To compensate for the loss of TEKE in a model, especially at low resolution, they added a parameterized vorticity confinement (VC) term (the last term of Eq. 57) to the horizontal momentum equation to preserve vorticity features in dissipative numerical schemes as follows:

$$\frac{D\mathbf{V}_H}{Dt} + f\mathbf{k} \times \mathbf{V}_H + \nabla\phi = \mu\nabla^2\mathbf{V}_H + \epsilon\hat{\mathbf{n}} \times |\zeta|\hat{\mathbf{k}} \quad (57)$$

$$\hat{\mathbf{n}} = \frac{\nabla_H\zeta}{|\nabla_H\zeta|}$$

In the last term of Eq. (57), $\hat{\mathbf{n}}$ is the normalized horizontal vorticity gradient, ϵ can be viewed as the tangential speed of the VC, $\epsilon \hat{\mathbf{n}}$ acts as an advective velocity, $|\zeta|\hat{\mathbf{k}}$ is the vertical component of relative vorticity, $\hat{\mathbf{n}}$ is in the normal direction pointing to higher vorticity or the left of flow in northern hemisphere, and $\hat{\mathbf{k}}$ is the direction vector pointing up. Therefore, the resulting VC force is always horizontal and along the direction of flow by adding back momentum to against model numerical diffusion, i.e., accelerating both cyclonic and anticyclonic flow in the mid-latitudes. The magnitude of the VC force is proportional to the vorticity. The parameter ϵ controls the strength of the confinement term and acts as a type of anti-diffusive velocity. $\epsilon = 0.6$ is used in the Sanchez et al.'s (2013) experiment. Some early tests, in which ϵ was proportional to the amount of kinetic energy dissipated, proved to be very unstable for the model. It is suggested that different formulations for ϵ , whose value is dependent on the local flow (Hahn and Iaccarino 2009), need to be explored in the future. A drawback of this method is that it might inadequately change the radial distribution of vorticity to potentially hurt a forecast.

Fig. 5 An example of stochastic perturbations (SPs) added to modify the model states of the 20-member NCEP GEFS. Shown are (upper panel) the temperature perturbation (unit: K) associated with ensemble number 20 and (lower panel) the corresponding perturbation size defined as the square root of “total energy” norm of the perturbation vector (unit: ms^{-1}), at 120-h integration time starting from 00Z, August 25, 2008

Whitaker et al. (2013, NCEP seminar) applied this VC method ($\varepsilon = 0.6$ is used) in an ensemble model (the NCEP global EPS) to simulate forecast uncertainty. They found that ensemble spread was increased mainly in the subtropics but not in the tropics, tropical cyclones became stronger, and the ensemble mean forecast error increased in a 5-day forecast. The VC method has been included in the NCEP global forecast system for EnKF-3DVAR hybrid data assimilation.

8. *First-order Markov chains – Generating space-time auto-correlated 2D random fields on the sphere.* Random fields with a specified space-time auto-correlation are often needed in ensemble forecasting and ensemble data assimilation, such as in the SPPT and SKEB, to control the noise characteristics of the random patterns. In the current implementations of the SKEB scheme and SPPT or physical parameter perturbations at the Canadian Meteorological Center (Li et al. 2008; Berner et al. 2009; Charron et al. 2010), first-order Markov chains are employed to generate a random time series. This method is described below. Only the two-dimensional case is described, but an extension to three spatial dimensions is straightforward.

Markov processes can be used as spectral coefficients of an expansion on spherical harmonics (in the horizontal). A two-dimensional random function on the sphere, $f(\lambda, \phi, t)$, correlated in space and time, with a probability density function (PDF) symmetric around the mean μ , can be defined as:

$$f(\lambda, \phi, t) = \mu + \sum_{l=L_{\min}}^{L_{\max}} \sum_{m=-l}^l a_{l,m}(t) Y_{l,m}(\lambda, \phi), \quad (58)$$

with

$$a_{l,m}(t + \Delta t) = e^{-\Delta t/\tau} a_{l,m}(t) + \sqrt{\frac{4\pi\sigma^2(1 - e^{-2\Delta t/\tau})}{(2l+1)(L_{\max} - L_{\min} + 1)}} R_{l,m}(t). \quad (59)$$

The independent variables λ , ϕ , and t are longitude, latitude, and time, respectively. The $Y_{l,m}$'s are spherical harmonics, with l being the total horizontal wavenumber and m the zonal wavenumber. The normalization convention is as follows:

$$\begin{aligned} \int_S d\Omega Y_{l,m}(\lambda, \phi) Y_{l',m'}^*(\lambda, \phi) &= \int_0^{2\pi} d\lambda \int_{-1}^1 d(\sin \phi) Y_{l,m}(\lambda, \phi) Y_{l',m'}^*(\lambda, \phi) \\ &= \delta_{l,l'} \delta_{m,m'}. \end{aligned} \quad (60)$$

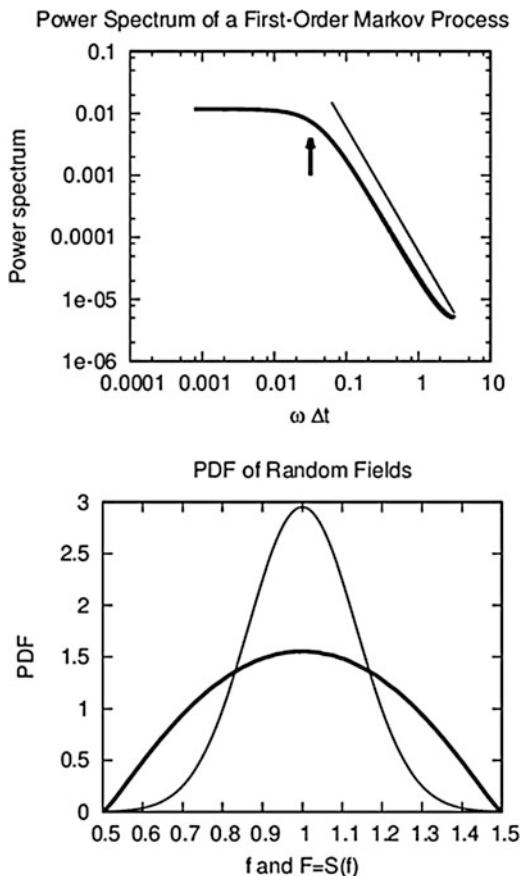
The integer parameters $L_{\min} \geq 1$ and $L_{\max} \geq L_{\min}$ define the spectral range of the random function. Their inverse can be interpreted in terms of spatial decorrelation length scales. The parameter τ is the decorrelation time scale of the spectral coefficients. For simplicity, it is defined here as a constant independent of l , although a

generalization as $\tau = \tau(l)$ might be useful. The parameter Δt is the time step of the numerical model. The complex $R_{l,m}$'s are uncorrelated random processes with a mean of zero and a variance $|R_{l,m}|^2$ of unity. The overbar denotes an ensemble mean or a time mean. The $R_{l,m}$'s can be, for example, Gaussian processes. The denominator in the square root of Eq. 60 has been chosen here to generate a white-noise signal (in space) with a specified mean global variance σ^2 (a modeler might want to impose another specified spatial spectrum by modifying this denominator). For real random fields, the condition $a_{l,m}(t) = (-1)^m a_{l,-m}^*(t)$ must apply. It can be shown that the correlation function is written

$$\frac{a_{l,m}(t+s)a_{l,m}^*(t)}{|a_{l,m}|^2} = e^{-s/\tau}, \quad (61)$$

where $s = n\Delta t$ and n a positive integer. The upper panel of Fig. 6 depicts the power spectrum of $a_{l,m}$ as a function of frequency. For time scales much larger than τ , $a_{l,m}$ is

Fig. 6 Upper panel: power spectrum of a first-order Markov chain as a function of frequency (thick line) and a reference -2 slope (thin line). The arrow indicates the position of $\Delta t/\tau$. Lower panel: the probability density function (PDF) of f with $\sigma = 0.135$ (thin line) and F (thick line)



white noise. For time scales smaller than τ , the spectrum has an approximate -2 slope.

One often needs to keep a perturbed physical parameter within some specified bounds. Equations 58 and 59 do not ensure that $f(\lambda, \phi, t)$ is bounded, say, between f_{\min} and f_{\max} . Moreover, the PDF of f is a Gaussian distribution when the $R_{l,m}$'s are Gaussian or when the sums over l and m in Eq. 58 are made with a large enough number of components (central limit theorem). A modeler might need a different PDF. These potential drawbacks can be fixed by applying a stretching to f to obtain F . For example, if one is interested in generating an F drawn from a uniform distribution, it can be shown that the stretching must be performed using the error function, provided that f is a Gaussian random field with mean μ and variance σ^2 :

$$F = \frac{(f_{\max} - f_{\min})}{2} \operatorname{erf}\left(\frac{f - \mu}{\gamma\sigma\sqrt{2}}\right) + \mu. \quad (62)$$

The parameter γ allows the shaping of the PDF of the stretched random function F . When $\gamma = 1$, the PDF of F is uniform. The lower panel of Fig. 6 shows the impact of stretching on the PDF when $\gamma = 1.5$, $f_{\min} = 0.5$, and $f_{\max} = 1.5$, i.e., a broadening of the PDF due to stretching.

9. *Factor separation method.* In a multi-physics-based EPS, an understanding of the relative impacts from different physics schemes as well as their interactions among schemes is desired. This could be done in different ways, such as a clustering technique (Johnson et al. 2011) and decomposition (Jankov et al. 2005). Jankov et al. (2005) demonstrated a factor separation method and found that the information derived from the method is useful in constructing and calibrating an EPS (Jankov et al. 2007). In this subsection, the factor separation method is introduced. However, one needs to keep in mind that a complete separation of factors is impossible in a nonlinear environment where a nonlinear interaction term always present.

The factor separation method was formulated by Stein and Alpert (1993). The method is summarized as:

$$f_{xy} - f_0 = (f_x - f_0) + (f_y - f_0) + \hat{f}_{xy} \quad (63)$$

$$\hat{f}_{xy} = f_{xy} - (f_x + f_y) + f_0 \quad (64)$$

where f_0 represents the control forecast with standard physics schemes (say, convection x_0 and microphysics y_0) and f_{xy} the experimental forecast using two alternative physics schemes x and y (convection x and microphysics y). Therefore, the term $f_{xy} - f_0$ is the difference from the control forecast due to changes in the two physics schemes. Equation (63) tells us that this difference can be attributed to three terms:

the change in scheme x alone $f_x - f_0$, in scheme y alone $f_y - f_0$, and the nonlinear interaction of schemes x and y $f_{xy}^* - f_{xy}$ is called the synergistic term and can be estimated by Eq. (64). Assuming a continuum of physical schemes, Eq. (63) is then equivalent to Taylor's series second-order expansion in two variables. The first two terms on the right-hand side of Eq. (63) represent the contribution of the first-order derivatives, while the third term (synergistic term) is a mixed second-order derivative (the unmixed second-order derivatives are zero). If the synergistic term is equal to zero, no interaction occurs between the two alternative schemes, which is an additive linear system.

In the work of Jankov et al. (2005, 2007), the impact of various physics schemes and their interactions on warm season precipitation caused by a continental meso-scale convective system (MCS) was evaluated using this method. Simulations of eight MCS events were performed using 18 WRF-ARW model configurations (members) consisting of three different convection treatments, three different micro-physics schemes, two different PBL schemes, and two different initializations over a 1500×1500 -km domain with 12-km grid spacing. They showed that this method is able to quantify the relative impacts of different physics schemes, e.g., a change in the convection scheme affects the rain rate the most, while both convection and microphysics are important for the rain volume depending on the initialization of the analysis. Information about the relative impacts and interactions were then used to construct four smaller ensembles. The performance of the four ensembles supported the results from the factor separation method, i.e., convection and microphysics have the largest impact on the simulated MCS rainfall. This demonstrated that the knowledge of which physics schemes exert the greatest impact on a forecast can allow for the design of smart ensembles that maximize forecast skill while minimizing the ensemble size.

10. *Perturbing model dynamic core.* Since the local change of a model state variable is a sum of physical and dynamical tendencies, the dynamical tendency can also be perturbed separately as we did for the physical tendency. For example, Koo and Hong (2014) perturbed the dynamic tendency in a global model to improve seasonal prediction. They added a random number to the dynamic tendency term to account for the inherent uncertainties associated with computational representations of the underlying partial differential equations that govern the atmospheric motion (their Eqs. 2a–c). Their stochastic forcing depends on forecast time and vertical layer. By making a comparison with the traditional approach of perturbing the physical tendency, they reported that the sensitivity of fluctuations in forecast variables to the magnitude of random forcing is found to be even greater in the case of a perturbed dynamical tendency. They also evaluated a simulated climate for a boreal summer. It demonstrates a significant enhancement in forecast skill in terms of the large-scale features and precipitation when both the dynamical and physical tendencies are simultaneously perturbed. This finding implies that model uncertainties can be addressed in terms of not only the physical parameterization but also the dynamical score.

4 Virtual Ensembles

Without enough computing resources to physically run a state-of-the-art EPS, some alternative virtual ensembles are proposed to quantitatively estimate forecast uncertainty. Five approaches are introduced in this section: a time-lagged ensemble, poor-man's ensemble, dual-resolution hybrid ensemble, neighborhood ensemble, and analog ensemble. A common advantage of all these methods is the minimal cost in constructing an ensemble of forecasts.

1. The *time-lagged ensemble* is proposed by Hoffman and Kalnay (1983). It pulls multiple forecasts, initiated at different times but all verified at same time, together to form an ensemble (i.e., a mixture of older and newer forecasts). The degree of run-to-run consistency is presumably a measure of forecast confidence. The advantage of this method is the inclusion of past information as well as the ready availability of forecast data at one NWP center. With this approach any operational NWP center automatically has an ensemble system if it runs at least one single model. Such a single model can run at its highest possible spatial resolution since there is no actual EPS model competing with it for computing resources. A limitation is on the size of the ensemble, since there are not many “good” older forecasts available due to the rapid degradation of forecast quality with forecast lead time. The more frequently a model is initialized to run, the more time-lagged members can be created. Given the inequality in the members’ quality, weights are normally assigned to different members based on their forecast age when producing ensemble products. The time-lagged approach has been used in operations and for research such as the NCEP operational seasonal ensemble forecast system (Saha et al. 2006) and aviation and high-resolution ensembles (Zhou et al. 2010; Du and Zhou 2017), as well as others (Lu et al. 2007; Brankovic et al. 2006; Mittermaier 2007).
2. A *poor-man’s ensemble* is a collection of many single-model forecasts from various available sources since a “poor” man cannot afford to run his own model forecasts (Wobus and Kalnay 1995; Ebert 2001). It can be viewed as the simplest version of a multi-model ensemble. The advantage of this method is its comprehensive sampling of possible uncertainty sources including different ICs, data assimilation systems, physics, and model dynamic cores, which often results in larger ensemble spread. A superb ensemble mean forecast is another advantage of this method, due to the apparent different biases in different members (Duan et al. 2012). As with the time-lagged ensemble, limited ensemble size and the inequality in members’ quality are two weaknesses of this method. Therefore, different weights are normally assigned to different members based on their past performance in producing final forecast products. For example, the so-called super-ensemble is actually a poor-man’s ensemble used to produce a deterministic forecast using a statistical linear regression technique (Krishnamurti et al. 1999). This multi-model-based linear regression approach can significantly improve forecast accuracy over the original forecasts by correcting model biases.

3. *The dual-resolution hybrid ensemble* is proposed by Du (2004). If one desires to have a high-res ensemble but cannot afford to run it at his desired resolution, this approach is a way to go if a low-res ensemble and a high-res single-model run are available. A synthetic downscaled high-res ensemble can be constructed using this hybrid ensembling approach, by combining the forecast variance from the low-res ensemble and single high-res forecast, as described below. At any grid point and forecast time, each low-res perturbed ensemble member (`Lres_mem`) is decomposed into a base forecast (low-res control forecast, `Lres_ctl`) and a forecast perturbation `Fpert` (Eq. 65):

$$\text{Lres_mem} = \text{Lres_ctl} + \text{Fpert}. \quad (65)$$

Using the high-res single run (`Hres_single`) as a new base forecast to replace `Lres_ctl` in Eq. (65), a new perturbed high-res ensemble member (`Hres_mem_p`) can be obtained as in Eq. (66):

$$\text{Hres_mem_p} = \text{Hres_single} + \text{Fpert}. \quad (66)$$

If Eq. (66) is the only way used to create new high-res members, it is referred as a one-sided approach (i.e., “addition” only), and the resulting ensemble is $\{\text{Hres_single}, \text{Hres_mem_p}\}$. The subscript p means “positive.” If “subtraction” is also used to obtain new members as in Eq. (67), it is referred as two-sided approach (i.e., both “addition” and “subtraction” are applied), and the resulting ensemble is $\{\text{Hres_single}, \text{Hres_mem_n}, \text{Hres_mem_p}\}$. The subscript n means “negative.”

$$\text{Hres_mem_n} = \text{Hres_single} - \text{Fpert}. \quad (67)$$

The ensemble size remains the same as the original in the one-sided approach, while it doubles in membership in the two-sided approach. Note that for precipitation and other humidity fields, a “positive value” constraint needs to be set during the calculations of Eqs. (66 and 67) to keep their values physical. Since the high-res forecast `Hres_single` is often more accurate with detailed spatial structures than the original low-res control forecast `Lres_ctl`, especially in the short range, the new hybrid ensemble normally outperforms the original low-res ensemble, especially for heavy rain events (Du 2004; Tang et al. 2015). A potential danger from this approach is the possible spatial mismatch between the new base forecast (from the high-resolution model) and the old forecast perturbation (from the low-resolution model) because they come from two models of differing resolutions. To minimize this mismatch, hybrid ensembling should be applied within the same model and at short ranges such as 1–3 days. This approach has been applied to both regional and global ensembles at NCEP. In the regional ensemble, a 16-km regional ensemble was combined with a 4-km single run to produce a new storm-scale ensemble. Improvements in heavy precipitation and surface wind forecasts are observed. In the NCEP global ensemble, the 27-km single high-res global model forecast was used with the 55-km global ensemble. Considering the decrease in the superiority of high-res over low-res model

forecasts with forecast lead time, a decaying weight function is used when combining high-res with low-res forecasts: ranging from 1.0 (i.e., 100% using high-res model info) to 0.0 (100% using low-res model info) over a 5-day period. This hybrid step is proven to greatly boost the NCEP GEFS performance.

4. The *neighborhood ensemble* is proposed by Theis et al. (2005) and Roberts and Lean (2008) for single forecast and by Schwartz et al. (2010) and Schwartz and Sobash (2017) for ensemble forecasts. As opposed to the downscaled hybrid ensemble, neighborhood ensemble is an upscaled ensemble. Although a very high-resolution storm-scale model may well simulate the detailed spatial structures of weather elements such as convective cells and extreme precipitation maxima, there are considerable uncertainties associated with the predicted locations, due to limited predictability. Therefore, beyond a certain spatial scale (critical scale), only probabilistic skill (not deterministic) presents, or an area averaged value (rather than individual grid point values) is more representative of reality. Forecasts at scales below this critical scale should be expressed in probabilistic form. Figure 7 demonstrates how this approach works: assume that a 1-km resolution model has no deterministic prediction skill for precipitation at scales smaller than 6 km (i.e., the critical scale in this case). Therefore, all grid points within the 3-km radius circle can form an ensemble of forecasts, so that probabilistic and ensemble mean forecasts at location A can be calculated based on this ensemble. Thus, the neighborhood method can transform a single high-res deterministic forecast into a probabilistic forecast or have the effect of turning a small ensemble into a larger ensemble. By upscaling in space, the resulting probabilistic or ensemble mean forecast should be more reliable than the original forecasts from the individual

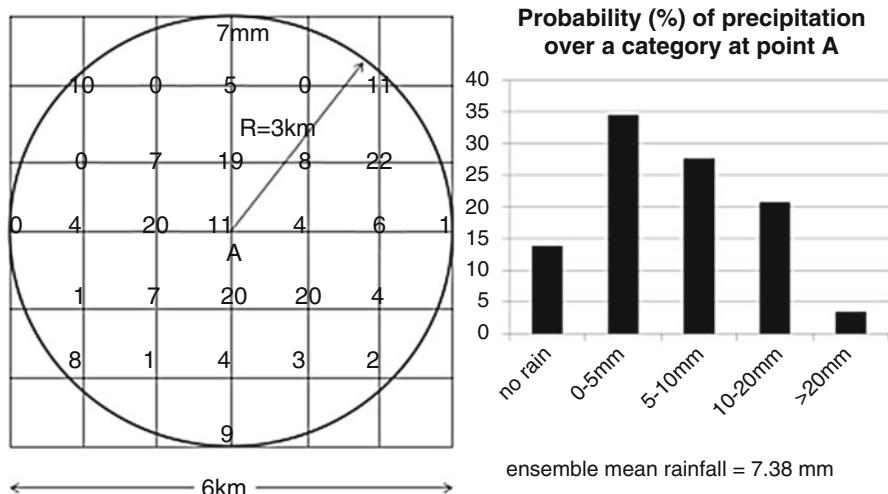


Fig. 7 An illustration using the “Neighborhood Approach” (a) to create probabilistic and ensemble mean precipitation forecasts at point A (b) in a high-resolution (1-km) model forecast

points. In this method, properly estimating the critical scale to distinguish deterministic forecast skill is key and depends on each model's capability as well as the predictability of an event (spread). Given large variation of ensemble spread in space and time, Dey et al. (2014, 2016) proposed the "ensemble agreement scale" technique to estimate an ensemble-spread (similarity of members)-dependent critical scale or impact radius on each grid point. Blake et al. (2018) applied this variable-radius (adaptive) technique to convection-allowing model ensemble and found that the resulting neighborhood probabilistic has been improved over the traditional fixed-radius approach for heavy rain forecasts during the 2017 NCEP Flash Flood and Intense Rainfall Experiment. In their study, the critical scale (impact radius) size ranges from 10 km for member forecasts that are in good agreement to 100 km when the members are more dissimilar.

Therefore, running a model even at a scale which might not have deterministic skill can still be justified in order to take advantage of more sophisticated physics or fine topography information, etc., as long as the model has probabilistic skill.

5. An *analog ensemble* is constructed by first matching up the current deterministic forecast or an ensemble of forecasts from an NWP model with similar past forecasts (e.g., a reforecast – single or ensemble) based on predetermined criteria (Hamill and Whitaker 2006; Hamill et al. 2008, 2015; Eckel and Delle Monache 2016). The verifying observation from each match is then used as an ensemble member. Therefore, the analog ensemble is really an ensemble of observations but not model forecasts. The degree of success in this approach obviously depends on how the analog is selected. Since an analog ensemble directly uses observations as members (and can be viewed as a model forecast with 100% error correction), the impact from the deficiencies of the IC and model, leading to an imperfect forecast, has automatically been corrected. In other words, the requirement for a high quality of IC and model may be eased. Besides this, another advantage of an analog ensemble over an NWP ensemble is little or no need for post-processing calibration of members (observations). Although an analog ensemble can capture flow-dependent error growth, it may miss the aspects of error growth that can be represented dynamically by multiple real-time model runs of an NWP ensemble. To combine the strengths of analog and NWP ensembles, a hybrid of the two has normally been used, i.e., finding m analogs for each member of a small n-member NWP ensemble, to produce a total of $m \times n$ members. Delle Monache et al. (2011, 2013) tested this hybrid approach in wind energy forecasting and compared the forecast skill between an analog ensemble, an NWP ensemble, and a hybrid of the two calibrated using logistic regression. They found that the hybrid outperforms the other approaches for probabilistic 2-m temperature forecasts yet underperforms for 10-m wind speed. The mixed results reveal a dependence on the intrinsic skill of the NWP members employed. In their study, the NWP ensemble is under-dispersed for both 2-m temperature and 10-m winds, yet displays some ability to represent flow-dependent error for the former though not the latter. Therefore, they concluded

that a hybrid of analog and NWP ensembles is a promising approach for efficient generation of high-quality probabilistic forecasts, but requires the use of a small and at least partially functional NWP ensemble. The 2012 Atmospheric River Experiment at NCEP's Weather Prediction Center compared a reforecast-based analog ensemble (also a hybrid version with 11 NWP ensemble members) and a multi-model (ECMWF, NCEP, and CMC global EPS)-based NWP ensemble in predicting heavy precipitation events at the 3.5–5.5-day range over the United States. Their results show that the analog ensemble is apparently superior to the multi-model NWP ensemble in medium-range heavy rain forecasts (Figs. 8 and 9, Du and Li 2014). For short-range (0–3 day) forecasts, a regime-dependent bias correction method was proposed and operationally implemented (but not activated) for the NCEP SREF (Du and DiMego 2008). It is a method aligned with the same idea of an analog ensemble, which uses the analog forecasts' errors

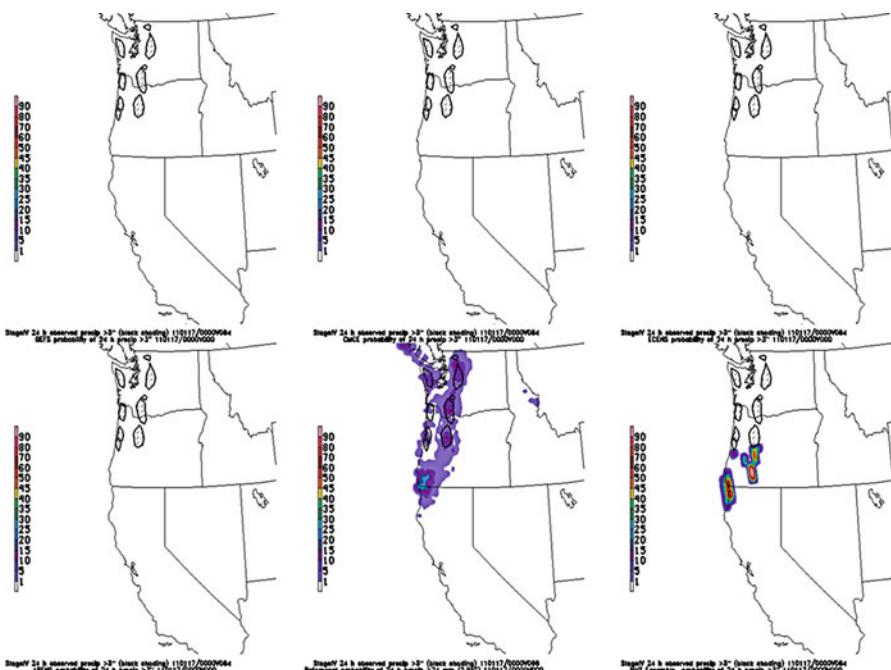


Fig. 8 A comparison between the “Reforecasting Analog Ensemble” and “Multi-Model Ensemble” approaches for a US west coast heavy rain case of 00z January 17, 2011 (one of the eight cases examined during the 2012 NCEP Weather Prediction Center's Atmospheric River Experiment). Truth (stage IV precipitation analysis, the shaded area): six main heavy rainfall centers exceeding 75 mm/24 h, scattered in the west of both Washington and Oregon. The 3.5-day probabilistic forecasts of rainfall exceeding 75 mm/24 h (the colored areas): NCEP GEFS (upper left), CMC CMCE (upper middle), ECMWF ECENS (upper right), a multi-model GEFS/CMCE/ECENS combined grand ensemble ARENS (lower left), a reforecasting analog ensemble reforecast (4-day forecast, lower middle), and a 7-km regional ensemble HMT-ENS (lower right). (Thanks to Mr. Thomas Workoff for plotting Figs. 8 and 9)

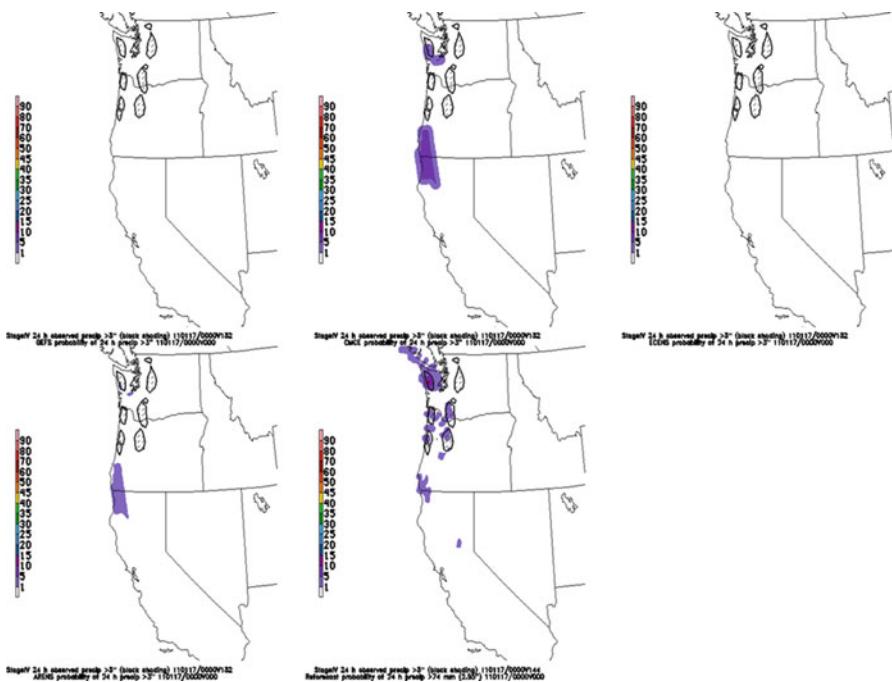


Fig. 9 Same as Fig. 8 but for the 5.5-day forecasts, as well as without the 7-km regional ensemble HMT-ENS

within the prior 20 days to estimate bias error (weighted mean) to calibrate the current forecasts. The spatial correlation of a forecast field between the current forecast and a past forecast is used to identify analogs. The results show that it works better than a simple (blind) equally weighted running mean of all past forecast errors, especially for regime transition periods and fast-varying fields like wind. An advantage of this regime-dependent bias correction method is it is cheap and it's also easy to handle the past forecast data (no need to access a long historical data archive), which is a critical factor in a real-time operational environment. The disadvantage is that no good analogs (but only relatively similar forecasts) might exist within such a short time period (20 days). This method is more in the post-processing arena rather than in creating an ensemble.

5 Ensemble Size

Due to limited computing resources available in real-time operations, ensemble size and model spatial resolution must be balanced against each other. Therefore, a frequently asked question is how many members are needed in an EPS and what

is the trade-off between ensemble size and model resolution. Based on an early study by Du et al. (1997), approximately ten members are enough for an ensemble to achieve most (90%) of the available increment in forecast skill for a large-scale precipitation forecast in terms of both ensemble mean and probabilistic forecasts. This conclusion is confirmed by other studies such as Talagrand et al. (1997). However, the Du et al. (1997) study is based on a coarse-resolution (80-km) model with parameterized convection; will the same conclusion hold in a storm-scale convection-allowing model? A study by Clark et al. (2011) showed that ten members seem also sufficient for a probabilistic precipitation forecast using a convection-allowing scale (4-km) model; at the same time they pointed out that the required membership will likely need to increase to obtain the maximum available benefit from the ensemble when forecast lead time increases or model resolution further increases. This agrees with the conclusion of Richardson (2001): more members are needed in an ensemble when the predictability of an event is lower. Richardson (2001) argued that probability scores such as the Brier score or Brier skill score have a theoretical cap or limit for a given ensemble size. The Brier skill score increases rapidly with the increase in ensemble size when the ensemble size is small ($<= 10$ members) and becomes nearly saturated when the ensemble size is larger ($> = 50$ members), which is particularly true for low-probability events. It implies that a probabilistic forecast cannot reach its full skill if the ensemble size is too small, especially for low predictability events. Increasing the ensemble size should be greatly beneficial when ensemble size is small, while the ensemble size impact will become smaller when the ensemble size is larger. For example, in a fog prediction when the ensemble size is increased from 5 to 10, forecast performance is noticeably improved (Zhou and Du 2010; Du and Zhou 2017); on other hand, Roquelaure and Bergot (2008) also demonstrated that there is little improvement in fog forecasts when going from a 30-member to 54-member ensemble.

Based on the studies mentioned above, the general rule is that for short-range weather forecasts, increasing the model resolution and using more sophisticated physics schemes is more beneficial than increasing the ensemble size alone as long as ensemble size is large enough, such as more than ten members. Clark et al. (2009, 2010) have demonstrated that a 5 (10)-member small ensemble using a 4-km convection-permitting model outperformed a 15 (30)-member large ensemble using a 20-km parameterized convection model in precipitation (convection) forecasts. On other hand, for medium-range or longer-range forecasts (with lower predictability), increasing the ensemble membership might be more beneficial than a resolution increase (Ma et al. 2012). Vertical resolution is also important to model performance (Aligo et al. 2009); it might be interesting to also study what the optimal trade-off will be between vertical resolution and ensemble size (studies not yet seen).

In real-world operations the actual ensemble size should depend on one's purpose as well as other nonscientific factors. For example, the membership required might be less for 500-hPa height than for convective precipitation forecasts, less for a

coarse-resolution model than a high-resolution model, less for an ensemble mean than a probabilistic forecast, less for prediction than for data assimilation purposes, and so on. It is also possible that there might be a discrepancy between practice and theory, e.g., a finite size ensemble might work sufficiently well in practice, but a huge or even infinite size ensemble might be required in theory. Therefore, there must be a compromise between efficiency and elegance (Mullen and Buizza 2002). Currently, the memberships of the NCEP and CMC ensembles (both global and regional) are around 20–25 members, while the ECMWF global EPS has 51 members. The ensemble data assimilation has normally around 80 members.

6 Ending Remarks

In this paper, many methods for creating an ensemble of forecasts are introduced. In reality, one single EPS might not satisfy all needs but multiple EPSs may be needed. These multiple supplementary EPSs could work seamlessly over a wide range of weather, climate, and water prediction problems. Each of these systems may be uniquely designed specifically to address unique problems. For example, an earth simulator-based climate EPS focuses on climate change due to human activity and natural variability. An atmosphere-ocean-coupled climate model-based seasonal EPS focuses on monthly to yearly scales of a dominant weather mode (warm or cold, wet or dry, etc.) in particular in association with large-scale abnormal episodes like El Niño-Southern Oscillation (ENSO). A global atmospheric model-based medium-range EPS focuses on 3–30-day large-scale flow patterns associated with baroclinic instability and serves as an early warming of weather events. A regional model-based short-range EPS focuses on 1–3-day detailed weather events with an emphasis on surface weather elements and sky conditions (clouds) associated with both baroclinic and convective instabilities. A cloud-resolving or convection-allowing model-based storm-scale EPS focuses on 1–24-h details of severe storms, including those associated with convective instability. A microscale ensemble with a model at a few meters of resolution focuses on in-cloud microphysics, turbulence, and planet boundary-layer structures. Specialized EPSs might also be needed for applications such as hurricane, marine and ocean forecasts, dispersion modeling and air quality, and space weather. EPSs at different scales (spatial and temporal) or for different missions need different strategies in perturbing the ICs and model. For example, in hurricane prediction both the environment and vortex (structure and intensity) need to be perturbed (Zhang and Krishnamurti 1999; Cheung and Chan 1999a, b). Finally, an ensemble model for decision-making based on weather forecast uncertainties also needs to be developed to deal with this complex society.

Acknowledgments Ms. Mary Hart is appreciated for her help to improve the readability of the manuscript. We thank Jack Kain and Binbin Zhou for their reviews.

References

- E.A. Aligo, W.A. Gallus, M. Segal, Summer rainfall forecast spread in an ensemble initialized with different soil moisture analyses. *Weather Forecast.* **22**, 299–314 (2007)
- E.A. Aligo, W.A. Gallus Jr., M. Segal, On the impact of WRF model vertical grid resolution on Midwest summer rainfall forecasts. *Weather Forecast.* **24**, 575–594 (2009)
- J.L. Anderson, An ensemble adjustment Kalman filter for data assimilation. *Mon. Weather Rev.* **129**, 2884–2903 (2001)
- J. Berner, F. Doblas-Reyes, T.N. Palmer, G.J. Shutts, A. Weisheimer, Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Phil. Trans. R. Soc. A* **366**, 2561–2579 (2008). <https://doi.org/10.1098/rsta.2008.0033>
- J. Berner, G.J. Shutts, M. Leutbecher, T.N. Palmer, A spectral stochastic kinetic energy backscatter scheme and its impact on flow-dependent predictability in the ECMWF ensemble prediction system. *J. Atmos. Sci.* **66**, 603–626 (2009)
- J. Berner, S.-Y. Ha, J.P. Hacker, A. Fournier, C. Snyder, Model uncertainty in a mesoscale ensemble prediction system: Stochastic versus multi-physics representations. *Mon. Weather Rev.* **139**, 1972–1995 (2011)
- J. Berner, T. Jung, T.N. Palmer, Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *J. Clim.* **25**, 4946–4962 (2012)
- E.G. Birgin, J.M. Martinez, M. Raydan, Nonmonotone spectral projected gradient methods on convex sets. *SIAM J. Optim.* **10**, 1196–1211 (2000)
- C.H. Bishop, Z. Toth, Ensemble transformation and adaptive observation. *J. Atmos. Sci.* **56**, 1748–1765 (1999)
- C.H. Bishop, B.J. Etherton, S. Majumdar, Adaptive sampling with the ensemble transform Kalman filter. Part I: theoretical aspects. *Mon. Weather Rev.* **129**, 420–436 (2001)
- B.T. Blake, J.R. Carley, T.I. Alcott, I. Jankov, M. Pyle, S. Perfater, B. Albright, An adaptive approach for the calculation of ensemble grid-point probabilities. *Weather Forecast.*, submitted **33**, (2018)
- M. Borges, D.L. Hartmann, Barotropic instability and optimal perturbations of observed non-zonal flows. *J. Atmos. Sci.* **49**, 335–354 (1992)
- N.E. Bowler, A. Arribas, K.R. Mylne, K.B. Robertson, S.E. Beare, The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**, 703–722 (2008)
- N.E. Bowler, A. Arribas, S.E. Beare, K.R. Mylne, G.J. Shutts, The local ETKF and SKEB: Upgrade to the MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **135**, 767–776 (2009)
- C. Brankovic, T.N. Palmer, F. Molteni, S. Tibaldi, U. Cubasch, Extended-range predictions with ECMWF models: time-lagged ensemble forecasting. *Q. J. R. Meteorol. Soc.* **116**, 867–912 (2006)
- R. Buizza, T.N. Palmer, The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.* **52**, 1434–1456 (1995)
- R. Buizza, J. Tribbia, F. Molteni, T.N. Palmer, Computation of optimal unstable structures for a numerical weather prediction model. *Tellus* **45A**, 388–407 (1993)
- R. Buizza, M. Miller, T.N. Palmer, Stochastic representation of model uncertainties in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908 (1999)
- M. Charron, G. Pellerin, L. Spacek, P.L. Houtekamer, N. Gagnon, H.L. Mitchell, L. Michelin, Toward random sampling of model error in the Canadian ensemble prediction system. *Mon. Weather Rev.* **138**, 1877–1901 (2010)
- J. Chen, J.-S. Xue, H. Yan, A new initial perturbation method of ensemble mesoscale heavy rain prediction. *Chin. J. Atmos. Sci.* **29**(5), 717–726 (2005). <https://doi.org/10.3878/j.issn.1006-9895.2005.05.05>

- K.W.C. Cheung, J.C.L. Chan, Ensemble forecasting of tropical cyclone motion using a barotropic model. Part I: perturbations of the environment. *Mon. Weather Rev.* **127**, 1229–1243 (1999a)
- K.W.C. Cheung, J.C.L. Chan, Ensemble forecasting of tropical cyclone motion using a barotropic model. Part II: perturbations of the vortex. *Mon. Weather Rev.* **127**, 2617–2640 (1999b)
- A.J. Clark, W.A. Gallus, M. Xue, F. Kong, A comparison of precipitation forecast skill between small convection-allowing and large convection-parameterizing ensembles. *Weather Forecast.* **24**, 1121–1140 (2009)
- A.J. Clark, W.A. Gallus Jr., M. Xue, F. Kong, Convection-allowing and convection-parameterizing ensemble forecasts of a mesoscale convective vortex and associated severe weather environment. *Weather Forecast.* **25**, 1052–1081 (2010). <https://doi.org/10.1175/2010WAF2222390.1>
- A.J. Clark, J.S. Kain, D.J. Stensrud, M. Xue, F. Kong, M.C. Coniglio, K.W. Thomas, Y. Wang, K. Brewster, J. Gao, X. Wang, S.J. Weiss, J. Du, Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon. Weather Rev.* **139**, 1410–1418 (2011). <https://doi.org/10.1175/2010MWR3624.1>
- G.K. Dai, M. Mu, Z.N. Jiang, Relationships between optimal precursors triggering NAO onset and optimally growing initial errors during NAO prediction. *J. Atmos. Sci.* **73**, 293–317 (2016)
- D. Dee, On-line estimation of error covariance parameters for atmospheric data assimilation. *Mon. Weather Rev.* **123**, 1128–1145 (1995)
- L. Delle Monache, F.A. Eckel, D. Rife, B. Nagarajan, K. Searight, Probabilistic weather predictions with an analog ensemble. *Mon. Weather Rev.* **141**, 3498–3516 (2013)
- M. Delle, L.T. Nipen, Y. Liu, G. Roux, R. Stull, Kalman filter and analog schemes to postprocess numerical weather predictions. *Mon. Weather Rev.* **141**, 3498–3516 (2011)
- S.R. Dey, A.G. Leoncini, N.M. Roberts, R.S. Plant, S. Migliorini, A spatial view of ensemble spread in convection permitting ensembles. *Mon. Weather Rev.* **142**, 4091–4107 (2014)
- S.R. Dey, N.M. Roberts, R.S. Plant, S. Migliorini, A new method for characterization and verification of local spatial predictability for convective-scale ensembles. *Q. J. R. Meteorol. Soc.* **142**, 1982–1996 (2016)
- F. Doblas-Reyes, A. Weisheimer, M. Déqué, N. Keenlyside, M. McVean, J.M. Murphy, P. Rogel, D. Smith, T.N. Palmer, Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**, 1538–1559 (2009)
- J. Du, Hybrid ensemble prediction system: a new ensembling approach. Preprints, in *Symposium on the 50th Anniversary of Operational Numerical Weather Prediction* (University of Maryland, College Park, 2004), June 14–17 2004, Am. Meteorol. Soc., CD-ROM (paper p4.2, 5pp). Available online <http://www.emc.ncep.noaa.gov/mmb/SREF/reference.html>
- J. Du, G. Deng, The utility of the transition from deterministic to probabilistic weather forecasts: verification and application of probabilistic forecasts. *Meteorol. Mon.* **36**(12), 10–18 (2010)
- J. Du, G. DiMego, A regime-dependent bias correction approach. in *19th Conference on Probability and Statistics* (New Orleans, 2008), Jan 20–24 2008, paper 3.2
- J. Du, J. Li, Application of ensemble methodology to heavy rain research and prediction. *Adv. Meteorol. Sci. Technol.* **4**(5), 6–20 (2014)
- Du, J., M. S. Tracton, Impact of lateral boundary conditions on regional-model ensemble prediction. in *Research Activities in Atmospheric and Oceanic Modelling*, ed. by H. Ritchie. Report **28**, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-No. 942, (1999), pp. 6.7–6.8
- J. Du, M.S. Tracton, Implementation of a real-time short-range ensemble forecasting system at NCEP: an update. Preprints, in *9th Conference on Mesoscale Processes* (Ft. Lauderdale, 2001), Am. Meteorol. Soc., pp. 355–356
- J. Du, B. Zhou, Ensemble fog prediction, in *Marine Fog: Challenges and Advancements in Observations, Modeling, and Forecasting*, ed. by D. Koracin, C.E. Dorman (Springer, Cham, 2017), pp. 477–509
- J. Du, S.L. Mullen, F. Sanders, Short-range ensemble forecasting of quantitative precipitation. *Mon. Weather Rev.* **125**, 2427–2459 (1997)

- J. Du, G. DiMego, M.S. Tracton, B. Zhou, NCEP short-range ensemble forecasting (SREF) system: multi-IC, multi-model and multi-physics approach. in *Research Activities in Atmospheric and Oceanic Modelling*, ed. by J. Cote, Report 33, CAS/JSC Working Group Numerical Experimentation (WGNE), WMO/TD-No. 1161, (2003), pp. 5.09–5.10
- J. Du, J. McQueen, G. DiMego, T. Black, H. Juang, E. Rogers, B. Ferrier, B. Zhou, Z. Toth, M.S. Tracton, The NOAA/NWS/NCEP short-range ensemble forecast (SREF) system: evaluation of an initial condition vs. multi-model physics ensemble approach. Preprints (CD), in *16th Conference on Numerical Weather Prediction* (Seattle, 2004), Am. Meteorol. Soc
- J. Du, G. Gayno, K. Mitchell, Z. Toth, G. DiMego, Sensitivity study of T2m and precipitation forecasts to initial soil moisture conditions by using NCEP WRF ensemble. 22nd WAF/18th NWP conference (AMS, Park City, 2007)
- J. Du, R. Yu, C. Cui, J. Li, Using a mesoscale ensemble to predict forecast error and perform targeted observation. *Acta Oceanol. Sin.* **33**(1), 83–91 (2014)
- W.S. Duan, Z.H. Huo, An approach to generating mutually independent initial perturbations for ensemble forecasts: orthogonal conditional nonlinear optimal perturbations. *J. Atmos. Sci.* **73**, 997–1014 (2016)
- W.S. Duan, M. Mu, Conditional nonlinear optimal perturbation: applications to stability, sensitivity, and predictability. *Sci. China Ser. D* **52**(7), 883–906 (2009)
- Y. Duan, J. Gong, J. Du, et al., An overview of Beijing 2008 Olympics Research and Development Project (B08RDP). *BMAS* **93**, 1–24 (2012)
- J. Dudhia, A nonhydrostatic version of the Penn State/NCAR mesoscale model: validation tests and simulation of an Atlantic cyclone and cold front. *Mon. Weather Rev.* **121**(5), 1493–1513 (1993)
- E.E. Ebert, Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.* **129**, 2461–2480 (2001)
- W. Ebisuzaki, E. Kalnay, Ensemble experiments with a new lagged average forecasting scheme. *WMO, Research activities in atmospheric and oceanic modeling. Report* **15**, (1991), pp. 6.31–6.32
- F.A. Eckel, L. Delle Monache, A hybrid NWP-analog ensemble. *Mon. Weather Rev.* **144**, 897–911 (2016). <https://doi.org/10.1175/MWR-D-15-0096.1>
- M. Ehrendorfer, J. Tribbia, Optimal prediction of forecast error covariances through singular vectors. *J. Atmos. Sci.* **54**, 286–313 (1997)
- E.S. Epstein, Stochastic-dynamic prediction. *Tellus* **21**, 739–759 (1969)
- R. Errico, D. Baumhefner, Predictability experiments using a high-resolution limited area model. *Mon. Weather Rev.* **115**, 488–504 (1987)
- G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte-Carlo methods to forecast error statistics. *J. Geophys. Res. Oceans* **99**(C5), 10143–10162 (1994)
- B.F. Farrell, The initial growth of disturbances in a baroclinic flow. *J. Atmos. Sci.* **39**(8), 1663–1686 (1982)
- B.F. Farrell, Optimal excitation of neutral rossby waves. *J. Atmos. Sci.* **45**(2), 163–172 (1988)
- B.F. Farrell, Optimal excitation of baroclinic waves. *J. Atmos. Sci.* **46**(9), 1193–1206 (1989)
- J.S. Frederiksen, A.G. Davies, Eddy viscosity and stochastic backscatter parameterizations on the sphere for atmospheric circulation models. *J. Atmos. Sci.* **54**, 2475–2492 (1997)
- G.H. Golub, C.F. Van Loan, *Matrix Computations*, 3rd edn. (John Hopkins, 1996). ISBN:9780008018-5414-9
- E.P. Grimit, C.F. Mass, Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest. *Weather Forecast* **17**, 192–205 (2002)
- J.P. Hacker, S.-Y. Ha, C. Snyder, J. Berner, F.A. Eckel, E. Kuchera, M. Pocernich, S. Rugg, J. Schramm, X. Wang, The U.S. Air Force weather Agency's mesoscale ensemble: scientific description and performance results. *Tellus A* **63**, 625–641 (2011)
- S. Hahn, G. Iaccarino, Towards adaptive vorticity confinement. 47th AIAA Aerospace sciences meeting including the new horizons forum and aerospace exposition, *Aerospace Sciences Meetings*. AIAA (2009)
- T.M. Hamill, J.S. Whitaker, Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Weather Rev.* **134**, 3209–3229 (2006). <https://doi.org/10.1175/MWR3237.1>

- T.M. Hamill, J.S. Whitaker, C. Snyder, Distance dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon. Weather Rev.* **129**, 2776–2790 (2001)
- T.M. Hamill, R. Hagedorn, J.S. Whitaker, Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: Precipitation. *Mon. Weather Rev.* **136**, 2620–2632 (2008). <https://doi.org/10.1175/2007MWR2411.1>
- T.M. Hamill, J.S. Whitaker, D.T. Kleist, P. Pegion, M. Fiorino, S.G. Benjamin, Predictions of 2010's tropical cyclones using the GFS and ensemble-based data assimilation methods. *Mon. Weather Rev.* **139**, 3243–3247 (2011). <https://doi.org/10.1175/MWR-D-11-00079.1>
- T.M. Hamill, M. Scheuerer, G. Bates, Analog probabilistic precipitation forecasts using GEFS reforecasts and climatology-calibrated precipitation analyses. *Mon. Weather Rev.* **143**, 3300–3309 (2015). <https://doi.org/10.1175/MWR-D-15-0004.1>
- R.N. Hoffman, E. Kalnay, Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus* **35A**, 100–118 (1983)
- D. Hou, E. Kalnay, K.K. Droegemeier, Objective verification of the SAMEX'98 ensemble forecasts. *Mon. Weather Rev.* **129**, 73–91 (2001)
- D. Hou, Z. Toth, Y. Zhu, A stochastic parameterization scheme within NCEP global ensemble forecast system. in *18th AMS Conference on Probability and Statistics* (Atlanta, 2006), Jan. 29–Feb. 2 2006. Available on line at http://www.emc.ncep.noaa.gov/gmb/ens/ens_info.html.pdf
- D. Hou, Z. Toth, Y. Zhu, W. Yang, Impact of a stochastic perturbation scheme on NCEP global ensemble forecast system. in *19th AMS Conference on Probability and Statistics* (New Orleans, 2008), 20–24 Jan 2008. Available on line at http://www.emc.ncep.noaa.gov/gmb/ens/ens_info.html.pdf
- P.L. Houtekamer, H.L. Mitchell, Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**, 196–811 (1998)
- P.L. Houtekamer, L. Lefavre, J. Derome, H. Ritchie, H.L. Mitchell, A system simulation approach to ensemble prediction. *Mon. Weather Rev.* **124**, 1225–1242 (1996)
- P.L. Houtekamer, X. Deng, H.L. Mitchell, S.-J. Baek, N. Gagnon, Higher resolution in an operational ensemble Kalman filter. *Mon. Weather Rev.* **142**, 1143–1162 (2014)
- Z.H. Huo, The application of nonlinear optimal perturbation methods in ensemble forecasting. Ph.D. Dissertation, University of Chinese Academy of Sciences (Beijing, 2016), p. 108
- I. Jankov, W.A. Gallus, M. Segal, B. Shaw, S.E. Koch, The impact of different WRF model physical parameterizations and their interactions on warm season MCS rainfall. *Weather Forecast.* **20**, 1048–1060 (2005)
- I. Jankov, W.A. Gallus, M. Segal, S.E. Koch, Influence of initial conditions on the WRF–ARW model QPF response to physical parameterization changes. *Weather Forecast.* **22**, 501–519 (2007)
- Z.N. Jiang, M. Mu, A comparisons study of the methods of conditional nonlinear optimal perturbations and singular vectors in ensemble prediction. *Adv. Atmos. Sci.* **26**, 465–470 (2009)
- A. Johnson, X. Wang, M. Xue, F. Kong, Hierarchical cluster analysis of a convection allowing ensemble during the hazardous weather testbed 2009 spring experiment. Part II: ensemble clustering over the whole experiment period. *Mon. Weather Rev.* **139**, 3694–3710 (2011). <https://doi.org/10.1175/MWR-D-11-00016.1>
- R. Kalman, A new approach to linear filtering and prediction problems. *Trans. ASME Ser. D J. Basic Eng.* **82**, 35–45 (1960)
- E. Kalnay, *Atmospheric modeling, data assimilation and predictability* (Cambridge University Press, 2003). 368pp
- E. Kalnay, A talk at Arakawa Symposium of 2007 American Meteorological Society annual meeting (San Antonio, 2007)
- D.T. Kleist, K. Ide, An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS, part I: system description and 3D-hybrid results. *Mon. Weather Rev.* **143**, 433–451 (2015a). <https://doi.org/10.1175/MWR-D-13-00351.1>
- D.T. Kleist, K. Ide, An OSSE-based evaluation of hybrid variational-ensemble data assimilation for the NCEP GFS, part II: 4D EnVar and hybrid variants. *Mon. Weather Rev.* **143**, 452–470 (2015b). <https://doi.org/10.1175/MWR-D-13-00350.1>

- M.-S. Koo, S.-Y. Hong, Stochastic representation of dynamic model tendency: formulation and preliminary results. *Asia Pac. J. Atmos. Sci.* **50**(4), 497–506 (2014). <https://doi.org/10.1007/s13143-014-0039-0>
- T.N. Krishnamurti, C.M. Kishtawal, T. LaRow, D. Bachiochi, Z. Zhang, C.E. Williford, S. Gadgil, S. Surendran, Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**, 1548–1550 (1999)
- C.E. Leith, Theoretical skill of Monte Carlo forecasts. *Mon. Weather Rev.* **102**, 409–418 (1974)
- J.M. Lewis, Roots of ensemble forecasting. *Mon. Weather Rev.* **133**, 1865–1885 (2005)
- X. Li, M. Charron, L. Spacek, G. Candille, A regional ensemble prediction system based on moist targeted singular vectors and stochastic parameter perturbations. *Mon. Weather Rev.* **136**, 443–462 (2008)
- J. Li, J. Du, Y. Liu, A comparison of initial condition-, multi-physics- and stochastic physics-based ensembles in predicting Beijing “7.21” excessive storm rain event. *Acta. Meteor. Sin.* **73**(1), 50–71 (2015). <https://doi.org/10.11676/qxb2015.008>. (in both Chinese and English)
- J. Li, J. Du, Y. Liu, J. Xu, Similarities and differences in the evolution of ensemble spread using various ensemble perturbation methods including terrain perturbation. *Acta. Meteor. Sin.* **75**(1), 123–146 (2017). <https://doi.org/10.11676/qxb2017.011>. (in both Chinese and English)
- D.C. Liu, J. Nocedal, On the limited memory method for large scale optimization. *Math. Program.* **45**, 503–528 (1989)
- E.N. Lorenz, Deterministic nonperiodic flow. *J. Atmos. Sci.* **20**, 130–141 (1963)
- E.N. Lorenz, A study of the predictability of a 28-variable atmospheric model. *Tellus* **17**, 321–333 (1965)
- E.N. Lorenz, *The Essence of Chaos* (University of Washington Press, Seattle, 1993), 240pp
- E.N. Lorenz, Predictability: a problem partly solved. in *Proceedings of Workshop on Predictability* (ECMWF, Reading, 1996), 18pp
- C. Lu, H. Yuan, B.E. Schwartz, S.G. Benjamin, Short-range numerical weather prediction using time-lagged ensembles. *Weather Forecast.* **22**, 580–595 (2007)
- J. Ma, Y. Zhu, R. Wobus, P. Wang, An effective configuration of ensemble size and horizontal resolution for the NCEP GEFS. *Adv. Atmos. Sci.* **29**(4), 782–794 (2012)
- S.J. Majumdar, C.H. Bishop, B.J. Etherton, Adaptive sampling with ensemble transform Kalman filter. Part II: filed program implementation. *Mon. Weather Rev.* **130**, 1356 (2002)
- A. Martin, V. Homar, L. Fita, J.M. Gutierrez, M.A. Rodriguez, C. Primo, Geometrid vs. classical breeding of vectors: application to hazardous weather in the Western Mediterranean. *Geophys. Res. Abstr.* **9**, European Geosciences Union (2007)
- P.J. Mason, D.J. Thomson, Stochastic backscatter in large-eddy simulations of boundary layers. *J. Fluid Mech.* **242**, 51–78 (1992)
- J.M. McLay, C.H. Bishop, C.A. Reynolds, The ensemble-transform scheme adapted for the generation of stochastic forecast perturbations. *Q. J. R. Meteorol. Soc.* **133**, 1257–1266 (2007)
- J. McLay, C.H. Bishop, C.A. Reynolds, A local formulation of the ensemble transform (ET) analysis perturbation scheme. *Weather and Forecast.* **25**, 985–993 (2010). <https://doi.org/10.1175/2010WAF2222359.1>
- M.P. Mittermaier, Improving short-range high-resolution model precipitation forecast skill using time-lagged ensembles. *Q. J. R. Meteorol. Soc.* **133**, 1487–1500 (2007). <https://doi.org/10.1002/qj.135>
- F. Molteni, T.N. Palmer, Predictability and finite-time instability of the northern winter circulation. *Q. J. R. Meteorol. Soc.* **119**, 269–298 (1993)
- F. Molteni, R. Buizza, T.N. Palmer, T. Petroliagis, The ECMWF ensemble prediction system: methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996)
- M. Mu, Z.N. Jiang, A new approach to the generation of initial perturbations for ensemble prediction: conditional nonlinear optimal perturbation. *Chin. Sci. Bull.* **53**(13), 2062–2068 (2008)
- M. Mu, W.S. Duan, B. Wang, Conditional nonlinear optimal perturbation and its applications. *Nonlin. Process. Geophys.* **10**, 493–501 (2003)

- M. Mu, W.S. Duan, D.K. Chen, W.D. Yu, Target observations for improving initialization of high-impact ocean-atmospheric environmental events forecasting. *Natl. Sci. Rev.* **2**, 226–236 (2015)
- S.L. Mullen, D.P. Baumhefner, Monte Carlo simulation of explosive cyclogenesis. *Mon. Weather Rev.* **122**, 1548–1567 (1994)
- S.L. Mullen, R. Buizza, The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. *Weather Forecast.* **17**, 173–191 (2002)
- S.L. Mullen, J. Du, Monte Carlo forecasts of explosive cyclogenesis with a limited-area, mesoscale model. *Preprints, 10th Conference on Numerical Weather Prediction* (Portland, 1994), July 18–22 1994, Am. Meteorol. Soc., pp. 638–640
- S.L. Mullen, J. Du, F. Sanders, The dependence of ensemble dispersion on analysis forecast system: implications to short-range ensemble forecasting of precipitation. *Mon. Weather Rev.* **127**, 1674–1686 (1999)
- National Research Council, *Completing the Forecasts: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts* (National Academies Press, Washington, DC, 2006), 124pp. <https://doi.org/10.17226/11699>
- P. Nutter, D. Stensrud, M. Xue, Effects of coarsely resolved and temporally interpolated lateral boundary conditions on the dispersion of limited-area ensemble forecasts. *Mon. Weather Rev.* **132**, 2358–2377 (2004a)
- P. Nutter, M. Xue, D. Stensrud, Application of lateral boundary condition perturbations to help restore dispersion in limited-area ensemble forecasts. *Mon. Weather Rev.* **132**, 2378–2390 (2004b)
- T. Palmer, R. Hagedorn (eds.), Predictability of weather and climate (Cambridge University Press, New York, 2006), 718pp
- T.N. Palmer, R. Buizza, M. Leutbecher, R. Hagedorn, T. Jung, M. Rodwell, F. Virat, J. Berner, E. Hagel, A. Lawrence, F. Pappenberger, Y-Y. Park, L. van Bremen, I. Gilmour, L. Smith, The ECMWF Ensemble Prediction System: recent and on-going developments. A paper presented at the 36th Session of the ECMWF Scientific Advisory Committee. ECMWF Research Department Technical Memorandum Note, vol. **540**. (2007). Available from ECMWF, Shinfield Park, Reading RG2-9AX, UK
- T.N. Palmer, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer, A Weisheimer, Stochastic Parameterization and Model Uncertainty. ECMWF Research Department Technical Memorandum, vol. **598**, (2009), p. 42. Available from ECMWF, Shinfield Park, Reading RG2-9AX, UK. <http://www.ecmwf.int/publications/>
- M.J.D. Powell, VMCWD: A FORTRAN subroutine for constrained optimization. DAMTP Report 1982/NA4 (University of Cambridge, UK, 1982)
- X. Qiao, S. Wang, J. Min, A stochastic perturbed parameterization tendency scheme for diffusion (SPPTD) and its application to an idealized supercell simulation. *Mon. Weather Rev.* **145**(6), 2119–2139 (2017). <https://doi.org/10.1175/MWR-D-16-0307>
- D.S. Richardson, Skill and relative economic value of the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **126**, 649–668 (2000)
- D.S. Richardson, Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Q. J. R. Meteorol Soc.* **127**, 2473–2489 (2001)
- N.M. Roberts, H.W. Lean, Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Weather Rev.* **136**, 78–97 (2008). <https://doi.org/10.1175/2007MWR2123.1>
- S. Roquelaure, T. Bergot, A local ensemble prediction system (L-EPS) for fog and low clouds: construction, Bayesian model averaging calibration and validation. *J. Appl. Meteorol. Clim.* **47**, 3072–3088 (2008)
- S. Saha, S. Nadiga, C. Thiaw, J. Wang, W. Wang, Q. Zhang, H.M. Van den Dool, H.L. Pan, S. Moorthi, D. Behringer, D. Stokes, M. Peña, S. Lord, G. White, W. Ebisuzaki, P. Peng, P. Xie, The NCEP climate forecast system. *J. Clim.* **19**, 3483–3517 (2006)

- C. Sanchez, K.D. Williams, G.J. Shutts, R.E. McDonald, T.J. Hinton, C.A. Senior, N. Wood, Toward the development of a robust model hierarchy: investigation of dynamical limitations at low resolution and possible solutions. *Q. J. R. Meteorol. Soc.* **139**, 75–84 (2013). <https://doi.org/10.1002/qj.1971>
- C.S. Schwartz et al., Toward improved convection-allowing ensembles: model physics sensitivities and optimizing probabilistic guidance with small ensemble membership. *Weather Forecast.* **25**, 263–280 (2010). <https://doi.org/10.1175/2009WAF2222267.1>
- C. Schwartz, R. Sobash, Generating probabilistic forecasts from convection-allowing ensembles using neighborhood approaches: a review and recommendations. *Mon. Weather Rev.* **145**, 3397–3418 (2017). <https://doi.org/10.1175/MWR-D-16-0400.1>
- G. Shutts, *A Stochastic Kinetic Energy Backscatter Algorithm for Use in Ensemble Prediction Systems*, Technical Memorandum, vol 449 (ECMWF, Reading, 2004)
- G. Shutts, A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **131**, 3079–3102 (2005)
- G. Shutts, T. Allen, Sub-gridscale parameterization from the perspective of a computer games animator. *Atmos. Sci. Lett.* **8**(4), 85–92 (2007). <https://doi.org/10.1002/asl.157>
- G. Shutts, M. Leutheuer, A. Weisheimer, T. Stockdale, L. Isaksen, M. Bonavita, Representing model uncertainty: stochastic parameterizations at ECMWF. *ECMWF Newslet.* **129**, 19–24 (2011)
- W.C. Skamarock et al., *A Description of the Advanced Research WRF Version 3*, NCAR Technical Note, vol NCAR/TN-475+STR (National Center for Atmospheric Research, Boulder, 2008). <https://doi.org/10.5065/D68S4MVH>
- U. Stein, P. Alpert, Factor separation in numerical simulations. *J. Atmos. Sci.* **50**, 2107–2115 (1993)
- J. Steinhoff, D. Underhill, Modification of the Euler equations for “vorticity confinement”: Application to the computation of interacting vortex rings. *Phys Fluid* **6**, 2738 (1994). <https://doi.org/10.1063/1.868164>
- D.J. Stensrud, H.E. Brooks, J. Du, M.S. Tracton, E. Rogers, Using ensembles for short-range forecasting. *Mon. Weather Rev.* **127**, 433–446 (1999)
- C. Sutton, T.M. Hamill, T.T. Warner, Will perturbing soil moisture improve warm-season ensemble forecasts? A proof of concept. *Mon. Weather Rev.* **134**, 3174–3189 (2006)
- O. Talagrand, R. Vautard, B. Strauss, Evaluation of probabilistic prediction systems. in *Proceedings, ECMWF Workshop on Predictability* (ECMWF, 1997). Available from ECMWF, Shinfield Park, Reading, Berkshire RG2 9AX, United Kingdom, pp. 1–25
- S. Tang, D. Wang, J. Du, J. Zhou, The experiment of hybrid ensemble forecast approach in short-range forecast for South China rainstorm. *J. Appl. Meteorol. Sci.* **26**(6), 669–679 (2015)
- W.J. Tennant, G.J. Shutts, A. Arribas, S.A. Thompson, Using a stochastic kinetic energy backscatter scheme to improve MOGREPS probabilistic forecast skill. *Mon. Weather Rev.* **139**, 1190–1206 (2011). <https://doi.org/10.1175/2010MWR3430.1>
- S.E. Theis, A. Hense, U. Damrath, Probabilistic precipitation forecasts from a deterministic model: a pragmatic approach. *Meteorol. Appl.* **12**, 257–268 (2005). <https://doi.org/10.1017/S1350482705001763>
- P.D. Thompson, Uncertainty of initial state as a factor in the predictability of large scale atmospheric flow patterns. *Tellus* **9**, 275–295 (1957)
- Z. Toth, E. Kalnay, Ensemble forecasting at NCEP: the generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**, 2317–2330 (1993)
- Z. Toth, E. Kalnay, Ensemble forecasting at NCEP: the breeding method. *Mon. Weather Rev.* **125**, 3297–3318 (1997)
- M.S. Tracton, E. Kalnay, Ensemble forecasting at NMC: practical aspects. *Weather Forecast.* **8**, 379–398 (1993)
- M.S. Tracton, J. Du, Z. Toth, H. Juang, Short-range ensemble forecasting (SREF) at NCEP/EMC. Preprints, in *12th Conference on Numerical Weather Prediction* (Phoenix, 1998), Am. Meteorol. Soc. pp. 269–272

- X. Wang, C.H. Bishop, A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.* **60**, 1140–1158 (2003)
- X. Wang, C.H. Bishop, S.J. Julier, Which is better, an ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Weather Rev.* **132**, 1590–1605 (2004)
- X. Wang, T.M. Hamill, J.S. Whitaker, C.H. Bishop, A comparison of hybrid ensemble transform Kalman filter-optimal interpolation and ensemble square-root filter analysis schemes. *Mon. Weather Rev.* **135**, 1055–1076 (2007)
- X. Wang, D. Barker, C. Snyder, T.M. Hamill, A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part I: observing system simulation experiment. *Mon. Wea. Rev.* **136**, 5116–5131 (2008a). <https://doi.org/10.1175/2008MWR2444.1>
- X. Wang, D. Barker, C. Snyder, T.M. Hamill, A hybrid ETKF-3DVAR data assimilation scheme for the WRF model. Part II: real observation experiments. *Mon. Wea. Rev.* **136**, 5132–5147 (2008b). <https://doi.org/10.1175/2008MWR2445.1>
- X. Wang, T.M. Hamill, J.S. Whitaker, C.H. Bishop, A comparison of the hybrid and EnSRF analysis schemes in the presence of model error due to unresolved scales. *Mon. Wea. Rev.* **137**, 3219–3232 (2009). <https://doi.org/10.1175/2009MWR2923.1>
- D. Wang, J. Du, C. Liu, Recognizing and dealing with uncertainty in weather-related forecasts. *Meteorol. Mon.* **37**(4), 385–392 (2011)
- X. Wang, D. Parrish, D. Kleist, J.S. Whitaker, GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP global forecast system: single resolution experiments. *Mon. Weather Rev.* **141**, 4098–4117 (2013). <https://doi.org/10.1175/MWR-D-12-00141.1>
- J. Wang, J. Chen, J. Du, Y. Zhang, G. Deng, Sensitivity of ensemble forecast verification to model bias. *Mon. Weather Rev.*, in press **146**, 781–796 (2018). <https://doi.org/10.1175/MWR-D-17-0223.1>
- T.T. Warner, R.A. Perterson, R.E. Treadon, A tutorial on lateral boundary conditions as a basis and potential serious limitation to regional numerical weather prediction. *Bull. Am. Meteorol. Soc.* **78**, 2599–2617 (1997)
- A. Weaver, P. Courtier, Correlation modeling on the sphere using a generalized diffusion equation. *Q. J. R. Meteorol. Soc.* **127**, 1815–1846 (2001). <https://doi.org/10.1002/qj.49712757518>
- M. Wei, Z. Toth, R. Wobus, Y. Zhu, C. Bishop, X. Wang, Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus A* **58**, 28–44 (2006)
- M. Wei, Z. Toth, R. Wobus, Y. Zhu, Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A* **60**, 62–79 (2008)
- M. Wei, C. Rowley, P. Martin, C.N. Barron, G. Jacobs, The US Navy's RELO ensemble prediction system and its performance in the Gulf of Mexico. *Q. J. R. Meteorol. Soc.* **140**, 1129–1149 (2014). <https://doi.org/10.1002/qj.2199>
- A. Weisheimer, T.N. Palmer, F.J. Doblas-Reyes, Assessment of representations of model uncertainty in monthly and seasonal forecast ensembles. *Geophys. Res. Lett. (Climate)* **38**, L16703 (2011). <https://doi.org/10.1029/2011GL048123>
- J. Whitaker, P. Piegion, T. Hamill, Representing model uncertainty in data assimilation (using ensembles), EMC/NCEP/NOAA seminar (2013). Available at <http://www.emc.ncep.noaa.gov/seminars/index.html>
- R. Wobus, E. Kalnay, Three years of operational prediction of forecast skill. *Mon. Weather Rev.* **123**, 2132–2148 (1995)
- Z. Zhang, T.N. Krishnamurti, A perturbation method for hurricane ensemble prediction. *Mon. Weather Rev.* **127**, 447–469 (1999)
- B. Zhou, J. Du, Fog prediction from a multimodel mesoscale ensemble prediction system. *Weather Forecast.* **25**, 303–322 (2010)
- B. Zhou, J. Du, G. DiMego, Introduction to the NCEP very short range ensemble forecast system (VSREF). in *14th Conference on Aviation, Range, and Aerospace, 90th AMS Annual Meetings* (Atlanta, 2010), 17–21 2010. Available at <http://www.emc.ncep.noaa.gov/mmb/SREF/VSREF-2010-AMS-J12.3.pdf>



Major Operational Ensemble Prediction Systems (EPS) and the Future of EPS

Roberto Buizza, Jun Du, Zoltan Toth, and Dingchen Hou

Contents

1	Operational, Global Medium-Range Ensembles (OG-ENS)	152
1.1	From Predictability Theory to Operational Ensemble Forecasting	153
1.2	The WMO THORPEX/TIGGE Project	154
2	The Nine TIGGE Operational, Global Medium-Range Ensembles	154
2.1	BMRC Australia	158
2.2	CMA China	159
2.3	CPTEC Brazil	160
2.4	ECMWF Europe	160
2.5	JMA Japan	163
2.6	KMA Korea	164
2.7	MSC Canada	165
2.8	NCEP United States of America	166
2.9	UKMO United Kingdom	168
3	Average Performance of the Operational, Global, Medium-Range Ensembles	169
3.1	Ensemble-Mean and Spread for One Specific Case (10 January 2013)	170
3.2	Average Performance over the Tropics and the Northern Hemisphere Extra-Tropics	174

R. Buizza (✉)

European Centre for Medium Range Weather Forecasts, Reading, UK

e-mail: roberto.buizza@ecmwf.int

J. Du

Environmental Modeling Center/ National Centers for Environmental Prediction (NCEP), NOAA, College Park, MA, USA

e-mail: jun.du@noaa.gov

Z. Toth

Global Systems Division, Earth System Research Laboratory, National Oceanic and Atmospheric Administration/OAR, Boulder, CO, USA

e-mail: zoltan.toth@noaa.gov

D. Hou

SAIC at NOAA/NWS/NCEP/EMC, Camp Springs, MA, USA

e-mail: dingchen.hou@noaa.gov

4 The Future of Ensemble Techniques	180
5 Final Remarks on Operational Global Medium-Range Ensembles	190
References	191

Abstract

Since the early 1990s, ensemble methods have been increasingly used to address predictability issues, and provide estimates of forecast uncertainties, e.g., in the form of a range of forecast scenarios or of probabilities of occurrence of weather events. Although there is an overall agreement on the main objectives of ensemble-based, probabilistic prediction, different methods have been followed to develop ensemble systems. In this chapter, we will review the methods followed at the major weather prediction centres to develop global ensembles, and we will highlight the links between the method followed and the ensemble forecast performance. The material presented in this chapter is based on the operational, global, medium-range ensembles operational at the time of writing (2014).

Keywords

Ensemble prediction · Predictability · Probabilistic forecasting · Forecast uncertainty

In this chapter we will review the existing, operational global ensemble prediction systems, which can be accessed via the TIGGE database. The chapter is organized in four sections:

- In Sect. 1, we will start by defining the meaning of an “operational global medium-range ensemble” (OG-ENS). More specifically, in Subsection 1.1 we will briefly review how these ensembles have been designed, linking their key characteristics to predictability theory. In Subsection 1.2 we will introduce the TIGGE archive, which makes all these OG-ENS accessible to the whole scientific community.
- In Sect. 2, we will then review the key characteristics of the OG-ENS, with each subsection focussing on each single ensemble.
- In Sect. 3 we will compare how the OG-ENS introduced in Sect. 2 perform, looking both at one specific case and at some statistical results.
- In Sect. 4, finally, we will discuss areas of research and development and illustrate how we think the ensembles will evolve in the forthcoming future.
- In Sect. 5 we will make few final remarks and conclude the chapter.

1 Operational, Global Medium-Range Ensembles (OG-ENS)

Before discussing the details of the operational, global, medium-range ensembles, let’s clarify the meaning of these three words, “operational,” “global,” and “medium-range”:

- *Operational* means that these ensembles produce forecasts on a daily basis.
- *Global* means that their forecasts cover the whole globe.
- *Medium-range* means that they provide forecasts at least up to 7 days, with some of them extending to 2 weeks or even to 1 month.

The OG-ENS ensembles are the topics of discussion in this chapter.

1.1 From Predictability Theory to Operational Ensemble Forecasting

Operational, global medium-range ensemble prediction begins in November/December 1992, when the National Centers for Environmental Prediction (NCEP, Toth and Kalnay 1993, 1997) and the European Centre for Medium-Range Weather Forecast (ECMWF, Palmer et al. 1993; Buizza and Palmer 1995; Molteni et al. 1996; Buizza et al. 2007) started producing global ensemble predictions as part of their operational products. ECMWF and NCEP were followed by the Meteorological Service of Canada (MSC, Houtekamer et al. 1996), who implemented its operational global ensemble prediction system in 1995. Following their example, six other centers started producing global ensemble prediction systems operationally soon afterward. These are the Australian Bureau of Meteorology (BMRC, Bourke et al. 1995, 2004), the Chinese Meteorological Administration (CMA), the Brazilian Center for Weather Prediction and Climate Studies (Centro de Previsao de Tempo e Estudos Climaticos, CPTEC), the Japanese Meteorological Administration (JMA), the Korean Meteorological Administration (KMA, Goo et al. 2003), and the UK Met Office (UKMO). This brings the total number of the operational OG-ENS to nine.

It is worth mentioning that there are many other centers (e.g., Meteo France, Met Norway, the COSMO Consortium established by the German, Greece, Italian, Polish, and Swiss National Meteorological Services, and the Spanish Instituto Nacional de Meteorologia) who have been running or testing short-range regional ensemble prediction systems. The discussion of these systems is beyond the scope of this chapter.

All nine OG-ENS have been designed to represent the effects on forecasts of observation uncertainties, imperfect boundary conditions and data assimilation assumptions (e.g., due to the data assimilation methods and underlying statistical assumptions), and model uncertainties (e.g., due to a lack of resolution, simplified parameterization of physical processes, effect of unresolved processes).

Before describing how the OG-ENS have been designed, let's briefly remind the key objectives of the TIGGE project, a WMO initiative that provides a unique database where all these ensembles can be accessed in quasi-real time (with quasi-real time, we mean that the OG-ENS are available 48 h after production; this delayed access was agreed by the nine contributing centers when TIGGE was established).

1.2 The WMO THORPEX/TIGGE Project

THORPEX is a World Meteorological Organization (WMO) World Weather Research Programme (WWRP) aiming to accelerate the improvements in the accuracy of 1-day to 2-week high-impact weather forecasts. TIGGE is the THORPEX Interactive Grand Global Ensemble (see Bougeault et al. 2010 and TIGGE and THORPEX in the list of references).

TIGGE started in 2004 with the objectives to develop a deeper understanding of the contribution of observation, initial, and model uncertainties to forecast error and to investigate new methods of combining ensembles from different sources and of correcting systematic errors. Since its inception, three centers have been acting as data collectors: ECMWF, CMA, and NCAR. At ECMWF, for example, users can access the ensemble forecast data from October 2006 for scientific research, by registering as a *tiggeuser* at the ECMWF TIGGE portal (<http://tigge.ecmwf.int>). The TIGGE archive contains ensembles from the nine operational, global medium-range ensembles mentioned above (BMRC, CMA, MSC, CPTEC, ECMWF, JMA, KMA, NCEP, and UKMO) and the Meteo France (MF) ensemble, which is also global but provides forecasts only up to 84 h.

The TIGGE archive has been an extremely valuable resource for research aiming to compare the performance of single ensemble prediction systems, identify their strengths and weaknesses, and to assess the potential value of combining different ensembles to generate multi-model/multi-analysis grand global ensemble products. As documented by the list of publications based on TIGGE data, they have proven to be very valuable also for research and development in the fields covered by the two other THORPEX working groups – Data Assimilation and Observing System (DAOS) and Predictability and Dynamical Processes (PDP) – and other WWRP working groups, including the Working Group on Numerical Experimentation (WGNE), Socio-Economic Research and Applications (SERA), and the Joint Working Group on Forecast Verification Research (JWGFVR).

2 The Nine TIGGE Operational, Global Medium-Range Ensembles

We said earlier that the nine OG-ENSSs have all been designed to represent the effects on forecasts of observation uncertainties, imperfect boundary conditions and data assimilation assumptions, and model uncertainties. To facilitate the comparison between the different approaches, we can group these sources of forecast errors into two broad classes, namely, initial and model uncertainties. The classes are broad and overlay, since it is impossible to separate clearly between initial and model uncertainties, given that initial conditions are constructed using model-based, data assimilation procedures. Thus, errors that are sometimes defined as “initial condition errors” may actually be partly due to model errors.

How similar, or different, are the strategies followed by the nine operational, global, medium-range ensembles to simulate initial and model uncertainties?

Table 1 lists the main characteristics of the nine ensembles, which together provide 550 forecasts every day, spanning a forecast range between 8 and 16 days, with grid spacing between 200 and 35 km. All ensembles simulate initial uncertainties, with none following exactly the same methodology. Broadly speaking, three OG-ENS use singular vectors (SVs; BMRC, ECMWF, and JMA), one uses bred vectors (BVs; CMA), one uses an ensemble Kalman filter (MSC), one uses perturbations computed using empirical orthogonal functions (EOFs; CPTEC), and three use initial perturbations computed by rotating and/or rescaling bred vectors (BVs; KMA, NCEP, and UKMO). Model uncertainties are simulated in six OG-ENS using different methods, as will be described later. In terms of spatial resolution, the OG-ENS cover a very wide range of scales: horizontally from about 200 km (CPTEC) to about 35 km (ECMWF, during the first 10 forecast days), and vertically from 19 vertical levels spanning the atmosphere up to 10 hPa (about 30 km), to 91 levels up to 0.1 hPa (about 80 km). Forecast length varies from 10 to 16 days. In terms of membership, each ensemble run includes at least 14 perturbed members, with maximum membership set to 50. Two centers update their ensembles four times a day, six run them twice, and one only runs it once a day.

Figure 1 visualizes the key characteristics of the nine OG-ENS along four key computing resources cost drivers: horizontal resolution squared (HR^2 , with HR in km; this takes into account, e.g., that a doubling of the resolution increases the number of grid points by a factor of 4), number of vertical levels (LEV), number of members per day (#M), and forecast length (FCD, in days).

These key cost drivers can be used to get an estimate of the relative cost of each ensemble. Values along these drivers have been normalized with respect to the most costly values, i.e., 35 for HR , 91 for LEV, 102 for #M, and 16 for FCD. The relative amount of resources required to produce the OG-ENS per day, r_n , has been defined as the product of the relative cost drivers:

$$r_n = (HR_n/35)^2 \cdot (LEV_n/91) \cdot (\#M_n/102) \cdot (FCD_n/16) \quad (1)$$

The ECMWF OG-ENS is the one requiring the highest amount of resources to be produced, due its highest resolution and largest membership, with the KMA being the second one (Fig. 1). More specifically, compared to the ECMWF OG-ENS, the KMA OG-ENS requires about 50%, the JMA and the UKMO OG-ENSs about 17%, the MSC OG-ENS and the NCEP OG-ENS about 6%, and the others less than 5%. These relative cost values should be taken into consideration when comparing the performance of the OG-ENS (see Sect. 3).

Generally speaking, each forecast of the TIGGE ensembles is given by the numerical integration of the model equations used at each center to describe the earth-system model. In other words, each single T-hour forecast starting at day d is given by the time integration of the model equations from initial time 0 to time T :

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_j(t) + P_j(t) + dP_j(t)] dt \quad (2)$$

Table 1 Main characteristics of the nine TIGGE OG-ENS, listed in alphabetic order by the production center (column 1): initial uncertainty method (column 2), model uncertainty simulation (Y/N, column 3), truncation and approximate horizontal resolution (column 4), number of vertical levels and top of the atmosphere in hPa (column 5), forecast length in days (column 6), number of perturbed members for each run (column 7), total number of members (including the control forecast) per day (column 8), and date since when data have been stored in TIGGE (column 9). Note that the BMRC OG-ENS data are available only up to the end of 2010, when BMRC stopped producing global ensemble forecasts

Centre	Initial unc. method (area)	Model unc.	Truncation (degrees, km)	# Vert Lev (TOA, hPa)	Fest length (d)	# pert	# runs	# mem	In TIGGE since
BMRC (AU)	SV(NH,SH)	NO	TL119 (1.5°; 210 km)	19 (10.0)	10	32	2 (00/12)	66	Sep-07/Jul- 10
CMA (CHI)	BV(globe)	NO	T213 (0.56°; 70 km)	31 (10.0)	10	14	2 (00/12)	30	May-07
CPTEC (BR)	EOF(40S:30N)	NO	T126 (0.94°, 120 km)	28 (0.1)	15	14	2 (00/12)	30	Feb-08
ECMWF (EU)	SV(NH, SH, TC) + EDA (globe)	YES	TL639 (0.28°; 35 km) TL319 (0.56°; 70 km)	91 (0.1) 15/32	0–10	50	2 (00/12)	102	Oct-06
JMA (JAP)	SV(NH, TR, SH)	YES	TL479 (0.38°; 50 km)	60 (0.1)	11	25	2 (00/12)	52	Aug-11
KMA (KOR)	ETKF(globe)	YES	N320 (0.35°; 40 km)	70 (0.1)	10	23	4 (00/06/12/ 18)	96	Dec-07
MSC (CAN)	EnKF(globe)	YES	600 × 300 (0.6°, 75 km)	40 (2.0)	16/32	20	2 (00/12)	42	Oct-07
NCEP (USA)	ETR(globe)	YES	T254 (0.70°; 90 km) T190 (0.95°; 120 km)	28 (2.7) 8–10	0–8	20	4 (00/06/ 12/18)	84	Mar-07
UKMO (UK)	ETKF(globe)	YES	N216 (0.45°; 60 km)	70 (0.1)	15	23	2 (00/12)	48	Oct-06

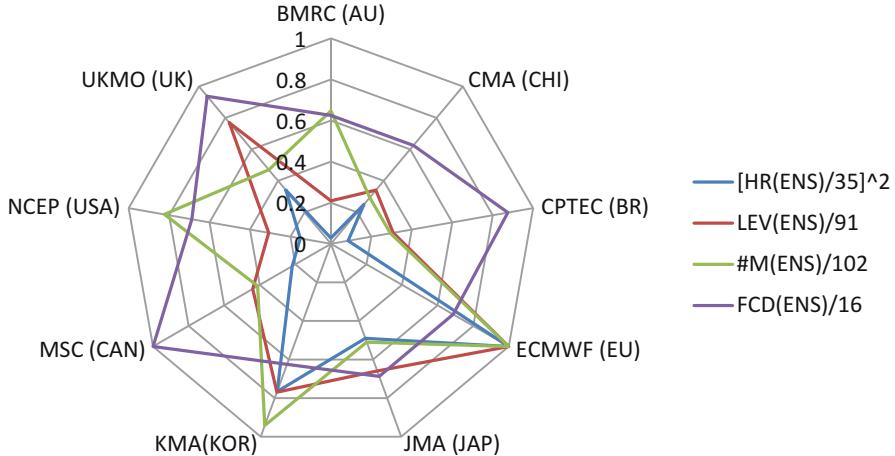


Fig. 1 Comparison of four key resource drivers of the nine OG-ENS: squared horizontal resolution (HR^2), number of vertical levels (LEV), number of perturbed members per day (#M), and forecast length (FCD, in days). Values for each driver have been normalized with respect to the most costly value (35 km for HR, 91 for LEV, 102 for #M, and 16 for FCD)

where A_j is the tendency due to the adiabatic processes (say advection, Coriolis force, pressure gradient force), P_j is the tendency due to the parameterized physical processes (say convection, radiation, turbulence, . . .), and dP_j represents the tendency due to unresolved processes.

In the MSC OG-ENS, each forecast integration starts from initial conditions defined by an independent data assimilation procedure:

$$e_j(d,0) = F[e_j(d - T_A; T_A), o_j(d - T_A, d)] \quad (3)$$

where $F[\dots]$ represents the result of the data assimilation process of merging a model first guess and observations spanning a time period T_A (the window covered by the data assimilation process) from $(d - T_A)$ to d . The first guess $e_j(d - T_A; T_A)$ is given by the T_A -hour time integration of the model equations from $(d - T_A)$ to d . The data assimilation process uses observations $o_j(d - T_A, d)$ relative to the time period from $(d - T_A)$ to d . More precisely, as it will be discussed in Sect. 2.3, each perturbed initial condition is selected among the 192 members of their ensemble Kalman filter.

In all the other OG-ENS, each forecast integration starts from initial conditions defined by adding a perturbation to an unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + de_j(d,0) \quad (4)$$

$$e_0(d,0) = F[e_0(d - T_A; T_A), o_0(d - T_A, d)] \quad (5)$$

where the unperturbed initial conditions are defined by the data assimilation process spanning the time period T_A . The initial perturbations $de_j(d, 0)$ are defined in different ways in each OG-ENS.

Equations (2, 3, 4, and 5) provide a nice unified description of how the j -th member of the 550 forecasts produced every day is generated, by integrating model equations starting from the perturbed initial conditions $e_j(d, 0)$ with tendencies defined by the j -th model (to be more precise, one should say that at each center, a forecast is given by the numerical integration of the sets of equations assumed to describe the earth-system at that specific center; the discussion of how similar or different these equations are is beyond the scope of this chapter).

2.1 BMRC Australia

The BMRC OG-ENS started production in July 2001 and stopped in July 2010, when BMRC decided to adopt the UK data assimilation and forecasting system. BMRC data are available in TIGGE from July 2007 to July 2010. They are planning to restart their medium-range ensemble production in 2015, when the migration from the BMRC to the newly adopted UK Unified Model environment has been completed.

In 2010, the BMRC OG-ENS comprised 33 members, 1 unperturbed and 32 perturbed ones (Bourke et al. 1995, 2004). Forecasts were run twice a day up to 10 days. The system had a spectral triangular truncation $T_L119L19$ (about 1.5° , 160 km in physical space, and 19 vertical levels), with the top of the atmosphere at 10 hPa. The forecast model included only a description of land and atmospheric processes (no wave or ocean model component was used).

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the BMRC OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_0(t)]dt \quad (6)$$

where A_0 and P_0 represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called with the same parameters, and no model error scheme) and the forecast length T is 10 days.

The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + de_j(d,0) \quad (7)$$

The unperturbed initial conditions were produced by interpolating to the ensemble resolution the initial conditions defined by the BMRC three-dimensional multivariate statistical interpolation scheme. The initial perturbations were defined by

T42L19 singular vectors (SVs). The individual perturbations were obtained as linear combinations of the singular vectors through an orthogonal phase space rotation followed by an amplitude scaling factor, as was done in the original ECMWF OG-ENS (Molteni et al. 1996). The ensemble did not simulate model uncertainties.

Thus, the initial perturbations were defined by a linear combination of SVs:

$$de_j(d,0) = \sum_{k=1}^{32} \alpha_{j,k} SV_k \quad (8)$$

The SVs were computed only over the Southern Hemisphere (SH, for all grid points with latitude $\lambda < 20^\circ\text{S}$) and scaled to have amplitude comparable to analysis error estimates.

2.2 CMA China

The CMA OG-ENS started production in 2001. In 2001 it was upgraded to a T213L31 resolution (about 0.56° , 70 km in physical space, and 31 vertical levels). CMA OG-ENS data have been available in the TIGGE archive since May 2007.

At the time of writing (May 2014), the CMA OG-ENS comprises 15 members, 1 unperturbed and 14 perturbed members (Su et al. 2014). Forecasts are run twice a day, at 00 and 12 UTC, up to forecast day 10. The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used). The ensemble does not simulate model uncertainties.

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the CMA OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_0(t)] dt \quad (9)$$

where A_0 and P_0 represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called with the same parameters, and no model error scheme) and the forecast length T is 10 days.

The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + de_j(d,0) \quad (10)$$

$$de_j(d,0) = BV_j(d,0) \quad (11)$$

The unperturbed initial conditions are produced by interpolating to the T213L31 analysis to the ensemble resolution. The initial perturbations are defined by bred vectors, as were defined in the original NCEP ensemble (Toth and Kalnay 1997).

2.3 CPTEC Brazil

The CPTEC OG-ENS has been producing ensemble forecasts since 2001. Its data have been available in the TIGGE archive since March 2008.

The CPTEC OG-ENS uses a spectral model, has 15 members (1 unperturbed and 14 perturbed), runs twice a day (at 00 and 12 UTC), and produces forecasts up to day 15. Initial perturbations are defined using empirical orthogonal functions (Coutinho 1999; Zhang and Krishnamurti 1999). The method consist of (a) computing 36-h bred vectors (by adding random perturbations to unperturbed initial conditions and running pairs of 36-h forecasts), (b) constructing a time series of these bred vectors, and (c) performing an empirical orthogonal function (EOF) analysis of this time series to obtain the fastest growing perturbations. These EOF-based perturbations are computed for the region 45°S–30°N and are added to the unperturbed analysis to simulate initial uncertainties.

At the time of writing (May 2014), the CPTEC OG-ENS has a T126L28 resolution, with a corresponding 0.94° grid (about 120 km) and 28 vertical levels up to 0.1 hPa. The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used). The ensemble does not simulate model uncertainties.

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the CPTEC OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_0(t)] dt \quad (12)$$

where A_0 and P_0 represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called with the same parameters, and no model error scheme) and the forecast length T is 15 days.

The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + de_j(d,0) \quad (13)$$

$$de_j(d,0) = EOF_j(d,0) \quad (14)$$

The unperturbed initial conditions are defined by the NCEP operational analysis, interpolated to the CPTEC resolution.

2.4 ECMWF Europe

The ECMWF OG-ENS has been producing ensemble forecasts from 19 November 1992. Its data have been available in the TIGGE archive since October 2006.

At the time of writing (May 2014), the ECMWF OG-ENS comprises an ensemble of 51 forecasts, 1 unperturbed and 50 perturbed ones. Forecasts are run with a variable resolution (Buizza et al. 2007): T_L639L91 (spectral triangular truncation T639 with a linear grid, which corresponds to about 35 km spacing in physical space and 91 vertical levels) during the first 10 days and T_L319L91 (i.e., about 70 km spacing) thereafter. Forecasts are run twice a day, with initial times at 00 and 12 UTC, up to 15 days; at 00 UTC on Mondays and Thursdays, the forecasts are extended to forecast day 32. The forecasts are coupled to the WAM wave model (Janssen et al. 2005, 2013) with a 55 km resolution and 24 directions and 30 frequencies up to day 10 and 12 directions and 25 frequencies afterward.

The latest update of the ensemble configuration was implemented in November 2013, when the top of the atmosphere was moved to 0.01 hPa and the number of vertical levels was increased from 62 to 91 and ENS forecasts have been coupled to a dynamical ocean model from initial time. The ocean model, NEMO (the Nucleus for European Ocean Modelling), uses the ORCA100z42 grid with a 1° horizontal resolution and 42 vertical layers. NEMO is a state-of-the-art modelling framework for oceanographic research, operational oceanography seasonal forecast, and climate studies, developed by the NEMO Consortium (<http://www.nemo-ocean.eu/>).

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the ECMWF OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_0(t) + dP_j(t)] dt \quad (15)$$

where A_0 and P_0 represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called with the same parameters), dP_j represents the model uncertainty simulated using the SPPT and SKEB stochastic schemes, and the forecast length T is 15 or 32 days. For the atmosphere, the initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + de_j(d,0) \quad (16)$$

The unperturbed analysis is given by the ECMWF high-resolution four-dimensional variational assimilations (4DVAR), run at T_L1279L137 resolution and with a 12-h assimilation window, and interpolated from the T_L1279L137 resolution to the T_L639L91 ensemble resolution.

The perturbations are defined by a linear combination of SVs and perturbations defined by the ECMWF ensemble of data assimilations (EDA, Isaksen et al. 2010):

$$de_j(d,0) = \sum_{a=1}^8 \sum_{k_a=1}^{50} \alpha_{j,k_a} SV_{k_a} + [f_{m(j)}(d-6,6) - < f_{m=1,25}(d-6,6) >] \quad (17)$$

The SVs are computed over up to eight areas (NH, all grid points with points with latitude $\lambda > 30^\circ\text{N}$; SH, all grid points with latitude $\lambda < 30^\circ\text{S}$; and up to six tropical regions where tropical depressions have been reported). The SVs, the fastest growing perturbations over a 48-h time interval (Buizza and Palmer 1995), are computed at T42L91 resolution. The SVs computed over the different areas are linearly combined and scaled to have an amplitude comparable to analysis error estimates provided by the ECMWF high-resolution 4DVAR.

SV-based initial perturbations are combined with perturbations defined by the ECMWF EDA (Buizza et al. 2008). The EDA includes 25 independent 4DVAR run at T_L399L137 resolution and with a 12-h assimilation window. Each EDA member is generated by an independent 4DVAR with perturbed observations, with observations' perturbations sampled from a Gaussian distribution with zero mean and the observation error standard deviation. Furthermore, each EDA member nonlinear trajectory is generated using also the SPPT stochastic scheme (see below). SVs and EDA-based perturbations are combined as described in Buizza et al. (2008). The EDA-based perturbations are defined by differences between 6-h forecasts from the most recent available EDA analyses, which are 6 h before the ENS initial time. Differences are computed between each of the 25 perturbed EDA analyses and their ensemble-mean (since only 25 EDA analyses are available, ENS members 26–50 use the same perturbations as members 1–25).

Considering the ocean, the initial conditions are defined by the 5-member ensemble of ocean analysis, produced by NEMOVAR, the NEMO three-dimensional variational assimilation system (Mogensen et al. 2012). Each ocean analysis is generated using all available in situ temperature and salinity data, an estimate of the surface forcing from ECMWF short-range atmospheric forecasts, sea surface temperature analyses, and satellite altimetry measurements. Each of the 5-member ensembles is created using perturbed versions of the unperturbed wind forcing provided by the high-resolution 4DVAR.

Model uncertainties are simulated only in the atmosphere, using two stochastic schemes (Palmer et al. 2009; Buizza et al. 1999). The stochastically perturbed parameterized tendency (SPPT) scheme is designed to simulate random model errors due to parameterized physical processes; the current version uses three spatial and time level perturbations. The stochastic backscatter (SKEB, Shutts 2005) scheme is designed to simulate the upscale energy transfer induced by the unresolved scales on the resolved scales.

Since March 2008, when the ECMWF medium-range and monthly ensembles were joined, a key component to the ECMWF OG-ENS has been the ENS re-forecast suite (Vitart et al. 2008; Leutbecher and Palmer 2008). The suite includes a 5-member ensemble run once a week with the operational configuration (resolution, model cycle, ...) for the past 20 years. These re-forecasts are used to estimate the model climate required to generate some ensemble products (e.g., the extreme forecast index or weekly average anomaly maps) and in general to calibrate the ENS forecasts. More precisely, for each date (e.g., 14 December 2012), 500 forecasts are defined by combining 5-member forecasts run for the 5 closest initial dates centered on the current date (in this case, 1, 7, 14, 21, and

28 December) of the past 20 years. These re-forecast ensembles start from the ECMWF reanalysis (ERA-Interim) instead of the operational one and use singular vectors of the day, but EDA-based perturbation computed for the current year since the EDA has been running only since 2010 (see Buizza et al. 2008 for more details).

2.5 JMA Japan

The JMA OG-ENS has been producing ensemble forecasts since March 2001. Its data have been available in the TIGGE archive since August 2011.

At the time of writing (May 2014), the JMA medium-range ensemble includes 25 forecasts with a resolution $T_L474L60$ (spectral triangular truncation with linear grid, 0.375° spacing which corresponds to about 50 km in physical space), with the top of the atmosphere at 0.1 hPa. The most recent change was introduced in March 2014, when the resolution was increased from T_L319 to T_L479 , and the daily configuration has been changed from producing 51 forecasts once a day, at 12 UTC, to producing 26 twice a day, at 00 and 12 UTC. Initial uncertainties are simulated using SVs, computed at $T63L40$ resolution, over the NH and the SH extratropics (north and south of 30°) with a 48-h optimization time interval and over the tropics (30°S – 30°N) with a 24-h optimization time interval. Model uncertainties are simulated using a stochastic scheme similar to the original ECMWF SPPT (Buizza et al. 1999). The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used).

The unperturbed initial conditions are provided by the JMA high-resolution 4DVAR, which has a resolution $T_L959L100$ (about 20 km in physical space), interpolated at the ensemble resolution.

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the JMA OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_0(t) + dP_j(t)] dt \quad (18)$$

where A_0 and P_0 represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called with the same parameters), dP_j represents the model uncertainty simulated using the JMA stochastic scheme, and the forecast length T is 11 days. The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + de_j(d,0) \quad (19)$$

The unperturbed analysis is given by the JMA high-resolution four-dimensional variational assimilations. The perturbations are defined by a linear combination of SVs computed over three regions, NH, SH, and tropics:

$$de_j(d,0) = \sum_{a=1}^3 \sum_{k_a=1}^{25} \alpha_{j,k_a} SV_{k_a} \quad (20)$$

It is worth mentioning that JMA also runs a monthly ensemble forecast with 50 members, with a T_L319L60 resolution (about 70 km in grid point space). The system uses bred vectors (Toth and Kalnay 1993, 1997), instead of SVs, to define the initial perturbations. Monthly products are issued once a week, on Fridays, and are based on the 25 forecasts started the previous Wednesday and Thursday at 12 UTC. The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used) and does not simulate model uncertainties.

2.6 KMA Korea

The KMA OG-ENS has been producing ensemble forecasts since March 2000. Its data have been available in the TIGGE archive since October 2007.

The original KMA OG-ENS used initial perturbations defined by a breeding method (Goo et al. 2003). It included 16 perturbed members, run with a T106L21 resolution, with the 16 initial perturbations defined by rotated bred vectors. The forecasts were run once a day (at 12 UTC) up to 10 days. It did not simulate model uncertainties and included only a description of land and atmospheric processes (no wave or ocean model component is used).

Since 2011, KMA has operationally implemented the Unified Model (UM) and related pre-/post-processing system imported from the Met Office (Kai and Kim 2014). Thus since then, the KMA OG-ENS has been practically the same as the UKMO OG-ENS (MOGREPS, see below).

At the time of writing (May 2014), the KMA OG-ENS is based on 24 members, 1 control and 23 perturbed ones, with initial perturbations generated using the ETKF with localization (Bowler et al. 2008; Kai and Kim 2014). It has a horizontal resolution of approximately 40 km and 70 vertical levels (N320L70). The unperturbed analysis is given by the KMA version of the UKMO 4DVAR system. Model uncertainties are simulated using stochastic physics schemes that consist of “random parameters” and “stochastic convective vorticity” schemes (Bowler et al. 2008).

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the KMA OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_0(t) + dP_j(t)] dt \quad (21)$$

where A_0 and P_0 represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called

with the same parameters), dP_j represents the model uncertainty simulated using the MA stochastic scheme, and the forecast length T is 10 days. The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + ETKF_j(d,0) \quad (22)$$

The unperturbed analysis is given by the KMA high-resolution four-dimensional variational assimilations. The perturbations are defined by ETKF perturbations.

2.7 MSC Canada

The MSC OG-ENS has been producing ensemble forecasts since February 1998. Its data have been available in the TIGGE archive since October 2007.

The MSC perturbed-observation approach attempts to represent all sources of uncertainty by adding random perturbations to as many system's components as possible. The Canadian ensemble Kalman filter (EnKF) has been used to supply initial conditions (Houtekamer et al. 2009, 2014), with sources of uncertainty simulated by means of different sets of random perturbation schemes (Houtekamer et al. 1996; Houtekamer and Lefavre 1997).

Throughout the years, the number of the MSC ensemble members has been increased from 8 to 16 and now stands at 20. The number of the members of the MSC ensemble Kalman filter used to define the initial conditions has also been increased from 48 to 96 and now stands at 192.

The most recent change was implemented in December 2013, when the treatment of the surface temperature of the sea (SST) was changed. In the new ensemble, the SST value evolves with the forecast time using the persistence of the anomaly (deviation from the climatology) method. Since this change, monthly forecasts (32 days) are produced once a week. Furthermore, re-forecasts based on 4-member ensemble run once a week for the past 18 years are being produced, so that in the near future the model climate can be estimated, and products such as the ECMWF extreme forecast index can be generated.

At the time of writing (May 2014), the MSC OG-ENS includes 21 members, 1 unperturbed and 20 perturbed, and is run twice a day (at 00 and 12 UTC) up to forecast day 16. Once a week, at 00 UTC of each Thursday, the ensemble is extended to 32 days. The initial conditions are obtained directly from the Canadian EnKF. Model uncertainties are sampled using four schemes: isotropic perturbations at the initial time, different physical parameterizations, stochastic physical tendency perturbations, and stochastic kinetic energy backscatter (Houtekamer et al. 2009). The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used). Since August 2011, the resolution of MSC OG-ENS has been $N600 \times N300$ (600 grid points in longitude and 300 in latitude), which corresponds to about 0.6° , 75 km.

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the MSC OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_j(t) + dP_j(t)] dt \quad (23)$$

where A_0 represents the “unperturbed” model dynamical core, P_j represents the physical tendencies, which varies in each member since different parameterization schemes and/or different parameters are used, and dP_j represents the model uncertainty simulated using different stochastic schemes.

The initial conditions are defined by one of the EnKF members:

$$e_j(d,0) = ENKF_j(d,0) \quad (24)$$

It is worth mentioning that the MSC and the NCEP OG-ENS forecasts are exchanged in real time to generate the multi-model products of the North American Ensemble Forecast System (NAEFS, http://weather.gc.ca/ensemble/naefs/index_e.html). NAEFS is a joint project involving the Meteorological Service of Canada (MSC), the United States National Weather Service (NWS), and the National Meteorological Service of Mexico (NMSM). NAEFS, launched in November 2004, provides users with operational products generated by blending the MSC and the NCEP OG-ENSs. The research/development and operational costs of the NAEFS system are shared by the three organizations (MSC, NWS, and NMSM), which make it more cost effective and result in higher quality and more extensive weather forecast products.

2.8 NCEP United States of America

The NCEP OG-ENS has been producing ensemble forecasts since December 1992. Its data have been available in the TIGGE archive since March 2007.

The original version of the NCEP OG-ENS simulated only initial uncertainties using bred vectors (Toth and Kalnay 1993, 1997). The breeding method involves the maintenance and cycling of perturbation fields that develop between two numerical model integrations and that once rescaled define the initial perturbations. In its original form with a single global rescaling factor, the bred vectors (BVs) represent a nonlinear extension of the Lyapunov vectors (Boffetta et al. 1998). In the NCEP OG-ENS, multiple breeding cycles were used, each initialized at the time of implementation with independent arbitrary perturbation fields (“seeds”). The original system was based on 10 perturbed ensemble members, run both at 00 and 12 UTC every day up to 16 days lead time. For both cycles, the generation of the initial perturbations was done in five independently run breeding cycles, originally started with different arbitrary perturbations, using the regional rescaling algorithm. Since then, the system has been upgraded several times.

At the time of writing (May 2014), the NCEP OG-ENS consists of four runs a day (at 00, 06, 12, and 18 UTC), with a T382L64 resolution up to forecast day 7, after

which it is run to 16 days at a T126L64 resolution. The unperturbed initial conditions are given by the truncated T382L64 NCEP analysis. Each run includes 1 unperturbed and 20 perturbed forecasts. The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used).

The initial perturbations are now generated using the ensemble transform (ET) with rescaling technique (ETR; Wei et al. 2006, 2008). The ETR method is an extension of the original breeding approach, in that they both dynamically cycle the perturbations. In an ensemble with only two members, both methods should produce the same results. To improve the simulation of initial uncertainties in cases of tropical storms, the initial condition perturbations also apply a tropical storm relocation method.

Model uncertainties are today represented using the stochastic total perturbation scheme (Hou et al. 2008), designed to represent model uncertainty by adding a stochastic forcing term to the total tendency. For each 6-h forecast period, this term is defined by a linear combination of the past ensemble tendencies. In the linear combination, the total tendencies are rescaled so that, on average, the ensemble standard deviation matches the error of the ensemble-mean.

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the NCEP OG-ENS:

$$e_j(d;T) = e_j(d;0) + \sum_{k=1}^N \Delta T_{j,k} + \Delta S_{j,k} \quad (25)$$

$$\Delta T_{j,k} = \int_{T_k}^{T_k+6} [A_{0,j}(t) + P_{0,j}(t)] dt \quad (26)$$

$$\Delta S_{j,k} = \sum_{m=1}^N w_{m,k} \Delta T_{m,k} \quad (27)$$

where $A_{0,j}$ and $P_{0,j}$ represent the “unperturbed” model dynamical and physical tendencies (i.e., there is only one dynamical core and one set of parameterizations called with the same parameters). The j -th subscript indicates that each finite-time tendency is different for each ensemble member, since it is computed starting from a different initial state.

Every 6 h, for each member the 6-h tendency $\Delta T_{j,k}$ is computed by integrating ahead in time the model equations for 6 h, starting from the initial state. Once the 6-h tendency $\Delta T_{j,k}$ has been computed, the stochastic perturbation $\Delta S_{j,k}$ is defined by a linear combination of the 6-h tendencies. It should be evident now why the scheme is called STTP, where “T” stands for “total”: since the original tendencies include also the dynamical tendencies, compared to the ECMWF method, this approach also perturbs the tendencies due to the dynamics. Once the STTP term has been computed, the initial states are then advanced by 6 h by adding $(\Delta T_{j,k} + \Delta S_{j,k})$. Then the process is repeated.

The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + ETR_j(d,0) \quad (28)$$

where the unperturbed analysis is given by the NCEP high-resolution four-dimensional variational assimilations.

2.9 UKMO United Kingdom

The UK Meteorological Office (UKMO) OG-ENS has been producing ensemble forecasts since August 2005. Its data have been available in the TIGGE archive since October 2006.

UKMO uses an ensemble transform Kalman filter approach (ETKF, Wei et al. 2006; Bishop et al. 2001; Bowler et al. 2007, 2008), to generate the initial perturbations. The ETKF is a simplified version of the ensemble Kalman filter, a data assimilation scheme which updates the mean state of the atmosphere and the error covariance in that estimate using background information obtained from an ensemble. The ETKF can be viewed as a transformation of the NCEP error breeding scheme.

In the UKMO OG-ENS, the initial perturbations are defined by a linear combination of the forecast perturbations from the previous cycle of the ensemble. The weights are calculated by considering the spread of the ensemble in the space of the observations (Wang et al. 2004), ensuring that the perturbations are centered on the control analysis and are orthogonal. As is normal in the EnKF, the perturbations are inflated to ensure that the ensemble has the correct spread for the next analysis time ($T + 12$ h in MOGREPS). The inflation factor is calculated online so the system will automatically retune itself to any model changes.

Two stochastic physics schemes are included to represent the effects of structural and sub-grid-scale model uncertainties: the random parameters (RP) scheme and the stochastic convective vorticity (SCV) scheme. The RP scheme treats a selected group of parameters as stochastic variables (Lin and Neelin 2000; Bright and Mullen 2002). A total of eight parameters from four different physical parameterizations are included: large-scale precipitation, convection, boundary layer, and gravity wave drag. The main aim of the SCV scheme (Gray and Shutts 2002) is to represent potential vorticity anomalies dipoles similar to the one typically associated with a mesoscale convective systems.

At the time of writing (May 2014), the UKMO OG-ENS includes 1 unperturbed and 24 perturbed members and runs twice a day with a 60 km resolution and 70 vertical levels and up to forecast day 15. The forecast model includes only a description of land and atmospheric processes (no wave or ocean model component is used).

If we now reconsider the general equations (2, 3, 4, and 5) that describe each ensemble member, this is how they should be qualified for the KMA OG-ENS:

$$e_j(d;T) = e_j(d;0) + \int_0^T [A_0(t) + P_j(t) + dP_j(t)] dt \quad (29)$$

where A_0 represents the “unperturbed” model dynamical tendencies (i.e., there is only one dynamical core), P_j represents the fact that the physical parameterization is integrated by perturbing some key parameters as defined by the RP scheme, dP_j represents the model uncertainty simulated using the SCV scheme, and the forecast length T is 15 days. The initial conditions are defined by adding perturbations to the unperturbed analysis:

$$e_j(d,0) = e_0(d,0) + ETKF_j(d,0) \quad (30)$$

The unperturbed analysis is given by the UKMO high-resolution four-dimensional variational assimilations, and the perturbations are defined by ETKF perturbations.

3 Average Performance of the Operational, Global, Medium-Range Ensembles

In Sect. 2 we have described the key characteristics of the nine global ensembles that can be accessed within TIGGE. The question we want to address now is whether the ensembles’ performances reflect the different designs. In other words, what is the sensitivity of the ensemble performance to the ensemble configuration?

We will compare eight of the TIGGE ensembles (all but the BMRC one, which has stopped production in 2010) first for a single case and then for whole seasons. First, we will look at similarities and differences in the ensemble-mean and the spread (measured by the ensemble standard deviation, i.e., the spread measured around the ensemble-mean) for the case of 10 January 2013, and then we will look at average performance indices for few seasons.

Similar comparisons were performed by other authors in the past. In the first of these comparisons, Park et al. (2008) concluded that there was a large difference between the performances of the single ensembles. For the 500 hPa geopotential height over the Northern Hemisphere (NH), in the medium-range (say around forecast day 5), the difference between the worst and the best control or ensemble-mean forecasts was about 2 days of predictability, while the difference between the worst and the best probabilistic predictions was even larger, about 3 days of predictability.

Hagedorn et al. (2012) investigated the possibility to combine all the available ensembles into a multi-model one and compared the performance of this multi-model ensemble with an ensemble defined by a calibrated ECMWF OG-ENS. This ECMWF calibrated ensemble was generated using re-forecast based on the past 18 years. They concluded that, considering the statistical performance of global probabilistic forecasts of 850 hPa and 2 m temperatures, a multi-model ensemble

containing nine ensembles from the TIGGE archive did not improve on the performance of the best single model, the ECMWF OG-ENS. However, a reduced multi-model system, consisting of only the four best ensemble systems, provided by Canada, the United States, the United Kingdom, and ECMWF, showed an improved performance. They also concluded that the ECMWF OG-ENS was the main contributor for the improved performance of the multi-model ensemble; that is, if the multi-model system did not include the ECMWF contribution, it was not able to improve on the performance of the ECMWF EPS alone. These results were shown to be only marginally sensitive to the choice of verification dataset.

Yamaguchi and Majumdar (2010) compared TIGGE ensemble tropical cyclone (TC) track predictions with tracks generated by a multi-center grand ensemble (MCGE). The 9 TIGGE ensembles were considered, and 58 TCs in the western North Pacific from 2008 to 2010 were verified. In the verification of TC strike probability, the Brier skill score of the MCGE was larger than that of the best SME, which was the ECMWF OG-ENS on a medium-range time scale, although this was not true on a short- to medium-range scale. In addition, the reliability was improved by the MCGE, especially in the high-probability range.

3.1 Ensemble-Mean and Spread for One Specific Case (10 January 2013)

First, let's consider the OG-ENS forecast for one specific date, the 10th of January 2013 (12 UTC), to illustrate how similar or different the ensemble-mean and the spread of the OG-ENS are. For this date, all but the BMRC OG-ENS are available in the TIGGE archive. This case was randomly selected within the most recent season, winter 2012–2013, for which at least eight ensembles were available, and for which routine scores were available in the ECMWF database.

Figures 2, 3, 4, 5, 6, 7, 8, and 9 show the ensemble-mean and the spread, measured by the ensemble standard deviation. The fields are shown over a Euro-Atlantic sector at initial time (12 UTC of 10 January 2014) and at $t + 48$ h. This geographical region has been chosen simply because it has a good size to appreciate synoptic-scale differences, while looking at the whole NH it would have been very difficult to appreciate similarities and differences between the eight ensembles.

If we start considering the ensemble-mean states, at initial time they are very similar, while at $t + 48$ h some small differences start to appear. By contrast, differences are much larger in terms of ensemble spread already at initial time. The spread of the CPTEC OG-ENS (Fig. 3) is zero, since by construction perturbations are defined only up to 30°N . The JMA OG-ENS (Fig. 5) has spread confined in some specific regions, since it is using only singular vectors, which are very localized in space (their spread is very similar to the one of the original ECMWF OG-ENS, which used only singular vectors).

The comparison between the JMA OG-ENS and the ECMWF OG-ENS (Figs. 4 and 5) illustrates the impact that had on the ECMWF spread replacing the evolved singular vectors with perturbations defined by an ensemble of data assimilations

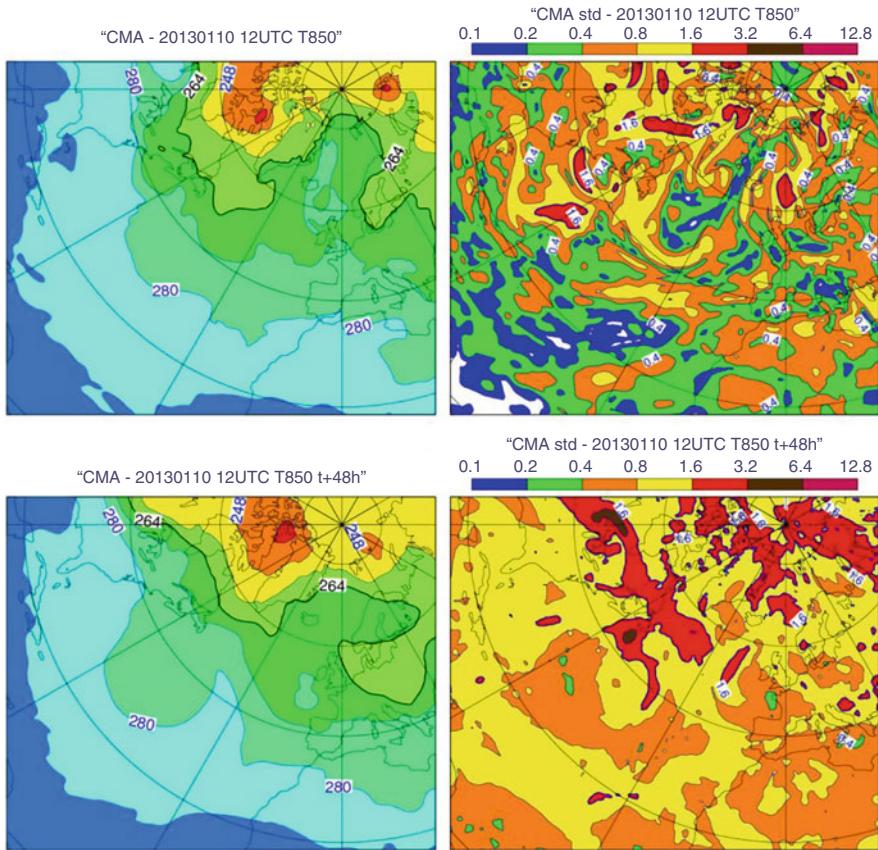


Fig. 2 CMA OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at $0.1^{\circ}, 0.2^{\circ}, 0.4^{\circ}, 0.8^{\circ}, 1.6^{\circ}, 3.2^{\circ}, 6.4^{\circ}$, and 12.8°

(see Buizza et al. 2008 for more details). The replacement led to the ECMWF initial perturbations being less localized both geographically and in the vertical and to include finer scales.

The KMA, MSC, and UKMO OG-ENSSs have the largest initial spread: this is linked to the fact that their initial perturbations, generated using the EnKF or ETKF methods, grow very slowly, slower than forecast error, and thus to get the right level of spread in the medium-range, they have to add rather large initial perturbations.

At forecast time $t + 48$ h, the spread values of all ensembles are rather similar, with local maxima around $3\text{--}6^{\circ}$. Also in terms of structures, maxima are localized in the same areas, where the jet stream is stronger or where cyclonic developments occur. This is not surprising since all these ensembles have been designed for the medium-range, say the 3–10 forecast range, and to achieve this, they have

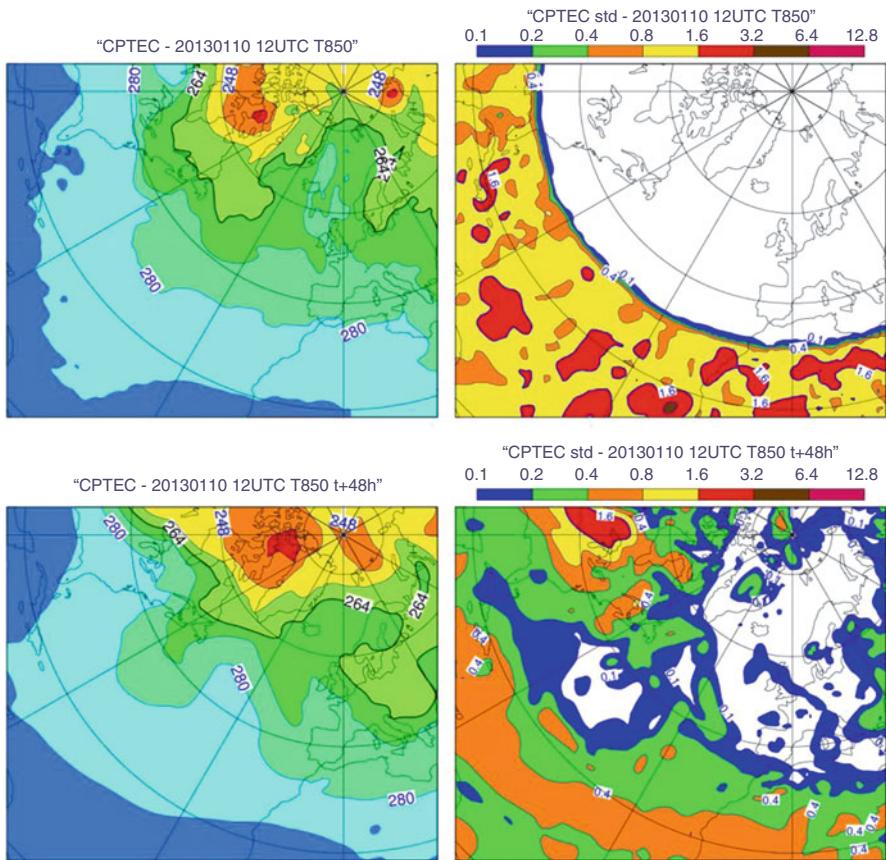


Fig. 3 CPTEC OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at 0.1° , 0.2° , 0.4° , 0.8° , 1.6° , 3.2° , 6.4° , and 12.8°

configured to have, on average, the right level of spread from about forecast day 2–3.

A proper assessment of the performance of an ensemble system has to be based on a large number of cases, to verify that, on average, the statistical properties of the system are as expected. To do this, we have computed the average ensemble spread for a whole season, winter 2012–2013 (i.e., December 2012 and January–February 2013). Figure 10 shows the winter 2012–2013 spread, measured for temperature at 850 hPa, over the Northern Hemisphere extra-tropics (NH, 20°N – 90°N) and the tropics (TR, 30°S – 30°N).

The NH results confirm the indications given by the case study, with the JMA OG-ENS showing the smallest initial spread and the CMA, KMA, MSC,

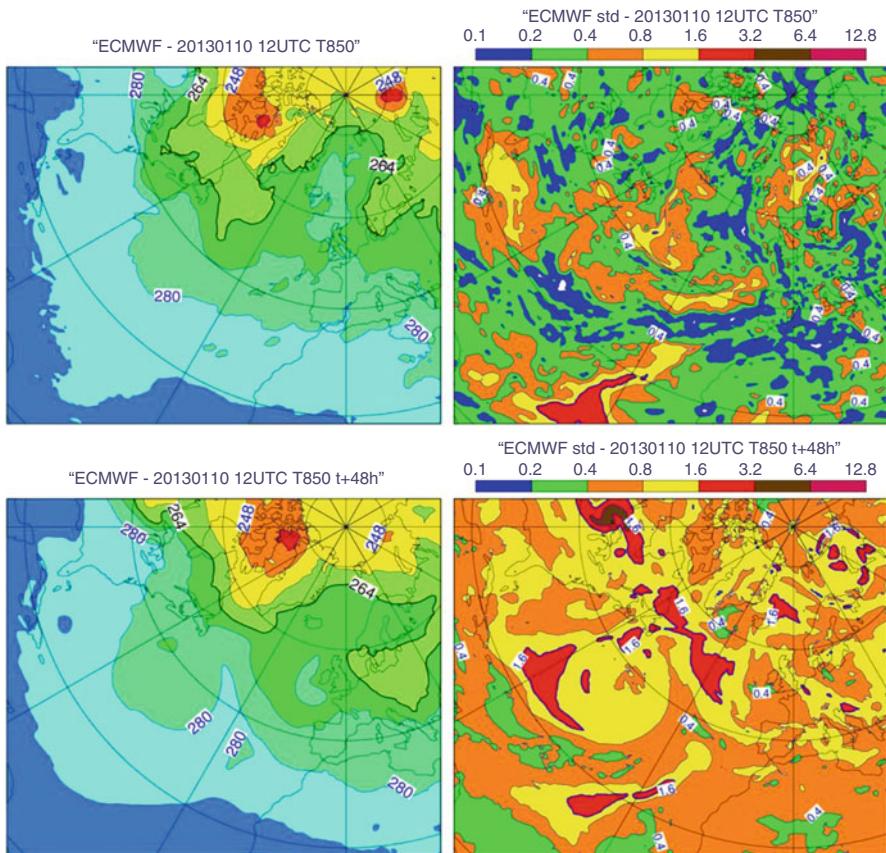


Fig. 4 ECMWF OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at 0.1° , 0.2° , 0.4° , 0.8° , 1.6° , 3.2° , 6.4° , and 12.8°

and UKMO OG-ENS the largest. If we compare the spread values at $t + 6$ h (first point) with the values at 48–72 h, it is very evident how much faster the JMA OG-ENS singular vector-based perturbations grow compared to the EnKF/ETKF ones of the CMA, KMA, MSC, and UKMO OG-ENSSs.

The TR results show a different picture, with the MSC OG-ENS showing the largest values. This is probably due to the fact that the MSC OG-ENS starts directly from EnKF initial conditions generated using four different methods to simulate model uncertainties: isotropic perturbations at initial time, different physical parameterizations, and two stochastic schemes.

These considerations confirm earlier results discussed, e.g., by Buizza et al. (2005), Park et al. (2008), and Hagedorn et al. (2012).

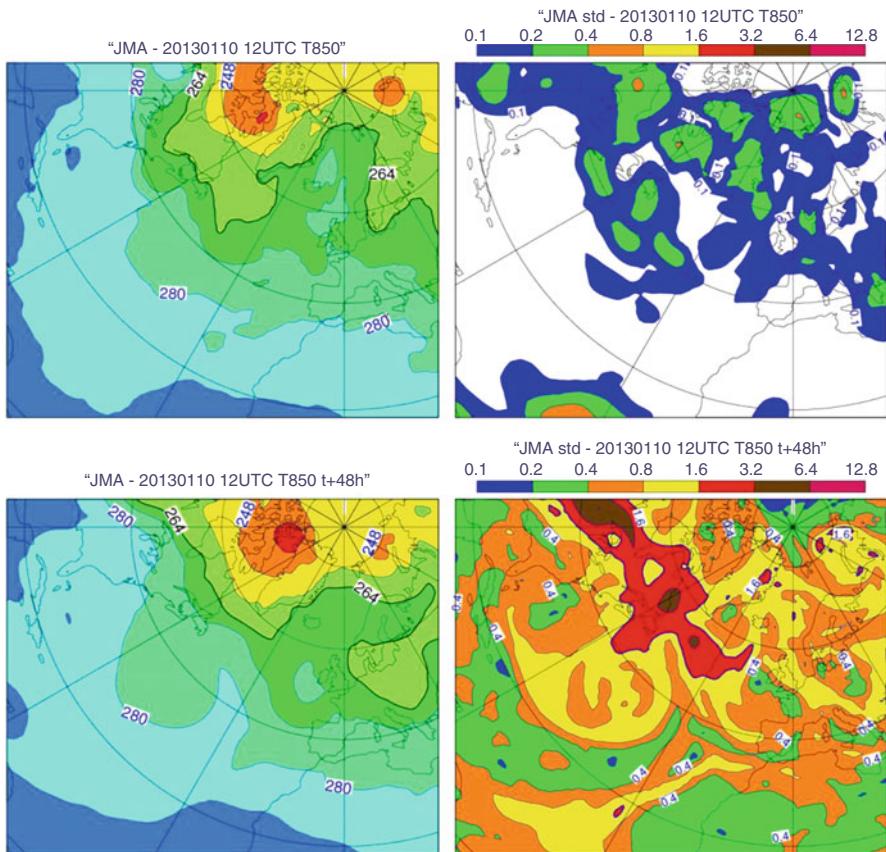


Fig. 5 JMA OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at 0.1° , 0.2° , 0.4° , 0.8° , 1.6° , 3.2° , 6.4° , and 12.8°

3.2 Average Performance over the Tropics and the Northern Hemisphere Extra-Tropics

In the previous section, we have seen that the ensemble's design influences the ensemble spread. Hereafter, we will look at the average ensemble skill. As in the previous section, first we will look at the winter 2012–2013 850 hPa temperatures over the NH and the TR regions. Then, we will also look at more recent seasons, for which routine performance measures are computed for few ensembles at ECMWF. In all these verifications, each ensemble has been verified against its own analysis.

Performance is going to be assessed looking at three key performance metrics: reliability, as measured by the match between the ensemble spread and the error of the ensemble-mean, the root-mean-square error (RMSE) of the ensemble-mean, and

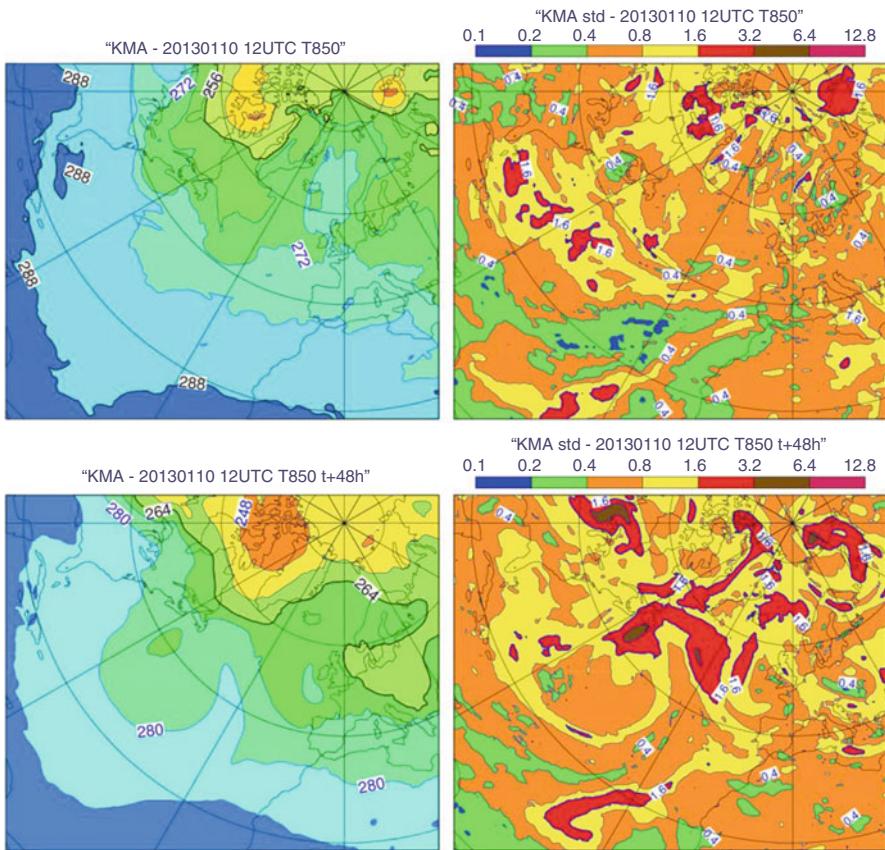


Fig. 6 KMA OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at $0.1^{\circ}, 0.2^{\circ}, 0.4^{\circ}, 0.8^{\circ}, 1.6^{\circ}, 3.2^{\circ}, 6.4^{\circ}$, and 12.8°

the accuracy of probabilistic forecast measured by the continuous ranked probability score.

Reliability is a very important attribute of ensembles. In a reliable ensemble, an event that is predicted to have a p% probability to happen in reality occurs p% of the times. Thus reliability can be measured by looking at average scatter diagrams of forecast probabilities versus occurrence frequencies. A reliable ensemble is also an ensemble that includes verification within its forecast distribution. For this to happen, the average spread of the ensemble measured by the standard deviation has to match, on average, the error of the ensemble-mean. The similarity between the ensemble spread and the ensemble-mean error is the first metric that we are going to analyze.

The second metric is the root-mean-square error (RMSE) of the ensemble-mean, the first-order moment of each ensemble probability distribution function.

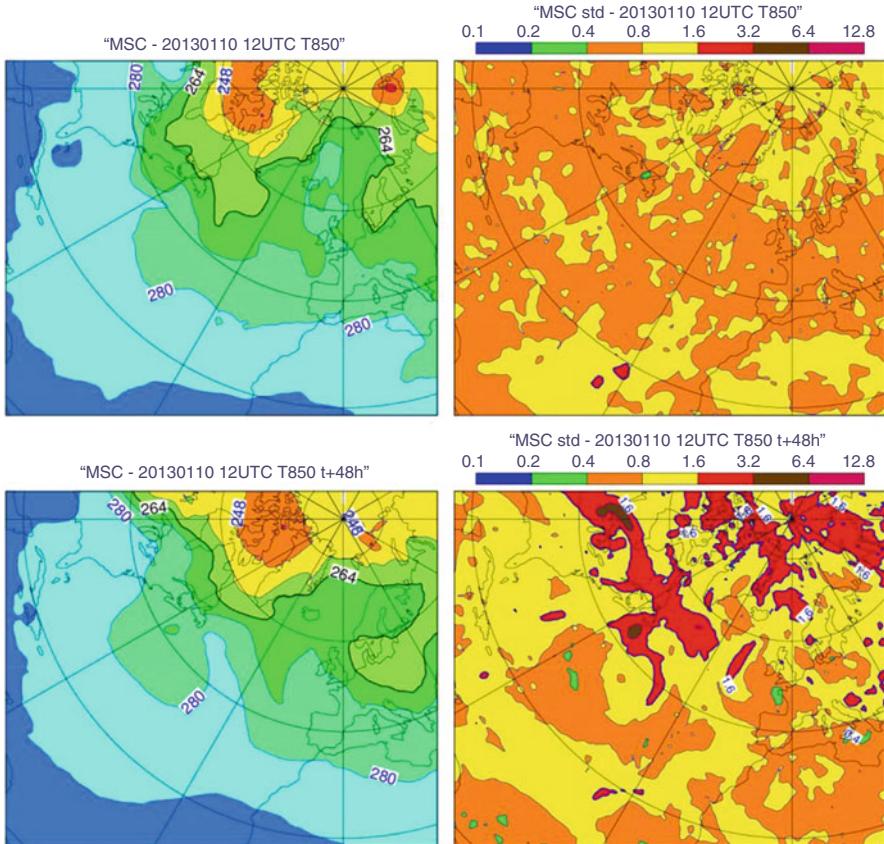


Fig. 7 MSC OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at $0.1^{\circ}, 0.2^{\circ}, 0.4^{\circ}, 0.8^{\circ}, 1.6^{\circ}, 3.2^{\circ}, 6.4^{\circ}$, and 12.8°

The third metric is the continuous ranked probability score (CRPS), the equivalent of the RMSE for probabilistic forecasts. It measures the average distance between the ensemble forecast probability density function and the observed density function, which is a delta function if observation errors are not taken into account (our case) or a very narrow distribution if they are taken into account. The CRPS is zero for a perfect forecast.

Figure 11 shows the RMSE of the eight TIGGE ensemble-mean forecasts (dashed lines) and the average ensemble spreads (already shown in Fig. 10). Over the NH, the plot shows firstly that the ECMWF OG-ENS ensemble-mean forecast (blue dashed line) has the lowest RMSE, while the CMA OG-ENS has the largest RMSE. Secondly, it shows that for the ECMWF OG-ENS the average spread and ensemble-mean error curves (solid and dashed blue curves) are the closest. They are followed by the MSC (black lines), NCEP (green lines), and JMA (orange)

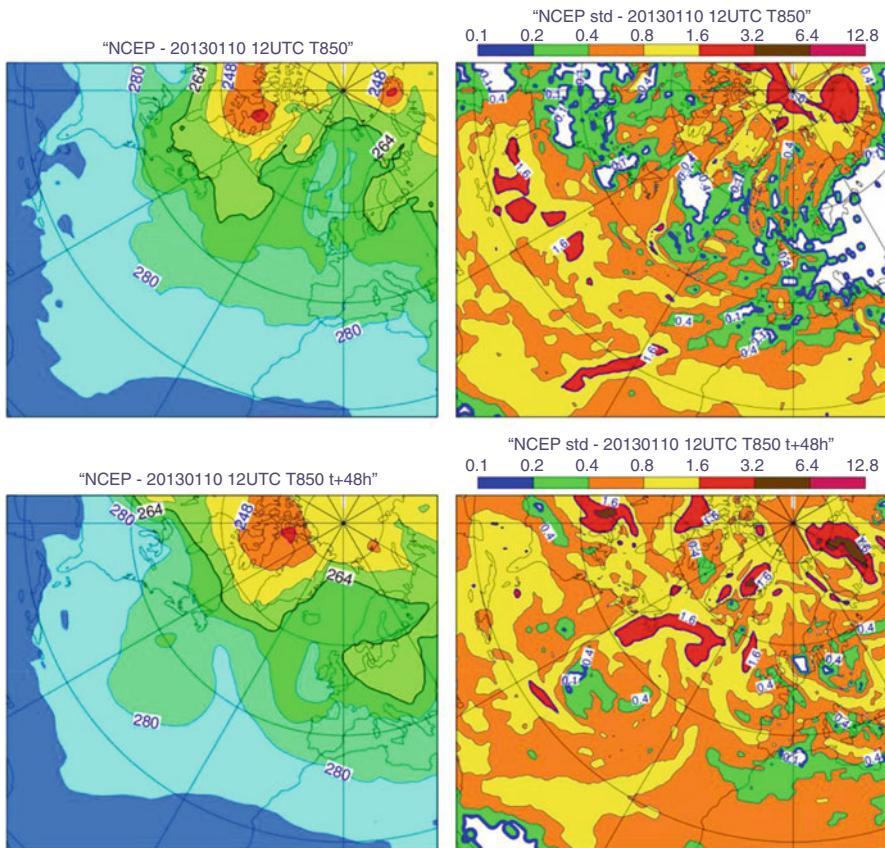


Fig. 8 NCEP OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at 0.1° , 0.2° , 0.4° , 0.8° , 1.6° , 3.2° , 6.4° , and 12.8°

OG-ENSs. Over the TR, the NCEP OG-ENS has the lowest RMSE. In terms of reliability, the MSC OG-ENS shows the closest match between the spread and ensemble-mean error curves, followed by the ECMWF, NCEP, and UKMO OG-ENS.

Figure 12 shows the CRPS of the eight ensembles for winter 2012–2013. Results show that over the NH, the ECMWF OG-ENS performs best, with the JMA and the NCEP OG-ENSs performing second best. Over the TR, results are similar between forecast days 2 and 6, while afterward the NCEP OG-ENS has the lowest CRPS.

Figures 13 and 14 show the most recent comparison of five of the TIGGE OG-ENS which are routinely monitored and compared at ECMWF, the JMA, MSC, NCEP, and UKMO ensembles. Results refer to the most recent summers and winters (JJA, Fig. 13, and DJF, Fig. 14), for the 500 hPa geopotential height and the 850 hPa temperature over the NH. The CRPSS is the CRPS skill score, computed using the observed

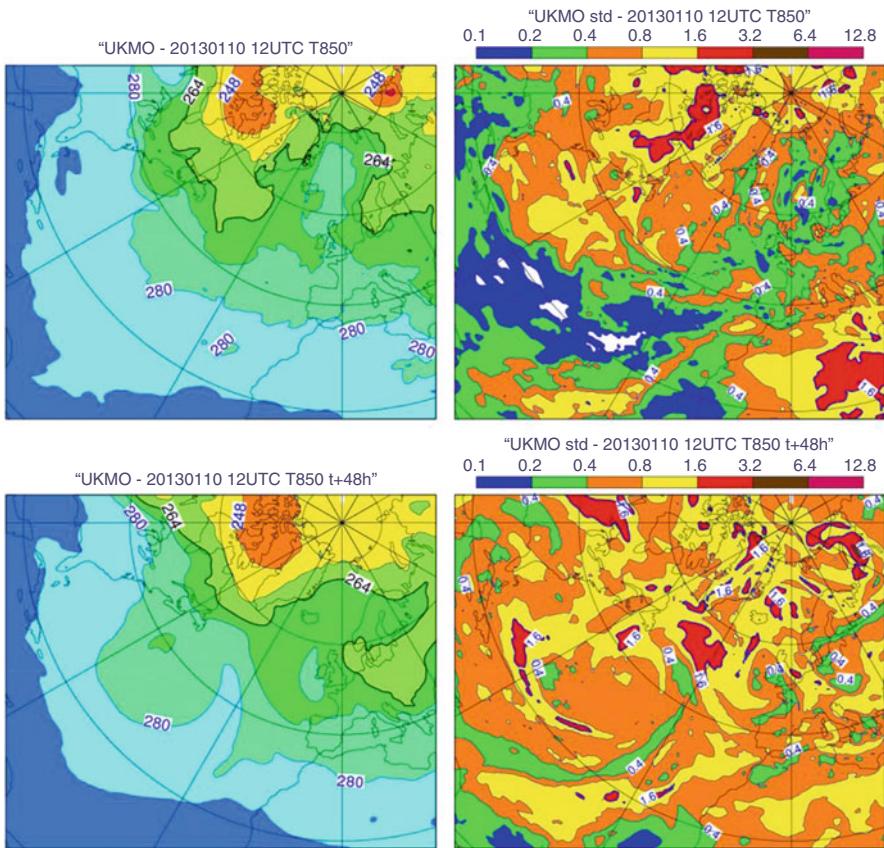


Fig. 9 UKMO OG-ENS ensemble-mean (left panels) and spread (right panels) at initial time (top panels) and after 48 h (bottom panels). The field shown is temperature at 850 hPa. The contour interval for the ensemble-mean is 8° , and the spread shading is at 0.1°, 0.2°, 0.4°, 0.8°, 1.6°, 3.2°, 6.4°, and 12.8°

climatology as reference. Results confirm that the ECMWF OG-ENS performs best, with differences in skill in the medium-range (say at about forecast day 7) of about 1 day (in other words, the ECMWF OG-ENS $t + 192$ h forecast has the same CRPSS as the second best OG-ENS $t + 168$ h forecast).

Finally, Figs. 15 and 16 show the most recent results for four of the TIGGE ensembles, the ECMWF, JMA, NCEP, and UKMO OG-ENS, for the prediction of 24-h accumulated precipitation, verified against observations at synoptic stations, for summer 2013 and winter 2013–2014. These results firstly indicate that the skill for the prediction of precipitation at synoptic stations is lower than the skill for the prediction of the large-scale flow, as represented, e.g., by the 500 hPa geopotential height and the 850 hPa temperature. While these latter forecasts are skillful (i.e., better than a probabilistic forecast based on climatology) for the whole 15-day forecast range, for precipitation they are skillful only for up to a maximum of

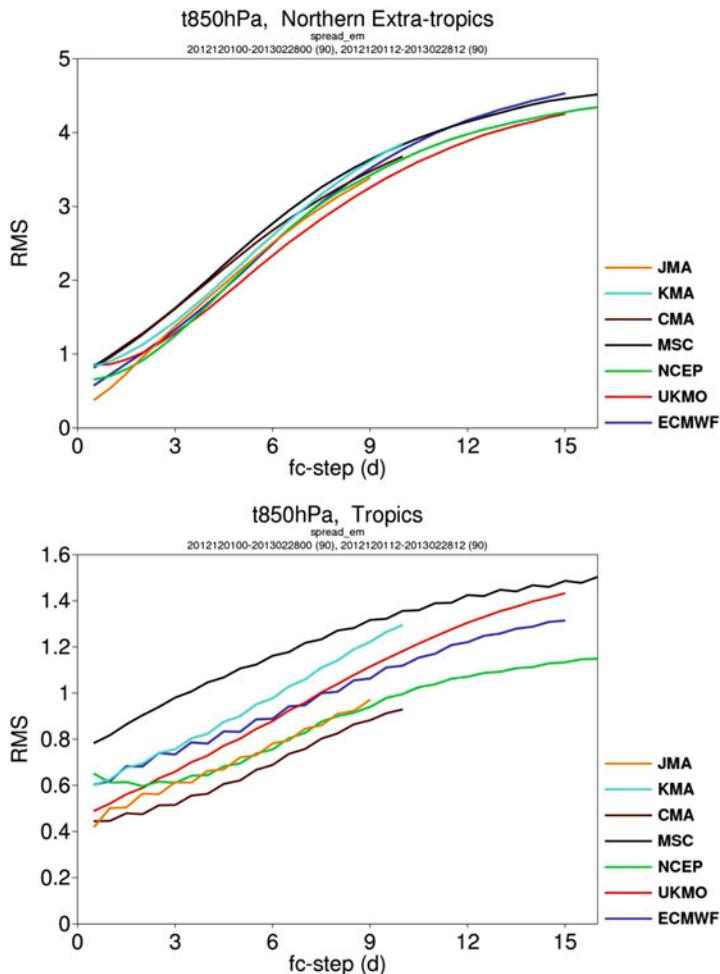


Fig. 10 Winter 2012–2013 average ensemble spread over the Northern Hemisphere (20°N – 90°N , top panel) and the tropics (30°S – 30°N , bottom panel) of seven TIGGE ensembles, for temperature at 850 hPa

10 days over the Northern Hemisphere extra-tropics. Over the tropics, only the ECMWF OG-ENS provided skillful forecasts in winter 2013–2014 but only up to forecast day 5.

The average performance metric confirms results published in the literature that, on average, the ECMWF OG-ENS is the best ensemble, followed by MSC, NCEP, and JMA. It is interesting to look back at the relative cost of the eight TIGGE ensembles and compare how performance and cost rankings are linked. The comparison indicates that best performer is also the most expensive system and that there is a 70% correlation between the ranks of the eight OG-ENS relative costs and performance.

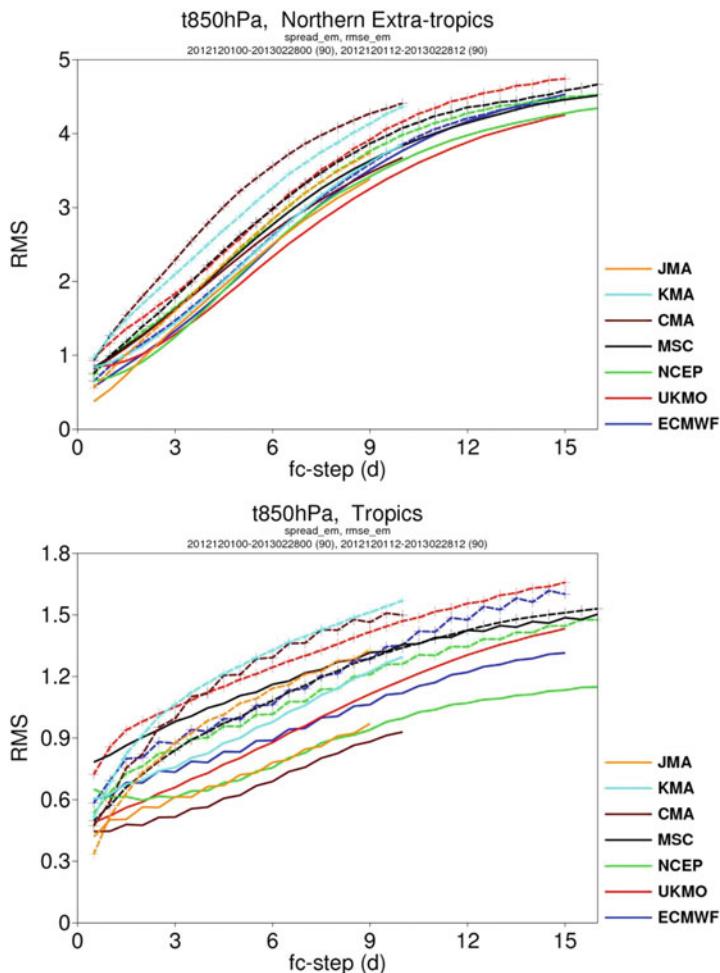


Fig. 11 Winter 2012–2013 average ensemble spread (solid lines) and root-mean-square error of the ensemble-mean (dashed lines) over the Northern Hemisphere (20°N – 90°N , top panel) and the tropics (30°S – 30°N , bottom panel) of eight TIGGE ensembles, for temperature at 850 hPa. The spread lines are the same as in Fig. 10. Each ensemble has been verified against its own analysis

4 The Future of Ensemble Techniques

In the last 25 years, we have witnessed a major shift of the approach followed in numerical weather prediction from a deterministic one, based on one single forecast, to a probabilistic one, whereby multiple ensembles are used to estimate the probability density function of initial and forecast states. 1992 saw the implementation of the first two operational ensemble systems at ECMWF in Europe and NCEP in the

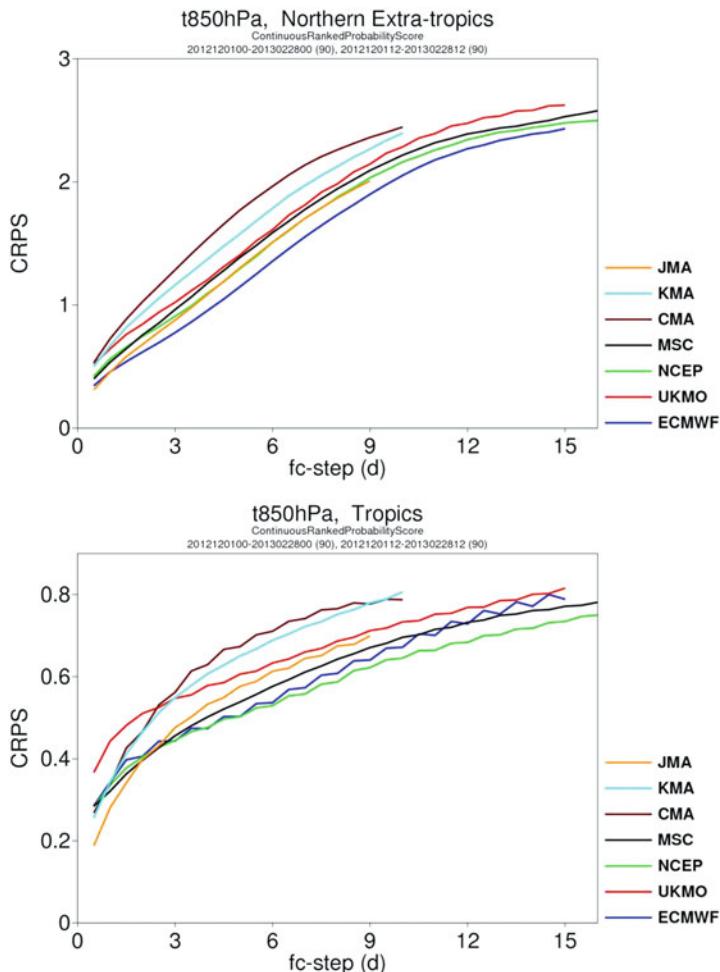


Fig. 12 Winter 2012–2013 average continuous ranked probability score (CRPS) over the Northern Hemisphere (20°N – 90°N , top panel) and the tropics (30°S – 30°N , bottom panel) of eight TIGGE ensembles, for temperature at 850 hPa. Each ensemble has been verified against its own analysis

United States of America. They were followed by MSC in Canada in 1995 and by others few years afterward. Building on the positive results of the global ensembles, limited area, short-range ensembles started being developed.

Today, it is widely accepted that a forecast has to be accompanied by uncertainty estimations, by confidence measures. These can be expressed in different ways, to provide the best possible service to any user, e.g., as a range of possible scenarios or as probabilities that events of interest can occur. Users interested in medium-range forecasts can have access to many hundreds of forecasts issued every day. Ensembles are also widely used to provide an estimate of the initial state uncertainty (analysis error) and in forecasting over all time scales. Decadal forecasts and climate

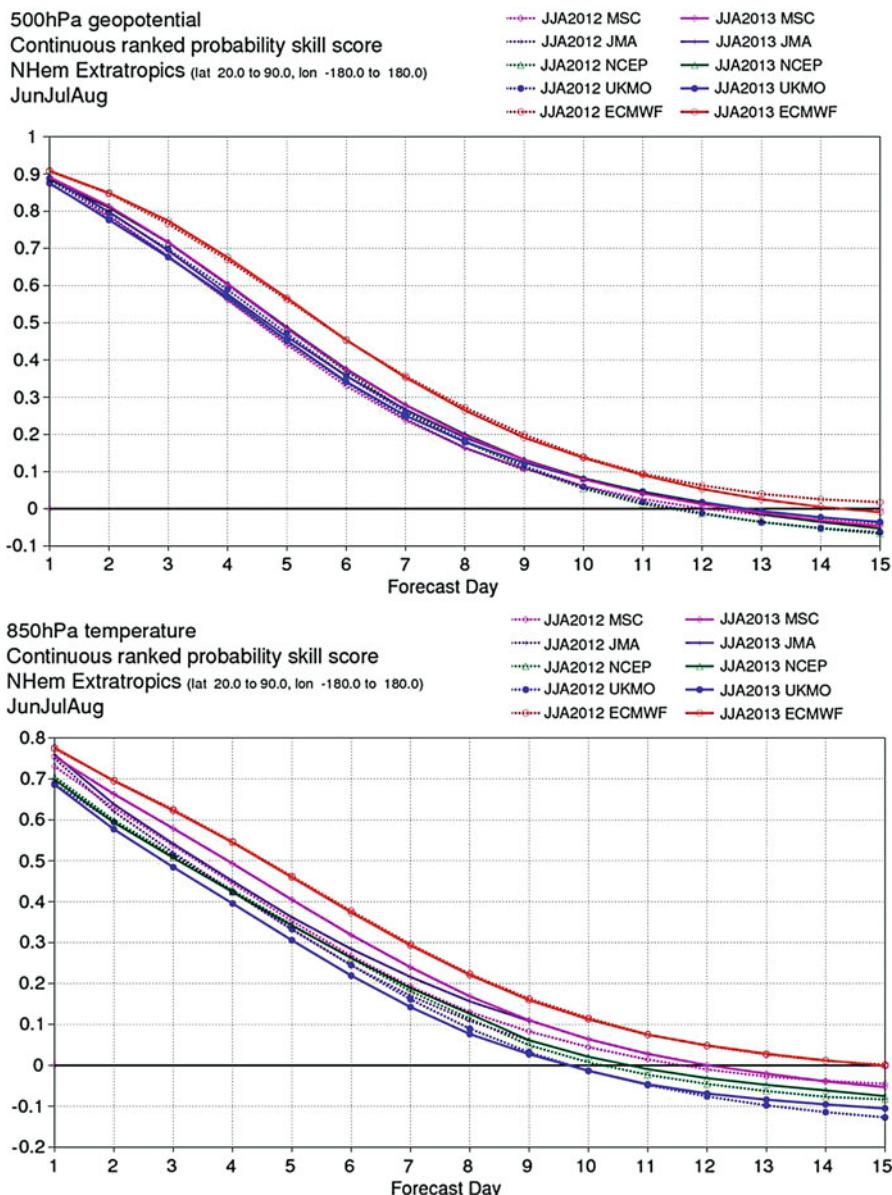


Fig. 13 Summer 2013 (JJA, solid lines) and summer 2012 (JJA, dashed lines) average continuous ranked probability skill score (CRPSS) over the Northern Hemisphere (20°N – 90°N) for the geopotential height at 500 hPa (top panel) and the temperature at 850 hPa (bottom panel) of five TIGGE ensembles from ECMWF, MSC, JMA, NCEP, and UKMO. Each ensemble has been verified against its own analysis

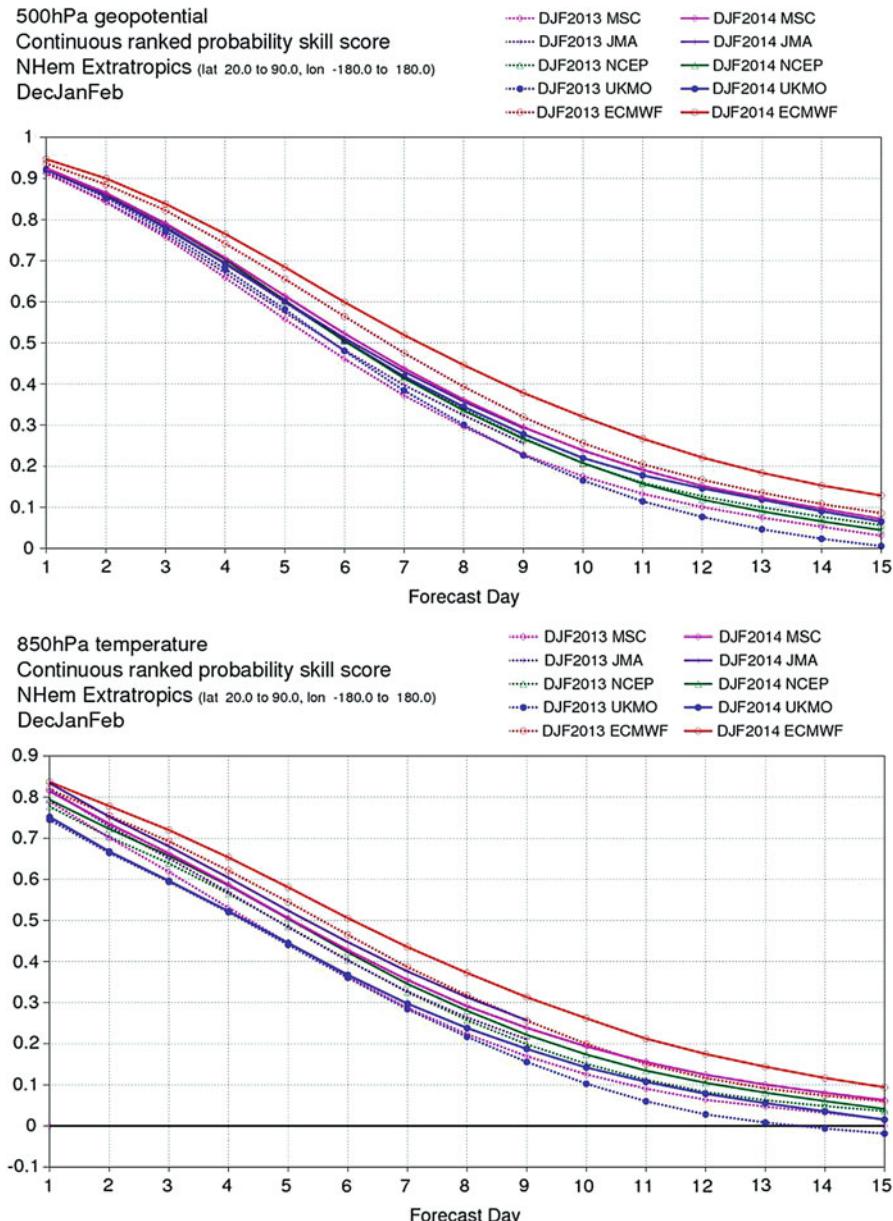


Fig. 14 Winter 2013/2014 (D13JF14, solid lines) and winter 2012/2013 (D12JF13, dashed lines) average continuous ranked probability skill score (CRPSS) over the Northern Hemisphere (20°N – 90°N) for the geopotential height at 500 hPa (top panel) and the temperature at 850 hPa (bottom panel) of five TIGGE ensembles from ECMWF, MSC, JMA, NCEP, and UKMO. Each ensemble has been verified against its own analysis

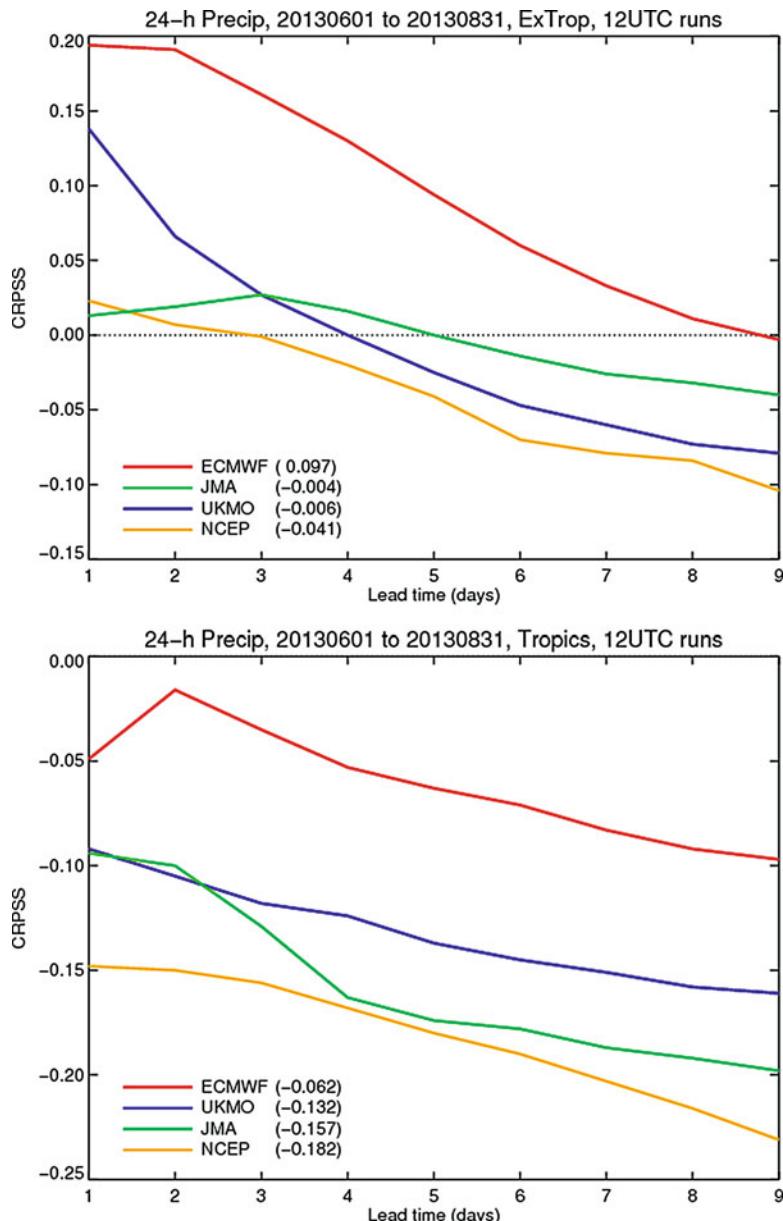


Fig. 15 Summer 2013 (JJA) average continuous ranked probability skill score (CRPSS) for 24-h accumulated precipitation over the Northern Hemisphere extra-tropics (20°N – 90°N , top panel) and the tropics (30°S – 30°N , bottom panel), for the four TIGGE ensembles from ECMWF, JMA, NCEP, and UKMO. Verification is against observations at synoptic stations

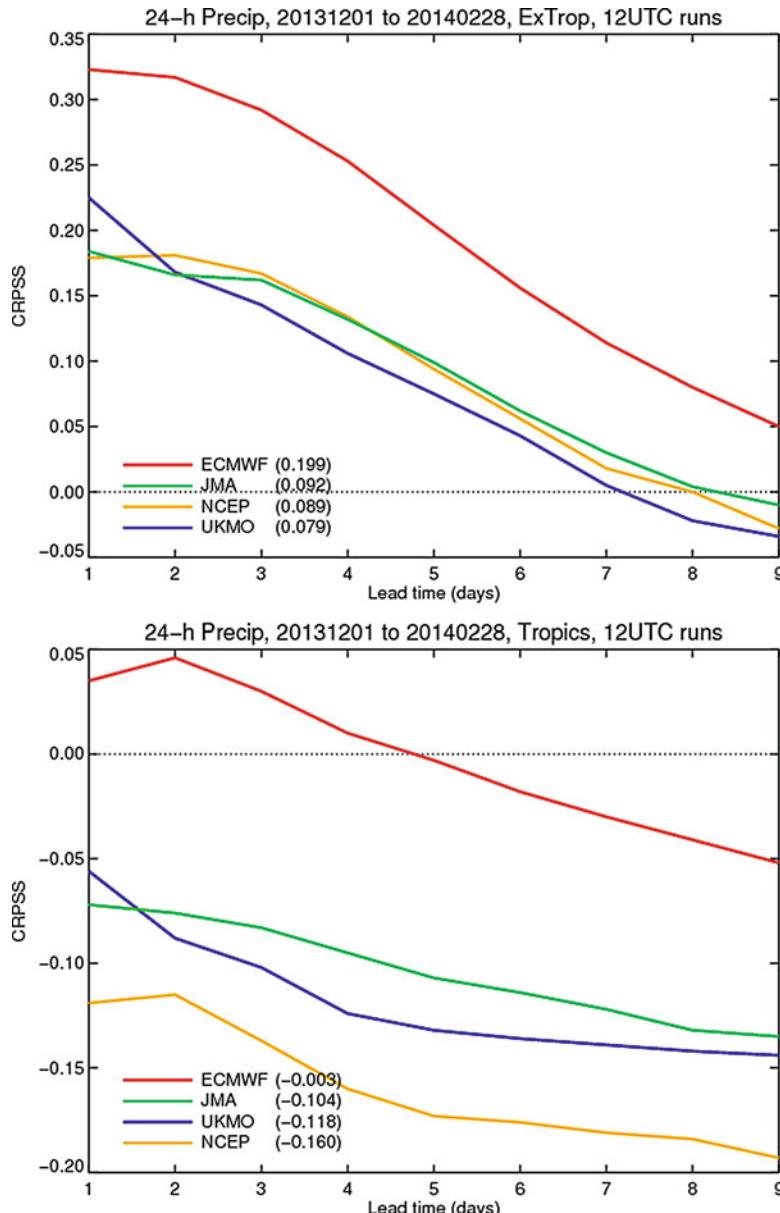


Fig. 16 Winter 2013–2014 (D13JF14) average continuous ranked probability skill score (CRPSS) for 24-h accumulated precipitation over the Northern Hemisphere extra-tropics (20°N – 90°N , top panel) and the tropics (30°S – 30°N , bottom panel), for the four TIGGE ensembles from ECMWF, JMA, NCEP, and UKMO. Verification is against observations at synoptic stations

projections are also based on ensembles, so that not only the most likely scenario but also its uncertainty can be estimated.

All ensembles have kept being improved throughout the years. As an example, let's consider the ECMWF OG-ENS. Figure 17 shows how the continuous ranked

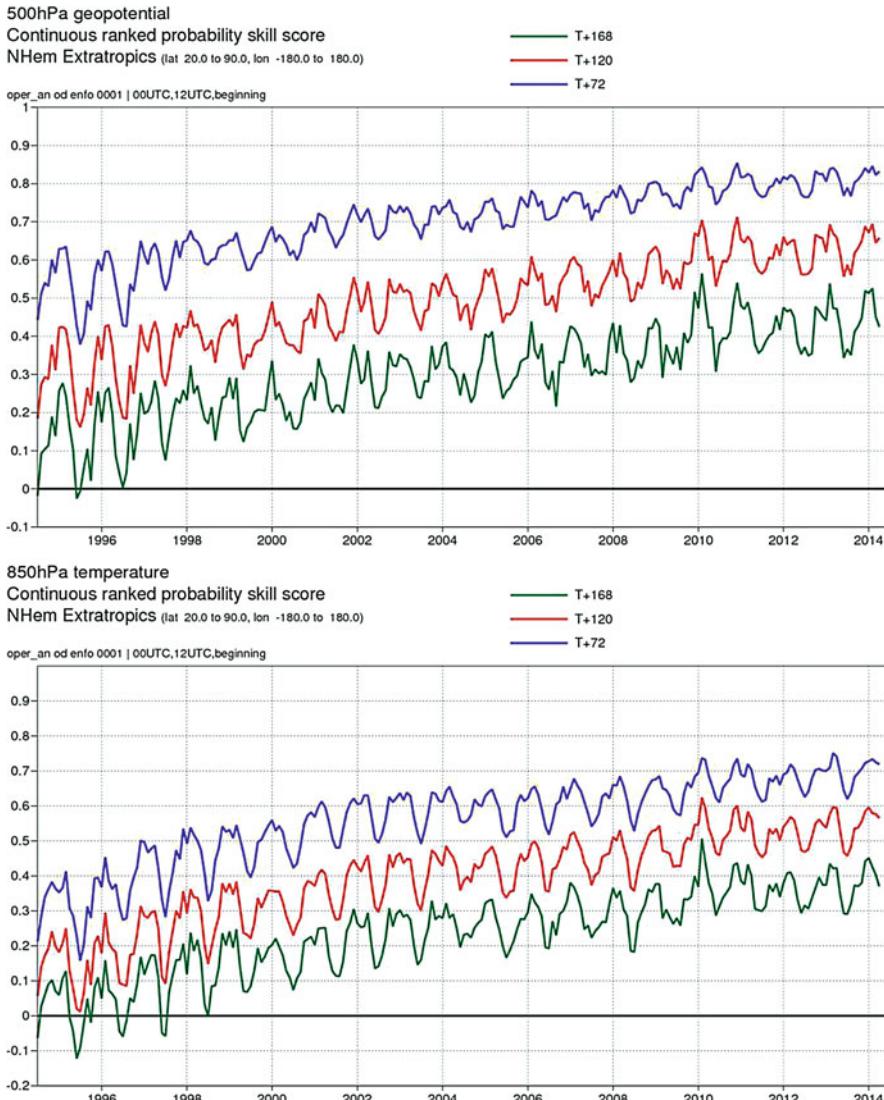


Fig. 17 Time series of the monthly average continuous ranked probability skill score (CRPSS) for the prediction of the 500 hPa geopotential height (top panel) and the 850 hPa temperature (bottom panel) of the ECMWF OG-ENS, over the Northern Hemisphere extra-tropics (20°N – 90°N), for the day 3 (blue line), day 5 (red line), and day 7 (green line) probabilistic forecasts. Verification is against ECMWF operational analyses

probability skill score for the 500 hPa geopotential heights and the 850 hPa temperatures have been improving from 1 May 1994 to date (1 May 1994 is the date when the ECMWF OG-ENS started being run every day, at 12 UTC; between 19 November 1992 and 30 April 1994, the ensemble was only run on Fridays, Saturdays, and Sundays, due to lack of computing resources to produce them daily). By comparing the three lines, it is clear that the forecast skill has been improving by about 1.5–2.5 days per decade.

Improvements have been due to a combination of model (better physical parameterizations and numerical schemes) and data assimilation upgrades (e.g., from the optimum interpolation scheme operational in 1992 to a three-dimensional and then four-dimensional variational system, which now uses a 12-h assimilation window). These changes, combined with the increased number and quality of satellite observations, have led to better initial conditions. Improvements have also been due to increases in horizontal and vertical resolution, made possible by the continuous increases of ECMWF computing systems, and from substantial upgrades in its configuration that have made the ensemble more reliable thanks to advances in the simulation of initial and model uncertainties. Table 2 lists the main changes of the ECMWF OG-ENS configuration.

Figure 17 shows that the improvements were larger between 1994 and 2004. This was the case mainly because the first versions of the ECMWF OG-ENS were rather under dispersive, and it took some years before the ensemble improved its reliability. 1998 saw major improvements in the simulation of initial uncertainties with the introduction of the evolved singular vectors to simulate errors that have been growing during the data assimilation time period (Barkmeijer et al. 1999) and the introduction of first version of the stochastic physics scheme (Buizza et al. 1999). 1999 and 2000 saw two major resolution increases, and 2002 saw the introduction of the initial perturbations in the tropics (Barkmeijer et al. 2001). After 2004, the rate of improvement has been slightly slower, but it is still positive at about 1–1.5 days per decade (the exact value depends on the variable, the region, and the forecast range considered).

In November 1992, the ensemble was based on 33 members, run at T63L19 resolution up to 10 days, with initial uncertainties simulated using initial-time singular vectors computed at T21L19 resolution over the whole globe and with a 36-h optimization time interval. The ensemble did not simulate model uncertainties. Today, the ensemble is based on 51 members, run at T639L91 resolution up to 10 days and extended up to 15 or 32 days with a T319L91 resolution. Initial uncertainties are now simulated using a combination of initial-time singular vectors computed at T42L91 resolution over difference regions and with a 48-h optimization time interval and perturbations defined by the ECMWF ensemble of data assimilations (EDA). The ensemble simulates model uncertainties using two schemes, the revised 3-time level stochastically perturbed parameterized tendencies (revSTP) and the backscatter (BS) schemes. The forecasts are also run coupled to an ocean model (NEMO) from initial time. Initial uncertainties in the ocean are simulated by using initial conditions from the 5-member ocean real-time analysis scheme (ORAS4, Mogensen et al. 2012).

Table 2 Time evolution of some key characteristics of the ECMWF OG-ENS between inception (Nov 1992) and today (May 2014). Columns 3–9 refer to the key characteristics of the initial perturbations added to simulate initial uncertainties: horizontal resolution (HRES) and number of vertical levels (VRES), optimization time interval used to compute the singular vectors (OTI), optimization area (Area), method used to simulate uncertainties that have been growing during the data assimilation period (past) and the future (future), and sampling technique used to define the initial perturbations. Columns 10–16 refer to the key characteristics of the forecasts: horizontal resolution (HRES) and number of vertical levels (VRES), forecast length (Tend, in days), number of members (#), simulation of model uncertainties (Mod Unc), coupling to an ocean model (Coupling), and size of the re-forecast suite in terms of members and years (refc suite)

	Description	Initial uncertainty estimation (SVs and EDA)						Model uncertainty estimation and forecast configuration							
		HRES	VRES	OTI	Area	Past	Future	sampI	HRES	VRES	Tend	#	Mod Unc	Coupling	refc suite
Nov 1992	Oper impl	T21	L19	36h	globe	NO	SVINI	simm	T63	L19	10d	33	NO	NO	NO
Feb 1993	SV LPO	-	-	-	NHK	-	-	-	-	-	-	-	-	-	-
Aug 1994	SV OTI	-	-	48h	-	-	-	-	-	-	-	-	-	-	-
Mar 1995	SV hor resol	T42	-	-	-	-	-	-	-	-	-	-	-	-	-
Mar 1996	IHH+SH SV	-	-	-	(NH+SH)X	-	-	-	-	-	-	-	-	-	-
Dec 1996	resol/mem	L31	-	-	-	-	-	-	TL159	L31	-	51	-	-	-
Mar 1998	EVO SV	-	-	-	SVEVO	-	-	-	-	-	-	-	-	-	-
Oct 1998	Stoch sch SPPT	-	-	-	-	-	-	-	-	-	-	-	-	STP	-
Oct 1999	vert resol	T40	-	-	-	-	-	-	-	-	-	-	-	-	-
Nov 2000	FC hor resol	-	-	-	-	-	-	-	TL255	-	-	-	-	-	-
Jan 2002	TC SVs	-	-	-	(NH+SH)X+TC	-	-	-	-	-	-	-	-	-	-
Sep 2004	sampling	-	-	-	-	-	-	Gauss	-	-	-	-	-	-	-
Jun 2005	rev sampI	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Feb 2006	resolution	L62	-	-	-	-	-	-	TL399	L62	-	-	-	-	-
Sep 2006	VAREPS	-	-	-	-	-	-	-	TL399(0-10)/ TL255(10-15)	-	15d	-	-	-	-
Mar 2008	VAREPS-mon	-	-	-	-	-	-	-	-	-	15d/32d	-	-	-	-
Sep 2009	Rev SPPT	-	-	-	-	-	-	-	-	-	-	-	-	HOPE from d10	5*18y
Jan 2010	hor resol	-	-	-	-	-	-	-	TL639(0-10)/ TL319(10-15)	-	-	-	-	-	-
Jun 2010	EDA EPS	-	-	-	EDA	-	-	-	-	-	-	-	-	revSTP	-
Nov 2010	Rev Stoch scheme	-	-	-	-	-	-	-	TL639(0-10)/ TL319(10-15)	-	15d/32d	-	-	-	-
Nov 2011	New ocean model	-	-	-	-	-	-	-	-	-	-	-	-	NEMO from d10	-
Jun 2012	Rev EDA-perf & refic suite	-	-	-	EDA	-	-	-	-	-	-	-	-	-	5*20y
Nov 2013	vert resol & coupling from d0	T42	L91	48h	(NH+SH)X+TC	EDA	SVINI	Gauss	TL639(0-10)/ TL319(10-15)	L91	15d/32d	51	revSTP+BS	NEMO from d0	5*20y

Some of the ECMWF OG-ENS characteristics have only slightly changed since its inception, e.g., the use of the singular vectors to simulate initial uncertainties, and the fact that they are still computed with a rather coarse resolution when compared to the resolution of the forecast model. Other aspects have changed substantially: a major change in the simulation of initial uncertainties was the replacement of the evolved singular vectors with EDA-based perturbations to simulate initial uncertainties that have been growing during the past 12 h, during the data assimilation period. Another major change has been the introduction of the simulation of model uncertainties: October 1998, when ECMWF introduced the first version of the SPPT scheme (Buizza et al. 1999), was the first time when it was recognized and accepted that stochastic terms, when added to the atmosphere “deterministic” equations, can improve the model forecasts by simulating missing terms. Another major advancement was the coupling of the ensemble to an ocean model in 1998, when the medium-range and the monthly ensembles were merged (Vitart et al. 2008). The number of members has increased only slightly, from 33 to 51. By contrast, horizontal resolution has increased by a factor of 10, from T63 to T639, with another major upgrade being planned for 2015, when the forecasts are supposed to move from a T639L91 (about 35 km) to a T1023 (about 20 km) horizontal resolution.

Since March 2008, when the medium-range and the monthly ensembles have been merged, focus has been widened from 3–10 to 3–30 days, to the so-called sub-seasonal forecast range. This is the range where major improvements have been detected in the past decade, thanks again to a combination of model improvements, the introduction of coupled ocean-land-atmosphere models, advances in data assimilation, and the development of better schemes to simulate model uncertainties. Vitart (2013) indicates that the skill of the ECMWF re-forecasts to predict the Madden-Julian Oscillation has improved significantly since 2002, with an average gain of about 1 day of predictability per year. The ability of the ECMWF model to simulate realistic MJO teleconnections has also improved dramatically over the past 10 years, with, for example, gains in predictability of about 1 week in the prediction of weekly average 2-m temperature anomalies over the Northern Hemisphere extratropics.

The results shown in the previous section and the brief look at the past evolution of the ECMWF OG-ENS configuration indicate that:

- (a) Ensemble-based probabilistic forecasting is providing more complete estimates of the state of the earth-system and of its future states (forecasts) since it includes uncertainty estimates; thus, its use will continue to increase.
- (b) To generate skillful ensemble forecasts, it is essential to have a good model and have good estimates of the initial conditions: this means that to further improve probabilistic forecasting models and data assimilation algorithms will have to be continuously improved.
- (c) Ensembles need to account properly all sources of forecast uncertainties to provide skillful and reliable forecasts: this suggests that to further increase their skill it is necessary to continue to improve the simulation of initial and model uncertainties.

-
- (d) Ensembles can provide skillful sub-seasonal and seasonal forecasts provided that the models include all relevant physical processes: this suggests that to further lengthen the range of skillful forecasts, forecast should use earth-system models that simulate all key processes in the ocean, cryosphere, land, and atmosphere.
 - (e) Ensemble re-forecasts (i.e., forecasts of past cases spanning many years generated using the operational system) are increasingly used to estimate the ensemble's model climatological distribution and assess the ensemble's predictability (Gneiting et al. 2005; Hamill and Whitaker 2007); they will become an integral part of the future systems.

Thus the future will very likely see further developments toward the use of more complex and fully coupled ensembles that simulate all relevant processes of the ocean, cryosphere, land, and atmosphere system. They will start from fully coupled initial conditions and will be able to simulate in a consistent and coherent ways initial and model uncertainties.

Grand, multi-system ensembles could possibly become more common, especially for the extended forecast range (say for the seasonal to interannual, and possibly longer, range). This is the range where results so far have indicated that blending forecasts generated using different systems could help filtering out the unpredictable signals and compensating for each single-system systematic biases. This has indeed been happening in the United States with NAEFS for the medium-range and has been happening in Europe for the seasonal range with EUROSIP, the EUROpean Seasonal to Interannual Prediction system, based on the four seasonal ensembles generated by ECMWF, Meteo France, NCEP, and UK Met Office. Multi-system forecast products could also be one of the main deliverables of the WMO sub-seasonal to seasonal (S2S) project, one of the proposed WWRP/THORPEX and WCRP joint research project to improve forecast skill and understanding on the sub-seasonal to seasonal time scale (see the WMO web site, http://www.wmo.int/pages/prog/arep/wwrp/new/documents/S2S_Implem_plan_en.pdf).

5 Final Remarks on Operational Global Medium-Range Ensembles

This chapter should have given the reader a brief but comprehensive overview of the global medium-range ensembles operational at the time of writing (May 2014). Readers now should have understood that there is not (yet at least) the “perfect recipe” to construct the best ensemble!

Different techniques can be used and have been used, to simulate initial and model uncertainties. The performance of the operational, global ensembles that we have analyzed reflects these choices but more importantly reflects the quality of the forecast model and of the data assimilation system used to provide the initial-time estimates of the state of the atmosphere and its uncertainty.

This chapter has not discussed how to design good quality models that can simulate all relevant processes of the earth-system, of the whole coupled ocean,

cryosphere, land, and atmosphere system. It has not discussed how to use good models to estimate the initial state of the earth-system by blending all available observations with the model first guess. The reader interested in these topics should look at the published literature on numerical weather and climate prediction and data assimilation.

My closing remark is that if we aim to predict the evolution of earth-system, a probabilistic approach must be followed, and ensembles are, at least so far, the only practical tool to estimate the initial time and forecast probability density function of the earth-system. This is the main reason why I expect them to be increasingly used at initial time and at all forecast ranges, since we will always have to deal with initial and model uncertainties in our predictions!

References

- J. Barkmeijer, R. Buizza, T.N. Palmer, 3D-Var Hessian singular vectors and their potential use in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **125**, 2333–2351 (1999)
- J. Barkmeijer, R. Buizza, T.N. Palmer, K. Puri, J.-F. Mahfouf, Tropical singular vectors computed with linearized diabatic physics. *Q. J. R. Meteorol. Soc.* **127**, 685–708 (2001)
- C.H. Bishop, B.J. Etherton, S.J. Majumdar, Adaptive sampling with the ensemble transform kalman filter. Part I: Theoretical aspects. *Mon. Weather Rev.* **129**, 420–436 (2001)
- G. Boffetta, P. Guliani, G. Paladin, A. Vulpiani, An extension of the Lyapunov analysis for the predictability problem. *J. Atmos. Sci.* **55**, 3409–3416 (1998)
- P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D. Chen, E. Ebert, M. Fuentes, T. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P. Silva Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, S. Worley, The THORPEX interactive grand global ensemble (TIGGE). *Bull. Am. Meteorol. Soc.* **91**, 1059–1072 (2010)
- W. Bourke, T. Hart, P. Steinle, R. Seaman, G. Embrey, M. Naughton, L. Rikus, Evolution of the bureau of meteorology's global assimilation and prediction system. Part 2: Resolution enhancements and case studies. *Aust. Meteorol. Mag.* **44**, 19–40 (1995)
- W. Bourke, R. Buizza, M. Naughton, Performance of the ECMWF and the BoM ensemble systems in the Southern Hemisphere. *Mon. Weather Rev.* **132**, 2338–2357 (2004)
- N.E. Bowler, A. Arribas, K.R. Mylne, K.B. Robertson, *Numerical Weather Prediction: The MOGREPS Short-Range Ensemble Prediction System. Part I: System Description*. UK Met. Office NWP Technical Report No. 497 (2007), p. 18
- N.E. Bowler, A. Arribas, K.R. Mylne, K.B. Robertson, G.J. Shutts, The MOGREPS short-range ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**, 703–722 (2008)
- D.R. Bright, S.L. Mullen, Short-range ensemble forecasts of precipitation during the southwest monsoon. *Weather Forecast.* **17**, 1080–1100 (2002)
- R. Buizza, T.N. Palmer, The singular vector structure of the atmospheric general circulation. *J. Atmos. Sci.* **52**, 1434–1456 (1995)
- R. Buizza, M. Miller, T.N. Palmer, Stochastic representation of model uncertainties in the ECMWF EPS. *Q. J. R. Meteorol. Soc.* **125**, 2887–2908 (1999)
- R. Buizza, P.L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, Y. Zhu, A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Weather Rev.* **133**, 1076–1097 (2005)
- R. Buizza, J.-R. Bidlot, N. Wedi, M. Fuentes, M. Hamrud, G. Holt, F. Vitart, The new ECMWF VAREPS (variable resolution ensemble prediction system). *Q. J. R. Meteorol. Soc.* **133**, 681–695 (2007)
- R. Buizza, M. Leutbecher, L. Isaksen, Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **134**, 2051–2066 (2008)

- M.M. Coutinho, Ensemble prediction using principal-component-based perturbations. Thesis in Meteorology, National Institute for Space Research (INPE), 1999, pp. 136 (in Portuguese)
- T. Gneiting, A.E. Raftery, A.H. Westweld III, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather. Rev.* **133**, 1098–1118 (2005)
- T.-Y. Goo, S.-O. Moon, J.-Y. Cho, H.-B. Cheong, W.-J. Lee, Preliminary results of medium-range ensemble prediction at KMA: Implementation and performance evaluation as of 2001. *Korean J. Atmos. Sci.* **6**, 27–36 (2003)
- M.E.B. Gray, G.J. Shutts, *A Stochastic Scheme for Representing Convectively Generated Vorticity Sources in General Circulation Models*. APR Turbulence and Diffusion Note No. 285. (Met Office, FitzRoy Road, Exeter EX1 3PB, UK, 2002)
- R. Hagedorn, R. Buizza, M.T. Hamill, M. Leutbecher, T.N. Palmer, Comparing TIGGE multi-model forecasts with re-forecast calibrated ECMWF ensemble forecasts. *Q. J. R. Meteorol. Soc.* **138**, 1814–1827 (2012)
- T.M. Hamill, J.S. Whitaker, Ensemble calibration of 500 hPa geopotential height and 850 hPa and 2-meter temperatures using re-forecasts. *Mon. Weather. Rev.* **135**, 3273–3280 (2007)
- D. Hou, Z. Toth, Y. Zhu, W. Yang, Impact of a stochastic perturbation scheme on NCEP global ensemble forecast system. In Proceedings of the 19th AMS Conference on Probability and Statistics, 21–24 January 2008, New Orleans, Louisiana. (2008)
- P.L. Houtekamer, L. Lefavre, Using ensemble forecasts for model validation. *Mon. Weather. Rev.* **125**, 2416–2426 (1997)
- P.L. Houtekamer, L. Lefavre, J. Derome, H. Ritchie, H.L. Mitchell, A system simulation approach to ensemble prediction. *Mon. Weather. Rev.* **124**, 1225–1242 (1996)
- P.L. Houtekamer, H.L. Mitchell, X. Deng, Model error representation in an operational ensemble Kalman filter. *Mon. Weather. Rev.* **137**, 2126–2143 (2009)
- P.L. Houtekamer, X. Deng, H.L. Mitchell, S.-J. Baek, N. Gagnon, Higher resolution in an operational ensemble Kalman filter. *Mon. Weather. Rev.* **142**, 1143–1162 (2014)
- L. Isaksen, M. Bonavita, R. Buizza, M. Fisher, J. Haseler, M. Leutbecher, L. Raynaud, Ensemble of data assimilations at ECMWF. *ECMWF Research Department Technical Memorandum No. 636* (2010). Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also <http://old.ecmwf.int/publications/>)
- P. Janssen, J.-R. Bidlot, S. Abdalla, H. Hersbach, Progress in ocean wave forecasting at ECMWF. *ECMWF Research Department Technical Memorandum No. 478* (2005). Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also <http://old.ecmwf.int/publications/>)
- P.A.E.M. Janssen, O. Breivik, K. Mogensen, F. Vitart, M. Balmaseda, J.-R. Bidlot, S. Keeley, M. Leutbecher, L. Magnusson, F. Molteni, Air-sea interaction and surface waves. *ECMWF Research Department Technical Memorandum No. 712* (2013). Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also <http://old.ecmwf.int/publications/>)
- J. Kai, H. Kim, Characteristics of initial perturbations in the ensemble prediction system of the Korea Meteorological Administration. *Weather Forecast* **29**, 563–581 (2014). <https://doi.org/10.1175/WAF-D-13-00097.1>
- M. Leutbecher, T.N. Palmer, Ensemble forecasting. *J. Comput. Phys.* **227**, 3515–3539 (2008)
- J.W.B. Lin, J.D. Neelin, Influence of a stochastic moist convective parameterization on tropical climate variability. *Geophys. Res. Lett.* **27**, 3691–3694 (2000)
- K. Mogensen, M. Alonso Balmaseda, A. Weaver, The NEMOVAR ocean data assimilation system as implemented in the ECMWF ocean analysis for System 4. *ECMWF Research Department Technical Memorandum No. 668* (2012). Available from ECMWF, Shinfield Park, Reading RG2-9AX (see also <http://old.ecmwf.int/publications/>)
- F. Molteni, R. Buizza, T.N. Palmer, The ECMWF ensemble prediction system: Methodology and validation. *Q. J. R. Meteorol. Soc.* **122**, 73–119 (1996)
- T.N. Palmer, F. Molteni, R. Mureau, R. Buizza, P. Chapelet, J. Tribbia, Ensemble prediction, in *Proceedings of the ECMWF Seminar on Validation of Models Over Europe*, Vol. 1 (ECMWF, Shinfield Park, 1993), 285 pp. Available from ECMWF, Shinfield Park, Reading RG2-9AX, UK

- T.N. Palmer, R. Buizza, F. Doblas-Reyes, T. Jung, M. Leutbecher, G.J. Shutts, M. Steinheimer, A. Weisheimer, Stochastic parameterization and model uncertainty. *ECMWF Research Department Technical Memorandum No. 598* (2009), p. 42. Available from ECMWF, Shinfield Park, Reading RG2-9AX, UK
- Y.-Y. Park, R. Buizza, M. Leutbecher, TIGGE: Preliminary results on comparing and combining ensembles. *Q. J. R. Meteorol. Soc.* **134**, 2029–2050 (2008)
- G. Shutts, A kinetic energy backscatter algorithm for use in ensemble prediction systems. *Q. J. R. Meteorol. Soc.* **131**, 3079–3100 (2005)
- X. Su, H. Yuan, Y. Zhu, Y. Luo, Y. Wang, Evaluation of TIGGE ensemble predictions of Northern Hemisphere summer precipitation during 2008–2012. *J. Geophys. Res. Atmos.* **119**, 7292 (2014). (in publication)
- Z. Toth, E. Kalnay, Ensemble forecasting at NMC: The generation of perturbations. *Bull. Am. Meteorol. Soc.* **74**, 2317–2330 (1993)
- Z. Toth, E. Kalnay, Ensemble forecasting at NCEP and the breeding method. *Mon. Weather Rev.* **125**, 3297–3319 (1997)
- F. Vitart, Evolution of ECMWF sub-seasonal forecast skill scores over the past 10 years. *ECMWF Research Department Technical Memorandum No. 694* (2013), p. 28. Available from ECMWF, Shinfield Park, Reading RG2-9AX, UK
- F. Vitart, R. Buizza, M. Alonso Balmaseda, G. Balsamo, J.R. Bidlot, A. Bonet, M. Fuentes, A. Hofstadler, F. Molteni, T.N. Palmer, The new VAREPS-monthly forecasting system: A first step towards seamless prediction. *Q. J. R. Meteorol. Soc.* **134**, 1789–1799 (2008)
- X. Wang, C.H. Bishop, S.J. Julier, Which is better, an ensemble of positive/negative pairs or a centered spherical simplex ensemble? *Mon. Weather Rev.* **132**, 1590–1605 (2004)
- M. Wei, Z. Toth, R. Wobus, Y. Zhu, C. Bishop, X. Wang, Ensemble transform Kalman filter-based ensemble perturbations in an operational global prediction system at NCEP. *Tellus A* **58**, 28–44 (2006)
- M. Wei, Z. Toth, R. Wobus, Y. Zhu, Initial perturbations based on the ensemble transform (ET) technique in the NCEP global operational forecast system. *Tellus A* **60**, 62–79 (2008)
- M. Yamaguchi, S.J. Majumdar, Using TIGGE data to diagnose initial perturbations and their growth for tropical cyclone ensemble forecasts. *Mon. Wea. Rev.* **138**, 3634–3655 (2010)
- Z. Zhang, T.N. Krishnamurti, A perturbation method for hurricane ensemble predictions. *Mon. Weather Rev.* **127**, 447–469 (1999)



Intraseasonal to Interannual Climate Variability and Prediction

Malaquias Peña, L. Gwen Chen, and Huug van den Dool

Contents

1	Introduction	197
2	ISI Variability: Highlights	200
2.1	El Niño-Southern Oscillation (ENSO) Phenomenon	201
2.2	Local Forcing and Feedbacks	207
2.3	Teleconnection Patterns	208
3	Rationale of ISI Predictions	209
3.1	Sources of Predictability	210
3.2	Observing Research	211
4	Modeling ISI Variability	212
4.1	Statistical Climate Prediction Models	212
4.2	Dynamical Climate Prediction Models	217
5	Climate Prediction Systems	220
5.1	Observational Data	221
5.2	Model Initialization and Data Assimilation Schemes	223
5.3	Ensemble Initial Perturbations	225
6	Product Generation Tools	226
6.1	Reanalysis and Hindcast Data Sets	226
6.2	Multi-Model Ensembles	227

M. Peña (✉)

Department of Civil and Environmental Engineering, University of Connecticut, Storrs, CT, USA
e-mail: malaquias.pena.mendez@noaa.gov

L. G. Chen

Earth System Science Interdisciplinary Center/Cooperative Institute for Climate and Satellites,
University of Maryland, College Park, MD, USA

CPC/NCEP/NWS/NOAA, College Park, MD, USA
e-mail: lichuan.chen@noaa.gov

H. van den Dool

CPC/NCEP/NWS/NOAA, College Park, MD, USA
e-mail: huug.vandendool@noaa.gov

6.3	Consolidation Methods	228
6.4	Decision Support Systems	229
7	Summary	233
	References	233

Abstract

This chapter outlines a set of topics essential to become aware of the science, methods, and procedures for operational prediction in the intraseasonal to interannual (ISI) time range. The quality of ISI predictions rely on three basic capabilities: observing networks to sample the Earth's climate system, an analysis scheme to summarize past and present observations into physically consistent time series of spatial fields, and a prediction method to project a present state of the climate system into the future. Observing networks provide essential data to estimate the true state of the climate system at regular time intervals and to measure physical climate processes and climate variability. Conventional observing networks are designed to sample the most relevant scales of variability and processes occurring in the climate system. Numerical analysis schemes generate physically consistent estimates of the state of the climate system based on observations; they vary in complexity from simple interpolation methods to modern data assimilation schemes. The dynamical approach to carry out ISI predictions use numerical schemes that couple atmosphere, land, ocean and cryosphere models. ISI predictions are also carried out using statistical methods or a combination of the two approaches. The computer burden associated with carrying out dynamical forecasts with comprehensive coupled models is high. Thus, operational centers perform those coupled model runs routinely out to a few seasons only.

Current prediction practices include running the coupled models in ensemble mode to account for the uncertainty in the forecasts and to filter out unpredictable signals through ensemble averaging. Furthermore, recognizing the difficulty for a single model to measure its own forecast limitations, it is common to combine ensembles from multiple independent models in a scheme called multi-model ensemble. The practice of incorporating reanalysis and hindcast data sets as tools to post-process raw forecast outputs considerably reduces forecast systematic errors, improves reliability, and enhances the estimation of potential skill and the detection of extreme events. The outputs of coupled global models often serve as input to downstream models such as limited-area high-resolution climate models, river routing, crop growth, and expanding the applicability of ISI forecasts to the regional and local level. Graphical interfaces that permit data analysis and smart decision support systems are becoming necessary to assist forecasters and decision-makers in their real-time endeavors.

As models become more skillful and reliable, the methods to generate climate numerical guidance have evolved from subjective approaches to increasingly objective and unsupervised numerical procedures. Nonetheless, human intervention typically increases the skill and value of the final products and is essential for product interpretation and communication to final users. Challenges to

realistically model the climate system are many, but those highlighted by the scientific community include better model representation of fine-scale processes in clouds, ocean eddies, and surface interactions and feedbacks and better coupled integration of model climate components. More skillful ISI forecasts are also conditioned to greater computer resources, more extensive and strategic observing systems, and effective data assimilation and model initialization schemes for the coupled climate prediction systems.

Keywords

Climate variability · ENSO · MJO · Teleconnections · Seasonal predictions · Predictability · Earth system models · Ocean-atmosphere coupling · Surface local feedbacks · Observing networks · Multi-model ensembles · Hindcasts · Reanalysis

1 Introduction

Seasonal prediction has a rather long history that began possibly with empirical anticipations of the rainy season in ancient civilizations. A fundamental step in this direction was the realization from astronomic observations of the existence of the annual cycle and its connection with surrounding phenologic and environmental changes. This awareness and the recording of events were essential to establish successful water irrigation and agricultural systems. Anticipating departures of the quasiperiodic rainy season causing floods or famine were critical to the survival of early societies. Today, predicting such departures with sufficient accuracy are needed for the planning of a wide range of economical and social activities but continue to be challenging. The difficulty to understand and to predict such departures, despite that the sun's radiation reaches a point on the Earth's surface periodically, is that climate is regulated by complex interactions among the components (atmosphere, land, ocean, cryosphere (snow and ice), lithosphere) of the Earth's system, which span many temporal and spatial scales. Studying departures, more commonly known as anomalies from a norm climate, allows characterizing their properties and understanding their attributions. Associating physical attributions to anomalies took centuries until inquisitive researchers were able to compile sufficient observations all over the globe to develop and test their theories. The study of anomalies continues to bear useful insights to infer causal mechanisms and to identify regular patterns and behaviors in the time series of the recorded climate data.

In the intraseasonal to interannual (ISI) time range, anomalies are usually computed with respect to the annual cycle. Computing the annual cycle of a climate variable requires making assumptions to deal with the fact that some climate variables are the result of nonlinear interactions therefore are not periodic; some have marked trends and some have short time series records. These complexities make the estimation of the annual cycle highly dependent upon the period of the data records used to compute it. Conventionally, the annual cycle is computed as a 30-year

average of data, e.g., from 1981 to 2010. Based on this convenient reference, the time series of a climate variable shows anomalies when the annual cycle is subtracted. Anomalies can be classified according to their amplitude, life span, and frequency. Defining thresholds to these three properties allows the extraction of extreme amplitude anomalies, persistent anomalies, and rare events, respectively. Extreme anomalies are often associated with hazardous conditions that affect society and the economy. In some applications, the 30-year norm is adjusted to make up for the nonlinear evolution of the climate system. One of these adjustments is the removal of trends aka detrending. Statistical analysis methods treat separately the trend, which is usually removed during the preprocessing stage of the data to make the time series stationary. Trends evolve much slower than the interannual, year-to-year variability, making them useful for long-term projections.

The climate mean state and its variability have been more thoroughly recorded with the establishment of global and regional observing networks. Global measurements of climate parameters have been facilitated by satellite observations, which started transmitting meteorological information in the late seventies. Data from historical records and routine and experimental observing networks are the main resource to describe, understand, and predict climate variability. The increasing amount of data demands techniques to remove errors, reduce redundancy, and synthesize and represent in mathematical form what has been observed. Advanced statistical methods are used to detect and understand relationships among climate quantities that are hidden in the multitude of spatial and temporal scales of the data. One important development in this area was treating the time series of data as the result of multiple oscillations as it is done in spectral analysis. The direct application of spectral analysis however is complicated because most processes are nonperiodic. Compounding this problem is that observations have errors of different types including capturing variability that may not be consequential to the large-scale evolution of the climate system. Thus, time series filters are needed to enhance certain important signals. The introduction of principal component decomposition strategies and empirical orthogonal functions (EOFs; Lorenz 1956) made it possible to identify spatial patterns of variability. These patterns reveal main physical processes but may require additional statistical processes such as truncation and rotation of the EOFs to give more physically interpretable results.

Time series of observational data reveal a multitude of phenomena in the ISI time range. The most prominent phenomenon at interannual time scale is El Niño-Southern Oscillation (ENSO), which affects local and regional weather patterns, ecosystems, and the economy. A prominent phenomenon in the intraseasonal time scales is the Madden-Julian Oscillation (Madden and Julian 1971). Hydrometeorological phenomena including droughts, floods, monsoons, and hurricane activity are influenced by large-scale phenomena and climate interactions that occur in the ISI time frame. In practical terms, analysis of time series of observational data focuses on key variables such as precipitation over certain geographical region and then infers from the data the possible causes or drivers of the variability observed. Causes of variability can be due to local processes or to remote mechanisms; the latter often referred to as teleconnections.

In parallel with statistical developments, the use of computers to obtain numerical solutions to dynamic and thermodynamic equations representing climate processes was conducted during the second half of the twentieth century (e.g., Manabe and Bryan 1969). Both statistical and modeling developments modernized the science of climate. In particular, dynamical models provided a new perspective from a Koeppen's catalog of "weather types" to a dynamical system perspective where the description of climate is made in terms of state variables of the combined Earth's system components. As computer capabilities increased, these methods and models grew in complexity. They are part of a data processing framework to form the numerical weather and climate prediction systems known today.

Climate prediction is an estimate of the conditions of future states of the climate system. In the context of ISI time scales, it is referred to as seasonal outlook or forecast to indicate that in practice predictions are made from one season out to a year in advance. These predictions are also referred to as short-term climate predictions to differentiate them from climate change and multi-decadal and beyond predictions. From a producer's perspective, short-term climate predictions have an important advantage over longer-term predictions: the length of the time series of global observations, which spans several decades, supplies several cases of ENSO events and other important phenomena to test their hypothesis and validate their predictions.

Short-term climate predictions attempt to foresee the strength, length, frequency, onset, and end of anomalies or prominent events. In operational settings, a small set of climate variables, which are deemed critical, are provided to users as real-time products and stored for further analysis. The most common variables include precipitation, near-surface temperature and humidity, and sea surface temperature (SST). In addition to predicted climate variables, forecast products include indexes such as El Niño 3, El Niño 3.4, the MJO signal, monsoon indexes, drought indexes, etc. These variables and products, also referred to as numerical guidances, allow human experts to monitor the state of the climate and make consensual statements about future weather or climate events for the general public or for specific decision-makers. Researchers and operational climatologists are working to meet user's demand for more accurate and informative forecasts, more products, and longer databases of past forecasts.

Ensemble forecasting methods are relatively recent tools to do weather and climate predictions (see chapter ► "[Overview of Weather and Climate Systems](#)," in this book). Ensembles changed the way in which monthly and seasonal prediction products were generated. The mean and spread of the forecasts are now part of the routine information provided to users. This is reflected in the diagram shown in Fig. 1. All numerical products for the medium range to seasonal forecasts are based on ensemble forecasting outputs. Typically products are organized as weeks 2, weeks 3 and 4 (or monthly), and seasonal forecasting. There is a variety of ongoing multi-model ensemble projects including the NAEFS, NUOPC, EUROSIP, and APEC Climate Center that cover that range of predictions. Attempts to consolidate them in an optimal fashion such as weighted averages and recalibration procedures abound. Ensemble averaging is an effective filtering approach that generally increases the skill of dynamical model forecasts.

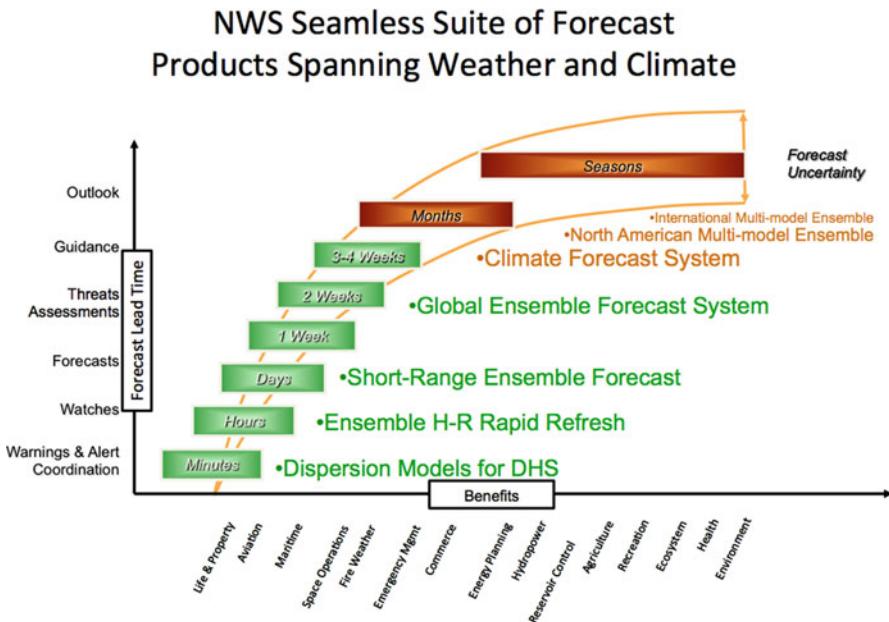


Fig. 1 Schematic of the US seamless suite of forecast products spanning weather and climate. (Adapted from (NCEP-NWS) presentations)

The following sections of this chapter continue the above discussion. Section 2 gives a concise account of the modes of ISI variability that have been detected in the long record of observations. Section 3 gives an overview of the rationale behind predictions, particularly ensemble seasonal predictions. Section 4 is devoted to the description of main types of modeling approaches of ISI variability. Modeling approaches were separated into statistical and dynamical models. Section 5 describes typical climate prediction systems, which are becoming extensions of the NWP systems described in previous chapters. Section 6 presents procedures and tools for product generation and a series of applications and case studies that have been developed in recent years.

2 ISI Variability: Highlights

As described in the introduction section of this chapter, a climate anomaly is a generic term indicating that one or more climate variables undergo a departure from their reference values. This section attempts to provide a description of those reference values or regular patterns in time and space beyond the annual cycle. Climate variables directly associated with the annual cycle are periodic, and many others are oscillatory in character; it is therefore common to decompose their time series into modes of variability. There are two difficulties in this decomposition. One

is that the nonlinear behavior of the climate system evolution produces a continuous spectrum with no evidence of sharp peaks except for the mid-latitude baroclinic instabilities and a few other quasiperiodic features. This difficulty makes it hard to readily identify the dominant modes and thus makes it necessary to perform time filtering in order to extract specific modes. The second problem is that climate variability is not spatially uniform and that variations in one region may be interrelated to variations in other remote regions. Tools that take into account spatial and time covariabilities are needed. One of the hallmarks of the analysis of climate time series is the use of empirical orthogonal functions (EOFs) to decompose the time series into preferred spatial patterns. The most prominent of these are known as modes of climate variability. A list of the more prominent modes of variability is given in Table 1. Except for the seasonal cycle, all patterns are quasiperiodic, and all cover extensive regions and are identified as spatial teleconnections.

2.1 El Niño-Southern Oscillation (ENSO) Phenomenon

ENSO is the strongest quasi predictable climate variation on the interannual climate time scale ranging from a few months to several years. ENSO is a fundamental oscillatory mode of the coupled ocean-atmosphere system manifesting itself as an irregular interannual oscillation of tropical Pacific sea surface temperatures (SSTs) and trade wind and sea-level pressure variation between the Indian and Pacific Oceans. The warm and cold phases of ENSO are referred to as “El Niño” and “La Niña” or warm and cold events, respectively. El Niño or La Niña is defined by the SST anomalies in the central and eastern equatorial Pacific. The “Southern Oscillation” part in ENSO is the accompanied atmospheric oscillation defined by the surface pressure differences in two points in the equatorial southern Pacific. ENSO affects regional and global climate and in turn affects the ecosystems in and around the tropical Pacific and the economies of several countries. Many research institutions continue to improve ENSO prediction through improvements of statistical prediction models and more extensively through the development and improvement of coupled global models.

The ENSO phenomenon was first detected in the nineteenth century in terms of sea-level pressure only (i.e., in the atmosphere) and immediately applied for long-range forecasting of monsoon rainfall in places like India, with mixed success. The connection to the ocean was suspected but not firmly established until the mid-1980s. In part it had to wait this long because global analyses of atmosphere and ocean were not available until 1980. The particularly strong El Niño of 1982/1983 came at the right time. A real-time event of that magnitude in combination with the availability of emerging global monitoring was an enormous stimulus for understanding and application in short-term climate prediction.

ENSO events come and go and last from a few months to a few years. There are also periods, sometimes years, without ENSO. An exact definition of an ENSO event has been hard to agree on. The NOAA criterion to define El Niño events is whenever SST exceeds 0.5 C in five successive 3-month periods. The criterion applies to the areal mean SST from 170 W to 120 W and 5S to 5 N, the so-called Nino 3.4 area.

Table 1 Definition of main patterns and indices of climate variability. (Adapted from A. Kaplan in Blunden et al. 2011)

Climate phenomenon	Index name	Index definition	Characterization/ comments
El Niño – Southern Oscillation (ENSO) Canonical, eastern pacific ENSO	NINO3	SST anomaly averaged over (5°S–5°N, 150°W–90°W)	Traditional SST-based ENSO index
	NINO3.4	SST anomaly averaged over (5°S–5°N, 170°W–120°W)	Used by NOAA to define El Niño/La Niña events. Detrended form is close to the 1st PC of linearly detrended global field of monthly SST anomalies (Deser et al. 2010)
	Cold Tongue Index (CTI)	SSTA (6°N–6°S, 180°–90°W) minus global mean SSTA	Matches “cold tongue” area, subtracts effect of the global average change
	Troup SOI	Standardized for each calendar month MSLP difference: Tahiti minus Darwin, $\times 10$	Used by Australian Bureau of Meteorology
	SOI	Standardized difference of standardized MSLP anomalies: Tahiti minus Darwin	Maximizes signal to noise ratio of linear combinations of Darwin/Tahiti records
	Darwin SOI	Standardized Darwin MSLP anomaly	Introduced to avoid use of the Tahiti record, considered suspicious before 1935
Central pacific El Niño (Modoki)	Equatorial SOI (EQSOI)	Standard difference of standard MSLP anomalies over equatorial (5°S–5°N) Pacific Ocean; east (130°W–80°W) minus west (90°E–140°E)	
	El Niño Modoki Index (EMI)	SSTA; [165°E–140°W, 10°S–10°N] minus $\frac{1}{2}$ [110°W–70°W, 15°S–5°N] minus $\frac{1}{2}$ [125°E–145°E, 10°S–20°N]	A recently identified ENSO variant: Modoki or central pacific El Niño (non-canonical)

(continued)

Table 1 (continued)

Climate phenomenon	Index name	Index definition	Characterization/ comments
Pacific decadal and interdecadal variability	Pacific Decadal Oscillation (PDO)	1st PC of the N. Pacific SST anomaly field (20°N – 70°N) with subtracted global mean	
	Intedecadal Pacific Oscillation (IPO)	The 3rd EOF3 of the 13-year low-pass filtered global SST, projected onto annual data	
	North Pacific Index (NPI)	SLP (30°N – 65°N , 160°E – 140°W)	
North Atlantic Oscillation	Lisbon/Ponta Delgada-Stykkisholmur/Reykjavik North Atlantic Oscillation (NAO) Index	Lisbon/Ponta Delgada minus Stykkisholmur/Reykjavik standardized MSLP anomalies	A primary NH teleconnection both in MSLP and Z500 anomalies (Wallace and Gutzler 1981); one of routed EOFs of NH Z500 (Barnston and Livezey 1987). MSLP anomalies can be monthly, seasonal, or annual averages. Each choice carries to the temporal resolution of the NAO index produced that way.
	Gibraltar-Reykjavik NAO Index	Gibraltar minus Reykjavik standardized MSLP anomalies	
	PC-Based NAO Index	Leading PC of MSLP anomalies over the Atlantic sector (20°N – 80°N , 90°W – 40°E)	
Annular modes: Arctic Oscillation (AO), a.k.a. Northern Annular Mode (NAM) Index, and Antarctic Oscillation (AAO), a.k.a. Southern Annular Mode (SAM) Index	PC-Based AO Index	1st PC of the monthly mean MSLP anomalies poleward of 20°N	Closely related to the NAO

(continued)

Table 1 (continued)

Climate phenomenon	Index name	Index definition	Characterization/ comments
	PC-Based AAO Index	1st PC of 850 hPa or 700 hPa height anomalies south of 20°S	
	Grid-Based AAO Index: 40°S–65°S difference	Difference between normalized zonal mean MSLP at 40°S and 65°S, using gridded SLP analysis	
	Grid-Based AAO Index: 40°S–70°S difference	Same as above but uses latitudes 40°S and 70°S	
	Station-Based AAO Index: 40°S–65°S	Difference in normalized zonal mean MSLP at 40°S and 65°S, using station data	
Pacific/North America (PNA) atmospheric teleconnection	PNA Pattern Index	$\frac{1}{4}[Z(20^{\circ}\text{N}, 160^{\circ}\text{W}) - Z(45^{\circ}\text{N}, 165^{\circ}\text{W}) + Z(55^{\circ}\text{N}, 115^{\circ}\text{W}) - Z(30^{\circ}\text{N}, 85^{\circ}\text{W})]$. Z is the location's standardized 500 hPa geopotential height anomaly	A primary NH teleconnection both in MSLP and Z500 anomalies
Atlantic Ocean thermohaline circulation	Atlantic Multi-decadal Oscillation (AMO) Index	10-year running mean of de-trended Atlantic mean SST anomalies (0° – 70°N)	Called “virtually identical” to the smoothed first rotated N. Atlantic EOF mode
	Revised AMO Index	As above, but subtracts global mean anomaly instead of de-trending	
Tropical Atlantic Ocean non-ENSO variability	Atlantic Nino Index. ATL3	SSTA (3°S – 3°N , 20°W – 0°)	Identified as the two leading PCs of detrended tropical Atlantic monthly SSTA (20°S – 20°N): 38% and 25% variance respectively for HadISSTI, 1900–2008 (Deser et al. 2010)
	Atlantic Nino Index PC-based	1st PC of the detrended tropical Atlantic monthly SSTA (20°S – 20°N)	

(continued)

Table 1 (continued)

Climate phenomenon	Index name	Index definition	Characterization/ comments
	Tropical Atlantic Meridional Mode (AMM)	2nd PC of the detrended tropical Atlantic monthly SSTA (20°S – 20°N)	
Tropical Indian Ocean non-ENSO variability	Indian Ocean Basin Mode (IOBM) Index	The 1st PC of the IO detrended SST anomalies (40°E – 110°E , 20°S – 20°N)	Identified as the two leading PCs of detrended tropical Indian Ocean monthly SSTA (20°S – 20°N): 39% and 12% of the variance, respectively, for HadISST1, 1900–2008 (Deser et al. 2010)
	Indian Ocean Dipole mode (IODM), PC-Based Index	The 2nd PC of the IO detrended SST anomalies (40°E – 110°E , 20°S – 20°N)	
	Indian Ocean Dipole Mode Index (DMI)	SST anomalies: 50°E – 70°E , 10°S – 10°N –(90°E – 110°E , 10°S – 0°)	
Cold Ocean – Warm Land (COWL) Variability	COWL Index	Linear best fit to the field of deviations of NH temperature anomalies from their spatial mean; the COWL pattern itself is proportional to the covariance pattern of the NH spatial mean with these deviations.	Useful for removing some effects of natural climate variability from spatially averaged temperature records.

While ENSO is mainly in the tropics, the impacts are felt through teleconnections in the entire tropics and in season-dependent fashion also in the mid-latitudes of both hemispheres. In terms of SST, ENSO is strongest in NH winter around December. The impact in NH mid-latitudes is felt strongest in the NH winter time, with maximum impact in February.

The most common and easy to communicate prediction tool is the ENSO composite. An example may help to clarify. In the middle of 2015, it was quite certain that winter 2015/2016 would be a strong El Niño. A prediction is then made empirically for the next winter by expecting conditions equal to the composite weather in 10–20 previous El Niño's. Figure 2 shows the composites of the warm and cold events.

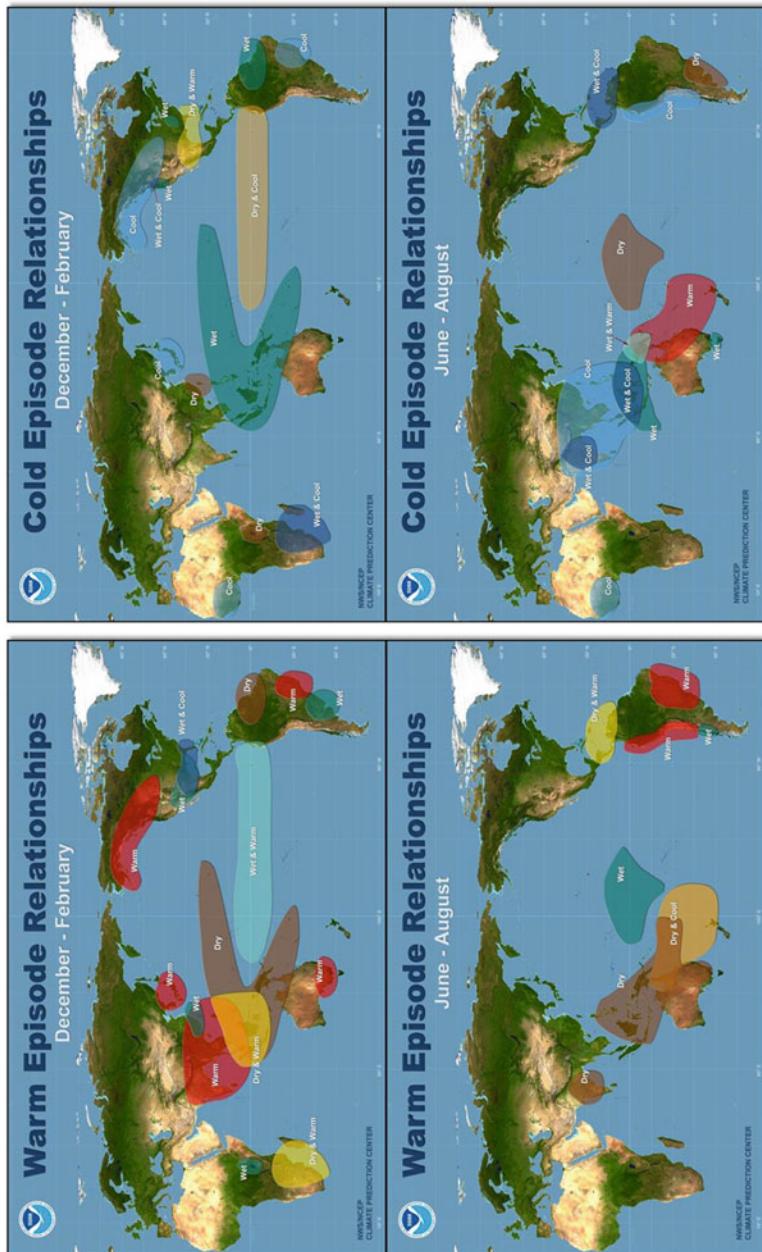


Fig. 2 Warm and cold events of ENSO and its remote connection with precipitation and temperature anomalies in DJF and JJA. Colors show areas with detectable impact

ENSO composites are quoted a lot in press interviews and include the following: “Within the tropics, the eastward shift of thunderstorm activity from Indonesia into the central Pacific during warm episodes results in abnormally dry conditions over northern Australia, Indonesia and the Philippines in both seasons. Drier than normal conditions are also observed over southeastern Africa and northern Brazil, during the northern winter season. During the northern summer season, Indian monsoon rainfall tends to be less than normal, especially in northwest India where crops are adversely affected. Wetter than normal conditions during warm episodes are observed along the west coast of tropical South America, and at subtropical latitudes of North America (Gulf Coast) and South America (southern Brazil to central Argentina).” Impacts in cold events are roughly (not exactly) the mirror image.

Nowadays coupled atmosphere-ocean models can also make decent forecasts of ENSO and its impacts worldwide. Models are especially accurate in reproducing precipitation composites and may become more useful because not all ENSO events are alike (Chen et al. 2017).

2.2 Local Forcing and Feedbacks

Numerous observational studies have indicated the existence of strong local forcing and interactions between the climate components. For example, high-resolution data of low-level winds in the tropical Pacific reveals strong interaction between wind stress and SST in regions of SST gradients. A robust atmospheric response in the form of surface pressure and precipitation to tropical SST anomalies is found in observations (e.g., Rasmusson and Carpenter 1982) and in simple models (Lindzen and Nigam 1987). The common finding is that tropical precipitation anomalies are associated with anomalous low-level moisture convergence, which is generally driven by pressure gradients derived from SST gradients. The atmospheric response to extratropical SST anomalies is difficult to determine because the atmospheric variability is larger than the signal of the forcing from the atmosphere to the ocean. Statistical techniques to extract ocean-atmosphere feedbacks and forcing directions from observations have been developed over the years. Basic lag and lead correlations indicate that in the extratropics, the atmosphere tends to drive the ocean (Peña et al. 2003). To extract the ocean forcing in observation data, some methods are devised to estimate the largest boundary-forced atmospheric signal. The methods are generalizations of the multivariate lag-correlation techniques that allow extracting the maximum response of the atmosphere to the SSTs (Frankignoul et al. 2011). These analyses consistently indicate that SST anomalies do exert influence on the atmosphere in the North Atlantic and North Pacific.

While climate variability studies and models emphasize the ocean-atmosphere interaction as a primary driver, more recent studies and realistic higher resolution modeling highlight the importance of land-atmosphere, ice-atmosphere, and sea-ice interactions at all time scales. Local feedbacks and forcings between the atmosphere and the land, ice, and lithosphere and their impact to weather and climate predictions have been the focus of major international research efforts for years. Understanding

and accurately modeling surface fluxes of heat, moisture, and momentum are critical to the success of climate prediction systems. In particular, the role of soil moisture for precipitation and temperature in the middle latitudes and transitional climate zones is useful sources of predictability for seasonal forecasting (Koster et al. 2000).

2.3 Teleconnection Patterns

Interannual variability in the climate system caused by ENSO and other coupled ocean-atmosphere oscillations occurs primarily in the tropical oceans. Although the effects are concentrated in the tropics, they often cascade into extratropical ocean and land regions. These remote effects, bridged by the atmospheric circulation, are called teleconnections (Wallace and Gutzler 1981). For example, changes in the equatorial easterly winds modify the upwelling and downwelling mechanisms in the upper layers of the ocean and the spatial structure of the SST of the equatorial Pacific; this in turn changes the location of regional deep convection and the large-scale circulation. An exact definition of teleconnection does not exist, but a working definition would be as follows. A teleconnection is a simultaneous significant (statistically significant and of practical importance) temporal correlation in a chosen variable between two locations that are far apart, where “far” means beyond the monopole of positive correlations that is expected to surround each grid point or observational site. “Beyond the local +ve monopole” implies we should first look for significant –ve correlation, keeping in mind there may be significant positive correlation at even greater distance. These teleconnections should exist in the original “raw” data and in that sense be “real.” By far the two most famous teleconnections in the extratropical NH are the North Atlantic Oscillation (NAO) and the Pacific North-American Pattern (PNA). The most important teleconnection with predictive implications is probably the global ENSO teleconnection.

Teleconnection patterns have been deduced via standard statistical methods applied to historical observed variables, gridded, or station data. The more prominent patterns found in historical records are quantified and monitored in real time through so-called indices that measure the strength of the mode. These indices could be just the value of a variable at some well-chosen base point (well chosen because it correlates well with the rest of world), an average over a small area, an expression with base point values entered with alternating signs like Iceland minus Azores for the NAO or Darwin minus Tahiti for the Southern Oscillation, or the coefficient of projection onto an empirical orthogonal function (EOF) or empirical orthogonal teleconnection (EOT). There is a fine distinction between teleconnection patterns and “modes of variability” as calculated by, for instance, EOF. EOFs are important because they explain as much as possible of the variance but in many instances fail the working definition of teleconnection given above. This is especially true for higher order EOFs, but even the first EOF may not satisfy the teleconnection working definition. An important EOF#1 with a name in the NH is the Arctic

Oscillation (AO). The AO is sometimes presented as a longitude invariant version of the NAO which, as the name suggests, is restricted mainly to the Atlantic domain. The AO would be similar to the zonal index cycle, a descriptive theory by Namias (1958) which imagined period's strong westerlies across mid-latitude with low-amplitude planetary waves and fast-moving cyclones, alternating with periods of weaker westerlies, high-amplitude planetary waves, and blocking.

The statistical techniques for teleconnections isolate the preferred large-scale patterns of the atmospheric circulation. When combined with lead and lag methods, they provide a partial description of possible origin and physical mechanisms for the remote correlations. Mechanistic interpretations of teleconnections have been developed to identify sources of the forcing. From this perspective teleconnections take the form of "standing waves" with fixed nodes and antinodes of low-frequency oscillation. Although the theory is incomplete, some of those oscillations may be related to propagating Rossby waves (e.g., Hoskins and Karoly 1981). The explanation of the PNA is often made in terms of Rossby wave dispersion. The explanation of the NAO has yet to be worked out to general satisfaction.

3 Rationale of ISI Predictions

In this section, the rationale and scientific basis to make ISI predictions is discussed. Considering the limit of atmospheric predictability, seasonal forecasts do not attempt to forecast a specific day or locality in which a particular weather event will occur. Instead, they forecast anomalies in monthly or seasonal averages of key climate variables. For example, forecasting "higher than normal" temperature over the Southeastern USA one season ahead may have useful prediction skill at the expense of not detailing the timing and specific location of the anomaly. In seasonal forecasting there are several aspects and levels of refinements at play that the expert takes into account to optimize the information contained in model forecast outputs. First, the "greater than normal" invokes a climate norm and a specific "binning" of the histogram of the event predicted. It is common to make predictions in terms of terciles even when outputs from ensemble forecasts can provide a greater refinement of forecast distribution function. One benefit of reducing the resolution to terciles is that averages of the fraction of ensemble members falling into each of the bins filter out non-predictable features. Figure 3 shows a schematic of the process of binning of an 81-member ensemble into terciles.

The size of the domain and the period used for the time average is also selected by the expert. From a statistical perspective, the process of time and spatial averaging a variable is equivalent to filtering transient, small-scale fluctuations in favor of a slow-evolving large-scale signal. Even in this situation, most variables are difficult to predict with sufficient accuracy, so it is important to understand and exploit the sources of predictability and to find sensible ways to deliver predictions taking into account that forecast errors at ISI time scales are large but contain useful information that users can take advantage of.

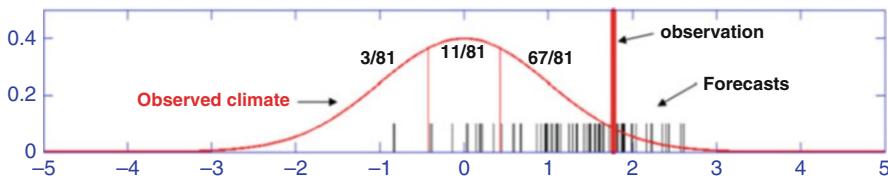


Fig. 3 Schematic of a tercile probabilistic forecast based on an 81-member ensemble. The black bars represent the values of each ensemble member. The ratio shown is the fraction of members falling on each tercile. The red curve is the observed climate computed from past observations

3.1 Sources of Predictability

The ISI variability is driven by a complex combination of many interconnected physical and dynamical processes operating on a wide range of time scales in the atmosphere and the rest of the climate components. Identifying and exploiting the sources of predictability in the spectrum of phenomena are necessary to skillfully predict beyond the limit of atmospheric predictability. Many variables associated with the thermodynamic state of the climate, particularly the upper layers of land, ocean, and ice, control the variability of surface fluxes. These variables evolve much slower than the internal variability of the atmosphere and are said to have large memory or “inertial memory.” Whenever the interaction between the atmosphere and the other components of the climate is strong enough, the inertial memory acts as a long-lasting forcing on the lower layers of the atmosphere, eventually constraining or modifying the statistics of the atmospheric variability. Presuming that atmospheric prediction models are capable of simulating the large-scale atmospheric response to SST forcing, inertial memory can contribute in making the statistics of the atmospheric variables more predictable. Inertial memory of the upper-ocean heat content can extend from seasons out to a year or more (Latif et al. 1998). Moisture captured in the top meter of soil after a precipitation event or snow melting can last from weeks to a few months depending on the type of soil and on the evaporation or transpiration rate at the surface. Soil moisture and snow cover have influence on the atmospheric processes, but it is usually more difficult to quantify due in part to competition with other mechanisms near the surface, feedbacks, and rapid balance adjustments (Dirmeyer 2005; Koster and Mahanama 2012).

A second source of predictability is the remote forcing exerted by teleconnection patterns. This forcing usually persists beyond the typical time scales of atmospheric internal variability, changing the regular atmospheric circulation patterns and flows. Finally, positive feedbacks between climate components reinforce an initial alteration enhancing the life span and/or the amplitude of climate anomalies. Several feedback processes have been identified: (a) the ice-albedo feedback, which increases ice or snow cover when albedo is high (more radiation reflected) and decreases rapidly when albedo is low (more radiation absorbed); (b) the Bjerknes feedback, which explains ENSO oscillation; (c) the wind-evaporation-SST feedback; and (d) the cloud feedback. These are some of the more common mechanisms found in the ocean-atmosphere interactions. Similar to the ice and ocean, the

land-atmosphere interaction contains multiple types of feedbacks. Large extensions of snow-covered regions and water reservoirs may produce some of the above feedbacks. Soil moisture-precipitation feedbacks have been reported in several investigations and are used to estimate and predict near-surface temperatures.

3.2 Observing Research

The capacity to make predictions beyond 2 weeks of the statistics of atmospheric variability is largely associated with the ability to model the life cycle of ENSO, MJO, and other phenomena as a first step and the propagation to other local and remote variables as described above. Observing research has been central to improve the description and understanding of phenomena through special oceanographic and meteorological observations carried out during observational campaigns in specific regions and periods. Those observations help refine model parameters and evaluate forecasts primarily. In addition, the special observations help defining observing priorities and recommendations for operational observing networks. For instance, in the 1980s it was recognized that observations in the tropical ocean-atmosphere interphase such as surface winds and SST were necessary to describe ENSO and associated large-scale phenomena. Through observing campaigns, it was appreciated that the broad spectrum of variability in both the ocean and the atmosphere and that the evolution of the large-scale ENSO phenomenon was conditioned to small-scale fluctuations such as episodic westerly wind bursts. In addition to surface winds and SST, the upper-ocean thermal structure, sea-level, and currents were added as essential. It was also recognized that quantitative understanding of the SST variability required improved estimates of surface heat fluxes and upper-ocean salinity variability and improved resolution of precipitation over the ocean (McPhaden et al. 1998).

Numerous observing experiments and the projects to support them have been organized for over a century to sample strategic regions and a defined set of phenomena. The first large-scale projects were the two International Polar Years in 1882–1883 and 1932–1933 with the goal to increase knowledge of the Arctic regions. The International Geophysical Year (1957–1958) had an emphasis in understanding the Arctic, Antarctic, and tropical regions. The success of these programs and the need for further studies led to the Global Atmospheric Research Program (GARP; 1968). The First GARP Global Experiment (FGGE) took place in 1979. FGGE's main objectives were to improve understanding of the global atmospheric motion for the development of realistic models for weather forecasting and to assess the limit of predictability. The successor of GARP was the 10-year Tropical Ocean-Global Atmosphere (TOGA) program (1985–1995).

TOGA made major advances toward understanding the ENSO, demonstrating the feasibility of operational ISI prediction of equatorial Pacific SST anomalies based on numerical models that simulated the physics of the coupled tropical ocean-atmosphere system. It clarified the nature of the remote, planetary-scale atmospheric response to these anomalies. In the early nineties, two other international projects

started: the World Ocean Circulation Experiment (WOCE) and the Global Energy and Water Cycle Experiment (GEWEX). GEWEX grew under the umbrella of the International Geosphere-Biosphere Program (IGBP). The US Global Ocean-Atmosphere-Land System (GOALS) program carried forward these efforts for the period 1995–2005 with the belief that the skill of operational climate prediction could be further increased by continued research on ENSO and by efforts to understand other elements of the climate system that contribute to the observed ISI variability. The Pan American Climate Studies (PACS) program, a component of the GOALS program, provided a phenomenological context for some of the GOALS research. The program promoted a better understanding and more realistic simulation on (1) boundary forcing of ISI variations, (2) the evolution of tropical SST anomalies, (3) the mean seasonal cycle, the intertropical convergence zone, (ITCZ) and the relevant land surface processes.

The international THORPEX program supported several observing research campaigns including the “Year of Tropical Convection” (YOTC), the International Polar Year (IPY), DIAMET and T-NAWDEX, and the THORPEX Pacific Asian Regional Campaign (T-PARC). T-PARC was one of the most comprehensive observing campaigns in the Pacific Ocean and Asian continent. Its main goal was the understanding of the extratropical transition of tropical cyclones and downstream impacts. A winter T-PARC component aimed at understanding the waveguide propagation of disturbances from Asia to North America, making use of targeted observation techniques.

4 Modeling ISI Variability

Approaches to model ISI variability can be classified as empirical (or statistical) and dynamical. In both, the complexity of the climate system is recognized but is approached in a different manner.

4.1 Statistical Climate Prediction Models

Advanced statistical techniques are used to identify the principal modes of climate variability on different space scales and time scales. These methods are exploited to provide seasonal to interannual predictions. The goal is to find from historical data records an optimal set of precursors (predictors) that predicts best the future evolution of a certain quantity (predictand). Since data records of ENSO and other important signals go back only to the 1950, only a few realizations are captured making it difficult to construct statistical prediction schemes that have both statistical significance, referred to as robustness, and skill. Removing artificial skill during the optimization process of the parameters of the model remains a problem. Success in statistical forecast schemes does not necessarily imply a causal relationship given the complexity and nonlinear interactions and mechanisms that exist in the climate system. Nonetheless, statistical prediction models remain successful for seasonal

prediction when the signal of the relevant predictors (e.g., Nino 3.4 index) is sufficiently strong and linear. While the list is virtually endless, below is a description of just three methods that have been used for about 25 years at least: canonical correlation analysis (CCA), constructed analog (CA), and linear inverse modeling (LIM). More methods are described in Van den Dool (2007), Chap. 8. All these methods use a data set describing past behavior of the system. The CCA and LIM have at their core the lagged cross-covariance matrix, while CA assigns weights to years in the past on the basis analogy and anti-analogy while taking collinearity among years into account.

Most empirical methods in short-term climate prediction are nowadays based on multiple linear regression “on the pattern level.” A primitive example is as follows. Suppose there are two data sets, $f(s,t)$ called the predictor and $g(s,t)$ called the predictand. One can perform two stand-alone EOF analyses of f and g and then do the regression between the time series of the leading modes in the predictand and predictor data sets. Klein and Walsh (1983) made an in-depth comparison of regression between EOF mode time series on the one hand and regression between the original data at grid points on the other – this was in the context of “specification” or downscaling. Using modes is efficient and cuts down on endless choices and thus protects against overfit, but it may not always help the skill.

For a more general approach, the time lagged covariance matrix must be discussed. One can define the elements of the time lagged covariance matrix C_{fg} as n_t

$$c_{ij} = \sum f(s_i, t) g(s_j, t + \Delta t) / n_t \quad (1)$$

where n_t is the number of time levels, a time mean of f and g was removed, and Δt is the time lag. C , non-square in general, thus contains the covariance between the predictor at any place in its domain and the predictand anywhere in its domain – local as well as nonlocal. From the time lag in g , our intention is clear: to predict g from f . However, some analyses (CCA, SVD, or MCA) do not go beyond establishing associations between f and g , leaving in the middle which predicts the other.

The matrix C_{fg} becomes the square C_{ff} or C_{gg} for $f = g$ and zero time lag.

4.1.1 CCA

The point of CCA (also LIM and MCA) is to discover spatial patterns in the predictor data set f and patterns in the predictand data sets g , the time series of which have a strong relationship. In a practical example, $f(s,t)$ could be SST in the tropics, and $g(s,t)$ could be 2-m temperature over the USA. These patterns are calculated from C_{fg} in a manner that resembles the calculation of EOF from say C_{ff} but with fewer orthogonality constraints. We thus expect for $f(s,t)$:

$$f(s,t) = \sum_{m=1}^{M_f} \alpha_m(t) e_m(s) \quad (2)$$

and for the predictand:

$$g(s,t) = \sum_{m=1}^{M_g} \beta_m(t + \Delta t) d_m(s) \quad (3)$$

Coupling the modes among the two data sets f and g , which have the same number of time levels but possibly different spatial domains and grids (also M_f need not equal M_g), can be described in terms of the properties of $\alpha_m(t)$ and $\beta_m(t + \Delta t)$ and $d_m(s)$ and $e_m(s)$, respectively.

The plain distinguishing feature of canonical correlation analysis (CCA) is that the correlation of $\alpha_m(t)$ and $\beta_m(t + \Delta t)$, denoted $\text{cor}(m)$, is maximized – the modes are ordered such that $\text{cor}(m) > \text{cor}(m + 1)$ for all m . Within each data set (homogeneous), we have for CCA.

$$\sum \alpha_k(t) \alpha_m(t) = 0 \quad \text{for } k \neq m \quad (4)$$

$$\sum \beta_k(t + \Delta t) \beta_m(t + \Delta t) = 0 \quad \text{for } k \neq m \quad (5)$$

i.e., orthogonal time series and across the data sets (heterogeneously):

$$\sum \alpha_k(t) \beta_m(t + \Delta t) = 0 \quad \text{for } k \neq m \quad (6)$$

$$\sum \alpha_k(t) \beta_m(t + \Delta t) = \text{cor}(m) \quad \text{for } k = m \quad (7)$$

where summation in the above four equations is over time. The $\text{cor}(m)$ can be found as the square root of the eigenvalues of the square matrix $M = C_{ff} C_{fg} C_{gg} C_{fg}$ (or from $C_{gg} C_{fg} C_{ff} C_{fg}$). Note that CCA's maps are not orthogonal.

CCA was not used much in meteorology until Barnett and Preisendorfer (1987). The main methodological twist in their paper is a prefiltering step where both f and g are truncated to just a few EOFs before calculating C . (Moreover, the EOF associated time series are standardized, as in a version of the Mahalanobis norm.) The pre-filtering greatly reduces CCA's susceptibility to noise. Additionally Barnett and Preisendorfer (1987) applied their adjusted CCA to the seasonal forecast and had the predictor data set cover four antecedent seasons. This method and this particular predictor layout have been popularized by Barnston (1994), and his work found short-term climate prediction application on nearly all continents.

In order to make forecasts for a desired lead time, CCA is calculated for every Δt . In contrast, for LIM the time lagged cross-covariance matrix is evaluated only for one specific Δt , and results are generalized to all imaginable Δt by analytic assumptions.

4.1.2 Constructed Analog

Probably the very first method that produced monthly global SST forecasts is the constructed analog method (van den Dool 1994; Van den Dool and Barnston 1995), CA-SST. These SST forecasts were popular because early atmosphere-ocean coupled models developed very large systematic error. So in the so-called tier

2 approach, a credible SST forecast was held underneath an atmospheric model as a prescribed time-evolving lower boundary condition.

The Idea

Because *natural* analogs are highly unlikely to occur in high degree-of-freedom processes, we may benefit from *constructing an analog* having greater similarity than the best natural analog. As described in Van den Dool (1994), the construction is a linear combination of past observed anomaly patterns such that the combination is as close as desired to the initial state. We then carry forward in time persisting the weights assigned to each historical case. All one needs is a data set of modest affordable length. Assume we have a data set $f(s, j, m)$ of, for instance, monthly mean data as a function of space (s), year ($j = 1, M$), and month (m). Given is an initial condition $f^{IC}(s, j_0, m)$, for example, the most recent state (monthly mean map), where j_0 is outside the range $j = 1, M$. A suitable climatology is removed from the data – henceforth f shall be the anomaly. A (linear) constructed analog is defined as:

$$f^{CA}(s, j_0, m) \equiv \sum_{j=1}^M \alpha_j f(s, j, m) \quad (8)$$

where the coefficients α are to be determined so as to minimize the difference between $f^{IC}(s, j_0, m)$ and $f^{CA}(s, j_0, m)$. Equation (8) is only a diagnostic statement, but since we know the time evolution of the f in year j (we know the next value (s) historically), we can make a forecast (of f or any other variable we believe relates to f) keeping the weights constant in time. Most commonly f would be global SST, and the forecast would be for global SST, but also 500 mb height, or precipitation and temperature over global land can be predicted this way. The CA method employs a non-orthogonal base and borrows the time tendencies from years in the past.

The Method of Finding the Weights

The first concerned is with solving Eq. 8. The problem is that the solution may not be unique, and the straightforward formulation given below leads to an (nearly) ill-posed problem.

Classically we need to minimize U given by:

$$U = \sum_s \left\{ f^{IC}(s, j_0, m) - \sum_{j=1}^M \alpha_j f(s, j, m) \right\}^2 \quad (9)$$

Differentiation w.r.t. the α_j leads to the equation

$$\mathbf{Q} \boldsymbol{\alpha} = \mathbf{a} \quad (10)$$

\mathbf{Q} is one rendition of the covariance matrix with elements $q_{ij} = \sum_m f(s, i, m) f(s, j, m)$ where the summation is over the spatial domain, $\boldsymbol{\alpha}$ is the vector

containing the unknown α_j , and the rhs is the vector \mathbf{a} containing elements a_j given by $a_j = \sum f^{IC}(s, j_0, m) f(s, j, m)$ where the summation is over the spatial domain. Note that α_j is constant in space – we linearly combine whole maps so as to maintain spatial consistency. Even under circumstances where Eq. (10) has an exact solution, the resulting α_j could be meaningless for further application, when the weights are too large and ultrasensitive to a slight change in formulating the problem. Applying ridging to Q can solve that practical problem (for details see Van den Dool (2007), Chap. 7.2). In essence that means solving (10) with constraint that $\sum \alpha_j^2$ is small, like 0.5.

A recent Nino 3.4 prediction by CA-SST is seen in Fig. 4, i.e., from the end of June 2015, ahead of the important El Nino in winter 2015/2016. CA-SST creates ensemble members by applying different EOF truncation (that is like CCA), creating an analog to a single time or several times in a row, and slight variation is the years that participate. Like LIM, CA does the “heavy” calculation just once and then produces the forecast for any lead by persisting the α_j . Examples of global forecasts for the same initial condition can be found here <http://www.cpc.ncep.noaa.gov/products/people/wd51hd/sst/201506/carealtime.html>. One can see the expected

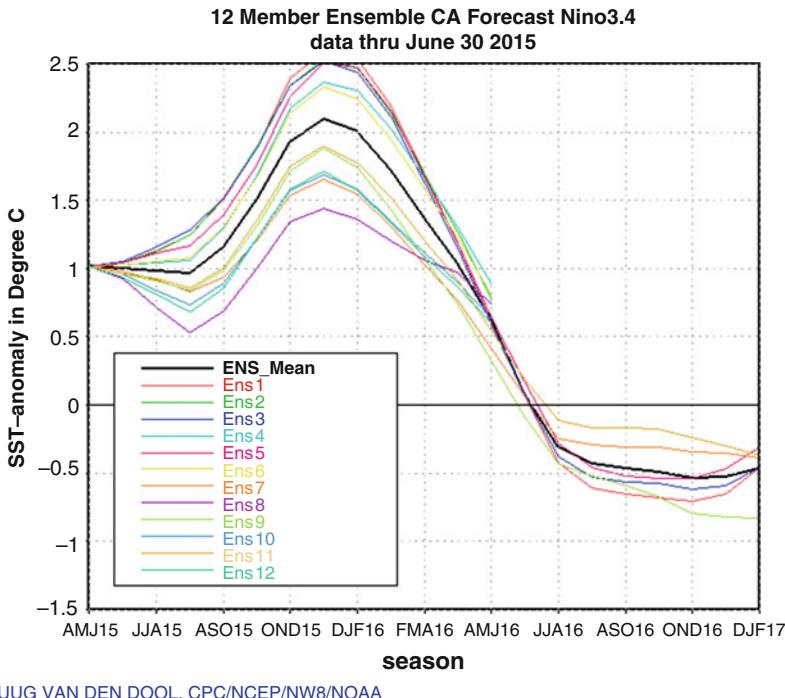


Fig. 4 A recent Nino 3.4 prediction by CA-SST from the end of June 2015, ahead of the important El Nino event in winter 2015/2016. The 12 ensemble members are generated by varying the EOF truncation, the number of antecedent seasons involved in the construction, etc. Ensemble mean forecast is in black

deep Aleutian low for lead = 6 months (JFM) a well-known element (teleconnection) in an El Nino winter.

Constructed analog is also applied to predict soil moisture, another erstwhile lower boundary condition (see Van den Dool et al. (2003)). As a downscaling technique, it has become popular for climate scenarios. Tippett and DelSole (2013) gave a thorough analysis of the relationship of CA with more traditional regression.

4.1.3 Linear Inverse Modeling

The linear inverse modeling (LIM) technique constructs the linear approximation of the climate system, including the linear parameterization of rapidly decorrelating nonlinearities, from the statistics of the system itself (e.g., Penland and Sardeshmukh 1995). Consider an atmospheric state vector x taken as the departure from a climatological norm. The evolution of x can be expressed as:

$$\dot{x} = \mathbf{L}x + \mathbf{N}(x) \quad (11)$$

where \mathbf{L} is a linear operator representing the dynamical equations and $\mathbf{N}(x)$ represents nonlinearities. In some cases, and for a long enough averaging interval, the nonlinear terms may be approximated as a second linear feedbacks of unresolved processes plus some white-noise process ($\mathbf{N}(x)dt \approx \mathbf{T}xdt + dr$) so that the system can modeled as:

$$dx = \mathbf{B}xdt + dr, \quad \mathbf{B} = \mathbf{L} + \mathbf{T}. \quad (12)$$

where \mathbf{B} contains both the linear dynamics and the linear approximation to the nonlinear dynamics, and the covariance matrix of the noise $\langle r r^T \rangle$ can be directly estimated from the observed statistics of x by multiple regression. Notice \mathbf{B} is obtained via the covariance structures in the data itself, which contain both linear and nonlinear dynamics. LIM has been successfully implemented to predict ENSO, tropical SST, and several other variables to describe extratropical atmospheric variability.

4.2 Dynamical Climate Prediction Models

Dynamical models are based on a set of mathematical equations describing the evolution of the climate system. In practice, dynamical models have two main numerical components, a dynamical core, which contains the equations of motion, and a set of physical models to mimic convection, the planetary boundary layer, and many other physical processes that are essential to model climate. Although advanced climate prediction models are often derived from NWP models and are becoming very much alike, there are several aspects that still distinguish them, some of which are highlighted in this section.

The first difference is of course that climate prediction models predict at much longer lead times, typically 7–10 months ahead, compare to the typical 16 days

global weather forecasts. In operational settings, where the time to complete the entire forecast are relatively short, climate prediction models must be configured to reduce the number of calculations compared to weather prediction models. The most common simplification is the reduction of the model's spatial resolution, which necessarily affects the way other model parameters (e.g., time steps, cloud, and turbulence) are customized. In addition, long lead integrations require a special treatment of model parameters to keep errors and instabilities from growing to large and render unphysical results. The present approach of parameterization in operational model is dependent upon model resolution. A second difference is the larger number of model components to represent as realistic as possible the ocean, ice, land, and other climate components. How sophisticated these climate model components are depends on the purpose of the model. For the ocean component in particular, three types of representations are used in current coupled atmosphere-ocean models: the skin surface, the mixed layer, and the full ocean. In the ocean skin surface approach, only surface fluxes and in particular SST fields are inserted or prescribed to the atmospheric model at the bottom boundary. In the NCEP atmospheric global model, a default SST is prescribed, which consists of the daily average of analysis anomaly that decreases at a rate of 90-days e-folding as the forecast lead time progresses. Other atmospheric models only persist the initial SST.

Ocean mixed layer (OML) models constitute the next level of complexity and attempts to model most of the air-sea interaction mechanisms. Conceptually, anomalous atmospheric forcing causes SST anomalies through changes in surface energy fluxes, vertical mixing in the ocean due to turbulence, and wind-driven vertical and horizontal motions associated with Ekman pumping and transport. Generally, ocean mixed layer models neglect temperature advection and focus instead in the vertical structure at every grid point in the ocean.

The success of the coupled modeling scheme is measured on how well it represents special modes of variability such as the annual cycle, the interannual variability, the diurnal cycle, and the major teleconnection patterns in either the observation or the reanalysis data sets. To do this comparison and optimizing model parameters, extensive numerical integrations, referred to as “running” the model, need to be performed. Then, time averages, spectral analysis, and other decompositions of the signal in the model outputs are obtained through statistical analysis and contrasted with observations.

4.2.1 Low-Hierarchy Dynamical Models

The large amount of informational data that need to be summarized into physical concepts and complex fully coupled climate models makes it necessary to create simplified numerical models that represent the most important mechanisms operating in the system or phenomena of interest. It may be that limitations in the knowledge, data, or computing capability restrict the development of more comprehensive models. Large-scale dynamical systems may be too large or have many degrees of freedom that prevent their analysis. Ideally, a large-scale dynamical system could be decomposed into independent and interconnected subsystems or low-hierarchy models.

Most low-hierarchy climate models are diagnostic or equilibrium models that are solved for specific state, for example, the energy balance model, which basically balances incoming (short wave) radiation against terrestrial (mostly long wave) radiation. These models are easy to implement but only provide temperature and associated heat fluxes over large regions and do not resolve well the heterogeneous properties of the surface. Single column (also radiative convective) models are 1-D models that resolve the vertical profile of temperature and associated radiative fluxes.

Another class of low-hierarchy dynamical models corresponds to those that attempt to model the ENSO behavior. Observations of the eastern equatorial Pacific SST and the thermocline depth support the notion that the statistics of El Niño evolution can be modeled using a set of simplified equations of what constitute the recharge oscillator (Burgers et al. 2005), which form as the equation of a classical damped oscillator. These models shed light on possible mechanisms that make ENSO more predictable in the fall than in the spring, for example.

4.2.2 Two-Way Interaction Models

A major source of atmospheric predictability beyond 2 weeks arises from the atmospheric sensitivity to the anomalous lower boundary conditions, particularly the SST (Lau 1985). Because of the ocean's larger thermal inertia, the anomalous SST is commonly assumed (e.g., in the Atmospheric Model Intercomparison Project, AMIP runs) to either strengthen or weaken atmospheric anomalies. The notion that global general circulation models were capable of predicting out to 1 month was pioneered by Shukla (1981) and Charney and Shukla (1981). This one-way (ocean-driving) interaction is still applied in the operational dynamical extended range forecasting (e.g., 15-day daily ensemble forecast at NCEP) and in "two-tier" coupled model forecast system utilized in early coupled models (e.g., Bengtsson et al. 1993), in which predicted SST anomalies are used to force an uncoupled atmosphere model to predict the atmospheric anomalies. This approach improves the seasonal and interannual predictions, primarily due to relatively skillful prediction of development of El Niño. However, the one-way interaction neglects the feedback effect of the atmosphere on the ocean. Moreover, most observational studies indicate that the atmosphere tends to force the ocean over the extratropics at least on intraseasonal time scales.

4.2.3 Coupled General Circulation Models

Fully coupled atmosphere, ocean, land, and cryosphere global circulation models are the most complex models applied to ISI predictions (Latif et al. 1993). Their primary function is to mimic the dynamics of the physical components of the climate system and their interactions. In these models, mechanisms involved in the exchange of heat, momentum, and water at the interface are of primary importance. Although understanding and proper representation of these interactions is not complete, some models produce realistic simulations of the annual cycle and major climate variability modes found in observations. Many research centers are developing and constantly improving their coupled models. These models contain a large number of parameters including expected greenhouse gas and aerosol forcing (Saha et al. 2010). Developing this type of models is complex and requires extensive runs,

diagnostics, and parameter adjustments to simulate past observations and analyses. Once the models satisfy defined criterion for accuracy, they are used for making seasonal predictions on a routine basis.

Despite these successes, basic questions regarding our ability to model the physical processes in the tropical Pacific remain open challenges in the forecast community. For instance, it is unclear how the MJO, westerly wind bursts (WWBs), intraseasonal variability, or atmospheric weather noise influence the predictability of ENSO or how to represent these processes in current models. Typically, coupled models do not adequately capture the diversity of ENSO events such as the recently identified different types of ENSO event (Ashok et al. 2007). There are also apparent decadal variations in ENSO forecast quality (Balmaseda et al. 1995), and the sources of these variations are the subject of some debate. It is unclear whether these variations are just sampling issues or are due to some lower frequency changes in the background state (Kirtman et al. 2005).

Chronic biases in the mean state of climate models and their intrinsic ENSO modes remain, and it is suspected that these biases have a detrimental effect on ENSO forecast quality and the associated teleconnections. Some of these errors are well known throughout the coupled modeling community. Three classic examples, which are likely interdependent, are (1) the double ITCZ problem, (2) the excessively strong equatorial cold tongue typical to most models, and (3) the subtropical eastern Pacific and Atlantic warm biases endemic to all models. Such biases may limit our ability to predict seasonal-to-interannual climate fluctuations and could be indicative of errors in the model formulations. Resolution may be one cause of some of these errors (e.g., Luo et al. 2005). Studies with models that employ higher resolution in both the atmosphere and ocean have demonstrated significant improvements in the mean state of the tropical Pacific and the simulation of El Niño and its teleconnections (e.g., Shaffrey et al. 2009).

5 Climate Prediction Systems

The routine seasonal forecast activity in most meteorological services relies on a numerical framework to process the data from observations to forecasts, products, and their dissemination. The forecast portion of this framework can be separated into three components: observational data processing, data assimilation scheme and model initialization, and numerical forecast method. This section discusses these components, which collectively are known as numerical weather and climate prediction systems. In past generations of prediction systems, the components were connected sequentially. That is, the dataflow was in one direction, but current numerical architecture designs allow interactions among these components. For instance, the forecast model influences the background error covariance of the data assimilation and can influence the data processing (e.g., targeted observations). Reliability and timely forecasts rely on software optimization and supercomputing systems, which are not discussed here (see, e.g., Bauer et al. 2015). As in NWP systems, numerical seasonal prediction systems have advanced in three major areas,

which continue to be challenging: physical process representation, model initialization, and ensemble forecasting. The section restricts the discussion to scientific aspects in prediction systems that are relevant to ISI variability.

5.1 Observational Data

Observations are essential to increase the understanding of climate variability and model improvements. In modern climate prediction systems, observations are used to generate model initial conditions and reanalysis data sets and to validate and improve forecasts. In real time, they are necessary to create the initial conditions of individual climate model components of the coupled prediction model. Non-real-time data are still assimilated into reanalysis systems and are used to optimize model parameters.

Acquiring climate observations of sufficient quality, coverage, and time continuity is a major challenge. It requires collective efforts of the research and the operational climate communities to design, maintain, and evaluate the observing networks and the support of international funding programs to carry on the projects. Observing systems, especially those from satellites and other remote sensing platforms, change sometimes dramatically when the life of the instruments end, making it challenging to provide a continuous climate record of the climate system.

The majority of observations to feed and to validate climate prediction systems come from the meteorological, hydrological, oceanographic, and related observational networks and systems established throughout the world. While climate prediction systems assimilate all the observations available within the data assimilation time window, there are several factors that make certain observations suitable for ISI phenomena. First, the measurements are not limited to the description of the atmosphere but aim to describe the full range of elements in the climate system such as ocean, ice, land use, soil moisture, trace gases, etc. Second, observations must be made for a sufficiently long period of time to sample a large number of events, for example, ENSO events. The record length must also be sufficiently large to create a reference climate. This is an important constraint in surface in situ observations because that require at least about 30 years of continuous, homogeneous, and good quality measurements. Climate observing network maintenance and operation requires careful planning particularly for long-term sustainability.

The ability to predict the seasonal variations of the tropical climate dramatically improved from the early 1980s to the late 1990s thanks in part to an improved observing system. During this period two of the largest El Niño events on record occurred: the 1982–1983 event and the 1997–1998 event. In the case of the former, there was considerable confusion as to what was happening in the tropical Pacific. As a result the NOAA Tropical Atmosphere Ocean (TAO) array of tethered buoys was implemented across the equatorial Pacific, providing essential observations of the ocean’s subsurface behavior. By contrast the development of the 1997–1998 El Niño was monitored very carefully, resulting in considerably better forecast (Table 2).

Table 2 A few major ocean observing network programs

Argo	Argo is a global array of 3,000 free-drifting profiling floats that measures the temperature and salinity of the upper 2000 m of the ocean. This allows, for the first time, continuous monitoring of the temperature, salinity, and velocity of the upper ocean, with all data being relayed and made publicly available within hours after collection.
DBCP	The Data Buoy Cooperation Panel (DBCP) is an official joint body of the World Meteorological Organization (WMO) and the Intergovernmental Oceanographic Commission (IOC), which was formally established in 1985. It consists of the data buoy component of the Joint WMO-IOC Technical Commission for Oceanography and Marine Meteorology (JCOMM). Principal objectives of the DBCP are (i) review and analyze requirements for buoy data, (ii) coordinate and facilitate deployment programs to meet requirements, (iii) initiate and support action groups, (iv) improve quantity and quality of buoy data distributed onto the Global Telecommunication System (GTS), (v) information exchange and technology development, and (vi) liaison with relevant international and national bodies and programs.
OceanSITES	OceanSITES is a worldwide system of long-term, deepwater reference stations measuring dozens of variables and monitoring the full depth of the ocean from air-sea interactions down to 5,000 meters. OceanSITES moorings are an integral part of the Global Ocean Observing System. They complement satellite imagery and ARGO float data by adding the dimensions of time and depth.
TAO	The Global Tropical Moored Buoy Array is a multinational effort to provide data in real time for climate research and forecasting. Major components include the TAO/TRITON array in the Pacific, PIRATA in the Atlantic, and RAMA in the Indian Ocean. The major phenomenological foci of this array are (1) El Niño/Southern Oscillation (ENSO) in the Pacific, (2) the interhemispheric dipole mode, equatorial warm events, and hurricane activity in the Atlantic, and (3) the monsoons, the Indian Ocean Dipole, and intraseasonal variability in the Indian Ocean.
GEOSS	Group on Earth Observations (GEO) is a voluntary partnership of governments and organizations that envisions “a future wherein decisions and actions for the benefit of humankind are informed by coordinated, comprehensive and sustained Earth observations and information.” GEO Member governments include 102 nations and the European Commission, and 95 Participating Organizations comprised of international bodies with a mandate in Earth observations. Together, the GEO community is creating a Global Earth Observation System of Systems (GEOSS) that will link Earth observation resources worldwide across multiple societal benefit areas – biodiversity and ecosystem sustainability, disaster resilience, energy and mineral resources management, food security and sustainable agriculture, infrastructure and transportation management, public health surveillance, sustainable urban development, water resources management – and make those resources available for better informed decision-making.

At that time the global monthly air-surface temperature and precipitation were compiled to validate global models. New assemblies of land-based precipitation and soil moisture started to become available, and traditional oceanographic atlases were supplemented by the compilations of surface variables from the Comprehensive Ocean-Atmosphere Data Set (COADS) project. However, other variables such as

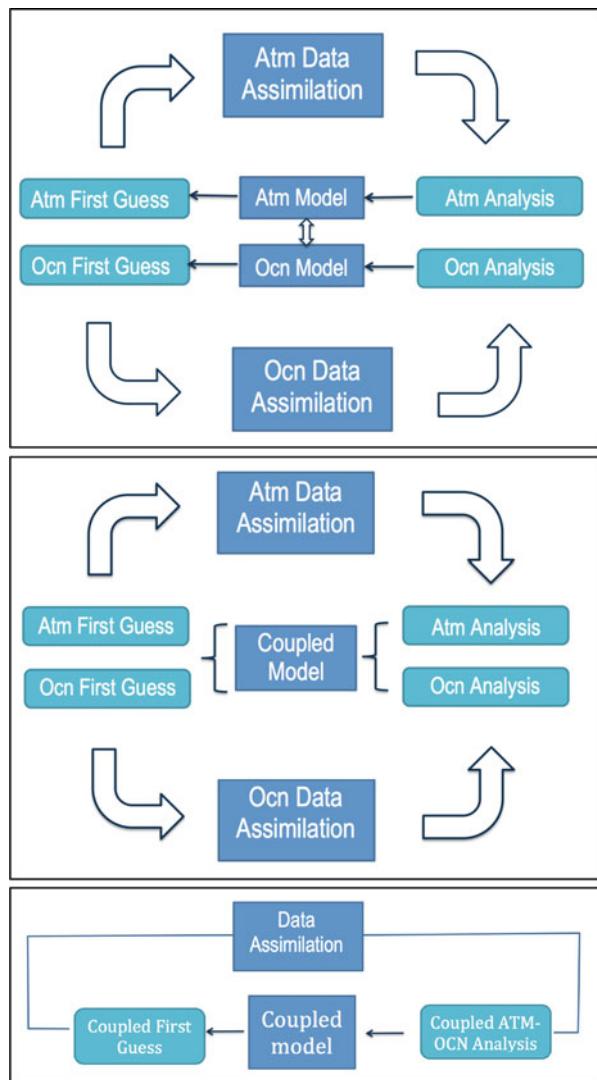
cloudiness, evaporation, runoff, surface heat flux and surface stress, etc. were inadequate for model validation. Reanalysis data sets have been the preferred choice to fill the observational gap.

5.2 Model Initialization and Data Assimilation Schemes

The prediction of ISI climate variability based on coupled models becomes an initial value problem, where initial conditions of each model component must be suitably prepared. As described earlier, climate predictability requires ocean, land, and cryosphere initial conditions and in particular the upper thermal structure. This is a complex undertaking given that not all measurements, such as ocean currents and temperature, arrive on time during the initialization stage. Surface variables and those derived by satellite and other remote sensing techniques are essential to fill important gaps in the observation coverage of ocean, lakes, land, forest, ice, and snow. Ideally, the initial conditions should contain a coherent representation of both the individual components and the phase relationship among climate components. The behavior of variables in some climate components such as land is highly heterogeneous making them difficult to represent with the low spatial resolution of the grids commonly used in the global coupled models. This is exacerbated by lack of sufficient observations to constrain the parameters of the model. The result is that the first guess (the short-range forecast) in those components contains large forecast errors making them difficult to use in data assimilation (DA) schemes unless a suitable bias correction technique is in place (Balmaseda et al. 2007).

Ocean DA schemes are generally adopted for monitoring ENSO because the variation of the heat content in the ocean interior of the equatorial Pacific is considered a good precursor of ENSO and essential for understanding the ENSO mechanisms. Assimilation of observations into an ocean model forced by prescribed atmospheric fluxes is the most common practice for initialization of the ocean component of a coupled model. DA schemes for the ocean, land, and ice model components use the same principles as those for the atmosphere. Simple assimilation schemes such as optimal interpolation techniques and nudging or relaxation to the observation techniques are not uncommon in the ocean, land, and ice model components of operational coupled data assimilation schemes. Figure 5 shows three types of data assimilation schemes in climate prediction systems, the last two of which use a coupled model. One, referred to as weakly coupled DA scheme, consists in having a DA scheme on each climate model component and then creates a coupled analysis by combining the analysis of each component and the background field from the coupled model – see Middle **Diagram**. In this case, the background atmospheric and ocean are in balance, but the analysis increments are calculated separately, and this are potentially unbalanced. The second type is referred to as strongly coupled DA scheme, which consists in having one single DA scheme for all the model components – see Bottom **Diagram**. This implies constructing coupled background error covariance matrix, applying an effective method to remove

Fig. 5 Schematic of different types of DA schemes in seasonal prediction systems. Top, uncoupled; middle, weakly coupled; bottom, strongly coupled. Arrows show the direction of the dataflow



coupled model biases, and imposing a series of physical constraints on coupled analysis increments.

Ensemble coupled data assimilation system has been suggested as a way to initialize coupled models for decades, but its relevance is bringing attention due to more realistic CGCMs. Assessments of weakly coupled (Saha et al. 2010) and strongly coupled (Chang et al. 2013) DA schemes indicate that the initial conditions generated for the coupled models are balanced with the prediction model improving climatological mean and variability of the model and reducing the drift in the deep ocean. However, some errors are not completely eliminated.

5.3 Ensemble Initial Perturbations

The chaotic nature of weather and climate requires its description and prediction to be characterized in terms of probabilistic functions. The ensemble generation community has addressed this issue successfully for the weather problem. Ensemble generation schemes exploit an important characteristic of chaotic dynamical system, which is that initial condition errors grow fastest on particular direction in the model's phase space, reducing the need to run huge ensembles to describe the probability density function of the forecast. A more complex problem arises when dealing with coupled chaotic systems with different time scales of evolution such as the ocean-atmosphere coupled system. Studies based on low-hierarchy models (e.g., Cai et al. 2003; Peña and Kalnay 2004) suggest the use of coupled breeding to separate the different time scales of evolution, which has been applied successfully in comprehensive coupled prediction systems. Other approaches to perturb the initial conditions use empirical singular vectors (Blumenthal 1991), to calculate the fastest growing errors in the coupled system. The most traditional approaches to ensemble generation of climate predictions are the lagged ensembles and the random perturbations of atmospheric variables that force the ocean at the initial time (Vialard et al. 2005). The most important challenge in the ensemble generation strategy is the representation of the model errors. Sampling the model errors have been suggested through the use of stochastic perturbation and the use of multi-model ensembles.

NWP systems aim to best represent synoptic and mesoscale weather events. However, key interactions, for example, at the air-sea-ice interface are assumed negligible. This is problematic on time scales beyond 2 weeks; modeling and predicting seasonal climate anomalies require a realistic treatment of the effects of sea surface temperature, sea ice, snow, soil wetness, vegetation, stratospheric processes, and chemical composition. The lack of such components of the Earth system in NWP systems may well be an impediment to improving forecasts on shorter time scales, particularly for high-impact weather. For example, the ocean mixed layer can precondition the atmosphere-ocean interface for subsequent extratropical and tropical storms. Seasonal prediction systems, on the other hand, typically include such coupled interactions, yet they fail to adequately resolve mesoscale weather systems. There is a wide range of scale interactions to be considered within the context of improving current operational weather and climate prediction systems.

Several model and data assimilation limitations are well known by developers. The more critical are the air-sea, air-land, and air-ice coupling and the model parameterization, particularly convection. Tropical convection exhibits a remarkable variability and organization across space and time scales, ranging from individual cumulus clouds to mesoscale cloud clusters to superclusters (families of mesoscale clusters) to synoptic-scale disturbances and even to planetary-scale circulations. The synoptic disturbances are often associated with equatorially trapped atmospheric waves, which, in turn, organize tropical convection. This hierarchy constitutes a highly nonlinear continuum of scale interaction. It follows that forecast skill in the tropics – on time scales of days, weeks, and beyond – is dependent upon both equatorial waves and convective organization, which contemporary weather and

climate prediction models do not realistically represent. This low skill is usually attributed to inadequacies in parameterizations of moist physical processes. Organized tropical convection is an important part of this deficiency, since it is neither represented by contemporary convective parameterizations nor adequately resolved in global models, especially climate models.

Details of the variability of the atmosphere cannot be predicted deterministically beyond a few days. This limits the ability to predict the details of variables in the other model components of the coupled model such as SST in the ocean or soil moisture in the land. There is, thus, uncertainty in the predicted forcing from the atmosphere to the other components, which is increased during the coupling process because of the uncertainty in the details of air-sea and air-land interaction processes. The ocean dynamics alone to a lesser extent compared to the atmosphere have instabilities that increase the growth of small errors in the initial conditions. Several approaches have been created to represent the uncertainty in the initial conditions of the coupled models. One approach that continues to be in use is the lagged ensemble generation strategy, which consists in collecting forecasts initialized at earlier times (Kalnay 2002).

6 Product Generation Tools

Operational product generation schemes vary from center to center. Next is a concise description of each of them and how they are utilized for product generation.

6.1 Reanalysis and Hindcast Data Sets

Reanalysis data sets are scientific tools for developing a comprehensive record of how climate states change over time. In them, observations and a numerical model that simulates one or more aspects of the Earth system are combined objectively to generate a synthesized estimate of the state of the system. A reanalysis typically extends over several decades or longer and covers the entire globe from the Earth's surface to well above the stratosphere, depending on the model's vertical resolution. The generation of a reanalysis needs a data assimilation scheme, a frozen forecast model, and all the archived past observations available that included those data sets used in real-time predictions and those that were gathered later on. Both the DA scheme and the forecast model are integrated in time spanning multi-decadal periods producing daily or 6-hourly analyses. The main advantage of using reanalysis as opposed to using historical real-time analysis outputs is that it removes long-term trends and discontinuity caused by model or DA scheme upgrades throughout the years. Major observing network changes produce undesirable discontinuities, which are difficult to remove with the current methods available. Reanalyses are also commonly used to initialize coupled prediction models because they contain a consistent ocean-atmosphere analysis. Reanalysis products are used extensively in climate research and services, including for monitoring and comparing current climate conditions with those of the past, identifying the causes of climate variations

and change, and preparing climate predictions. Information derived from reanalyses is also being used increasingly in commercial and business applications in sectors such as energy, agriculture, water resources, and insurance.

A hindcast is a data set of historical seasonal forecasts generated with a fixed forecast model that typically extend to several decades (e.g., Saha et al. 2014). The fixed forecast model is the same used to provide the real-time seasonal predictions. The main purpose is to generate a model climatology for each initial day of the year and lead time. The model climatology has several purposes. The most used is to remove systematic errors of the first and second moments of the distribution of the variable of interest. The model climatology serves as a reference to create index of severity of predicted anomalies. Such strategy is used to compute the extreme forecast index (EFI) for seasonal anomalies. Another application is the preparation of the a priori skill mask, which is computed as two-dimensional masks to cover regions where the particular forecast model performed lower than an acceptable threshold, and highlights the area where the model has performed well in the past. This mask is useful for the forecasters to be able to sense the level of trust put into the anomalies predicted by the model.

6.2 Multi-Model Ensembles

Collaborative multiagency activities have resulted in the multi-model approach, which consists in gathering in one place all the ensemble forecasts from the participating institutions. The goals of these efforts are to capitalize on the modeling investments made independently by each institution to produce routine real-time multi-model ensemble ISI predictions. The data concentrated follows a prespecified data format and schedule of reception of data and algorithms to automatically generate forecast products. Calibration, evaluation, and consolidation of multi-model ensemble prediction system are done in an experimental basis. A key to the success of prediction is the use of hindcast data sets, which consist in retrospective ensemble forecasts that are consistent with the real-time forecast on each model. Data provider institutions benefit from this collaboration in a number of ways, including a near real-time and ongoing evaluation of their prediction system by numerous users with a wide range of potential applications and peer-to-peer exchange of modeling and initialization strategies, inter-comparisons of model, etc.

Multi-model ensemble forecasts have been found to have higher skill than forecasts made by individual models (Kirtman et al. 2014). The NMME project organized by NOAA/CPC is aimed at improving subseasonal to seasonal (S2S) forecast skill by blending predictions from different models. Figure 6 shows one product of the NMME project.

In general, a multi-model ensemble (MME) prediction system approach provides more useful probability density functions than those obtained from a single one when their skills are similar and forecasts are nonredundant (e.g., Peña and van den Dool 2008). Moreover, the MME approach identifies which outcomes are model independent and hence likely to be robust. Weather and climate ensembles produce

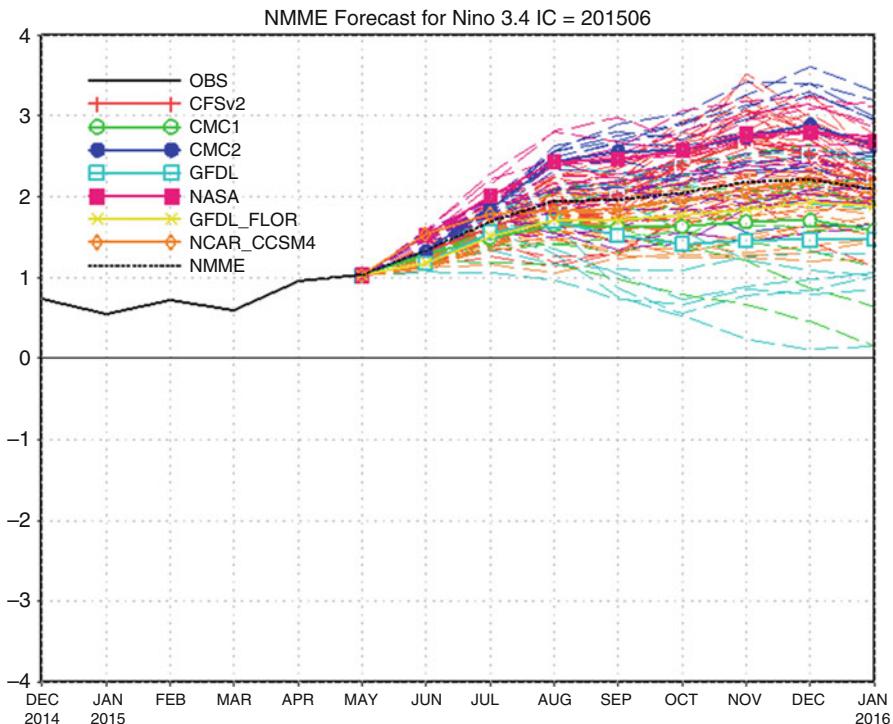


Fig. 6 ENSO plume forecasts from the NMME project corresponding to forecasts initialized about June 1, 2016. Colors correspond to the ensemble model indicated. Black line corresponds to the Nino 3.4 index observation; dotted black line is the multi-model average. (Source: <http://www.cpc.ncep.noaa.gov/products/NMME>)

biased outputs both in its ensemble mean and in its spread. It is a nontrivial task to characterize statistically these deficiencies and to utilize the information to process model output for forecast applications, which have a wide range of time horizons. The skill and uncertainty of weather and climate forecasts are highly space and time-scale dependent, making traditional post-processing approaches only partially useful. Accounting for this dependency is critical for many ensemble applications that are sensitive to the space-time variability of weather and climate. Evaluating biases and forecasting skill on the subseasonal time scale requires extensive hindcast data sets. To this end, several research institutions and universities have joined efforts to produce and centralize forecast data sets (TIGGE for weather, NMME and IMME for the seasonal) in uniform formatting to facilitate its assessment and use (Kirtman et al. 2014).

6.3 Consolidation Methods

Forecasts arising from a combination of multiple models of similar skill generally outperform forecasts from individual models. This is true both for forecasts

produced by the same model but with perturbed initial states and for forecasts produced by models that differ in numerics or physics or both, which may be run at different institutions. Efforts to make the best single forecast out of a number of forecast inputs, a consolidation forecast, have resulted in different types of combination approaches. The best forecast minimizes the average of an error metric (e.g., the root-mean-square error) over a series of past events. Simple (equal weights) multi-model averaging of ensemble members (MMA) can produce forecasts consistently more accurate and more probabilistically reliable than forecasts from any single participating model (Hagedorn 2005). Other more sophisticated consolidation methods have been developed to further improve forecast skill; however, whether these methods can outperform MMA is still a matter of debate, the main obstacle being a lack of sufficiently long data sets of retrospective forecasts. Independently on whether these methods can improve upon MMA, given that more and more prediction systems are becoming available to forecasters and other users, an objective procedure is necessary to deal with the information overload (Chen and van den Dool 2017). For this, optimal weights should be determined for all input models taking into consideration their individual past performance and collinearity among models. Judging the success of a consolidation is difficult because so many diverse issues play a role: hindcasts, overfit, and collinearity. Figure 7 shows the performance comparison of different consolidation strategies as described in Peña and Van den Dool (2008) for monthly prediction of SST in the tropical Pacific.

6.4 Decision Support Systems

Information and decision support systems (DSS) link existing databases, monitoring of vegetation and climate (with ground observations and remote sensing), weather and seasonal climate forecasts, and simulation tools. DSS have been developed in many countries to improve climate risk management in the agricultural sector. The initial stages are establishing and developing products and tools to assist in the preparedness and response to extreme anomalies. Formulation and implementation of activities are suited to the needs of the users and to the computer and observing capabilities available. A series of DSS projects have sprawled in other sectors or to address specific user's needs some include:

6.4.1 Drought Monitoring and Prediction

Drought is the most costly natural hazard that can lead to widespread impacts, including agricultural losses, water and food crises, and wildfires. Because of its slow-evolving nature and greater areal extent, many drought monitoring and prediction systems (DMAPS) have been developed using climate information and forecasts. Depending on the area of interest and desired temporal and spatial resolutions, DMAPS vary from simple meteorological drought prediction system (utilizing only precipitation forecasts) covering the globe to complex drought information system that integrates satellite data and multi-model ensemble forecasts to provide regional information of drought indicators from meteorological, hydrological, and agricultural

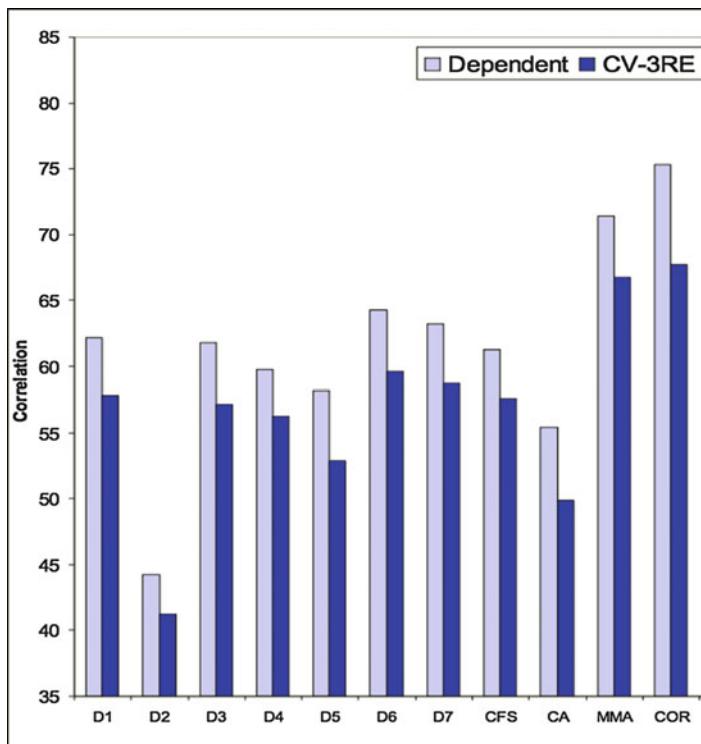


Fig. 7 Ensemble mean anomaly correlation skill performance of monthly tropical Pacific SST based on nine seasonal prediction models. D1 through D7 are DEMETER model participants. The CFS and CA are the 2008 operational NCEP models. Light color bars correspond to AC using all the data available and dark color when a cross validation method is applied to estimate the true skill. The MMA bars correspond to equal weight average and COR to weighted skill average. (Modified from Peña and Van den Dool (2008))

aspects. A well-known and successful DMAPS example is the US Drought Monitor (USDM) and Drought Outlook (USDO) available at <http://www.drought.gov>.

Since 1999, NOAA National Centers for Environmental Prediction (NCEP) has been producing operational drought monitoring and forecast products – USDM and USDO – to aid preparedness planning and mitigation efforts. To support this mission, NCEP, together with its partners from academia and federal agencies, has developed the North American Land Data Assimilation System (NLDAS) and NMME Standardized Precipitation Index (SPI) prediction system, to monitor drought in real time and predict drought from a month to a season in advance.

NLDAS consists of four uncoupled land surface models (i.e., Noah, Mosaic, VIC, and SAC-SMA) driven by observation-based atmospheric forcing, including precipitation, radiation, and low-level winds, to generate land surface states and fluxes at a horizontal resolution of 1/8 degree (Mitchell et al. 2004). For Phase-2 NLDAS

(Xia et al. 2012a, b), the precipitation forcing is derived from CPC unified gauge-based precipitation analysis (Chen et al. 2008), and the majority of atmospheric forcing (e.g., 2-m air temperature, 2-m specific humidity, and 10-m wind speed) is obtained from the North American Regional Reanalysis (NARR; Mesinger et al. 2006). Output variables from NLDAS, such as soil moisture, runoff, evaporation, and snow water equivalent, are used to calculate drought indices (e.g., SPI, SMP, and SRI) to provide comprehensive information on drought conditions from multiple aspects (Mo 2008). Anomaly fields of selected variables are updated daily and available at CPC drought information website (<http://www.cpc.ncep.noaa.gov/prod/ucts/Drought>) to assist in real-time drought monitoring.

The NMME SPI prediction system (Chen et al. 2013) consists of six model forecasts from USA and Canada modeling centers, including the CFSv2, FLOR, GEOS-5, CCSM4, CanCM3, and CanCM4 models. Before calculating SPI, monthly-mean precipitation forecasts from each model were bias corrected and spatially downscaled to regional grids of 0.5-degree resolution over the contiguous USA based on the probability distribution functions derived from the hindcasts (Wood et al. 2002). The corrected precipitation forecasts were then appended to the CPC unified precipitation analysis to form a precipitation time series for computing 1-, 3-, 6-, and 12-month SPIs. The ensemble SPI forecasts are the equally weighted mean of the six model forecasts. The NMME SPI prediction system has been in operation since December 2012. New forecasts (with lead time up to 6 months) are issued every month according to the NMME forecast schedule and are available at http://www.cpc.ncep.noaa.gov/products/Drought/Monitoring/spi_outlooks.shtml. Figure 8 provides a snapshot of the 3-month SPI forecast with initial condition on 1–5 April 2013 when 2011–2016 California droughts intensified. NMME SPI predictive skill is regionally and seasonally dependent, and the 6-month SPI forecasts are skillful (with anomaly correlation above 50%) out to 4 months.

6.4.2 Streamflow Forecasting Framework Using Climate and Hydrological Models

Water resources planning and management efficacy are subject to capturing inherent uncertainty stemming from climatic and hydrological data and models. Streamflow forecasts, critical in reservoir operation and water allocation decision-making, contain uncertainty due to initial conditions, model structure, and model processes. Accounting for these propagating uncertainties is still a challenge. Enhancement in climate forecasting skill and the availability of multiple ensembles, each with equally likelihood to occur provides impetus to deliver probabilistic streamflow forecasts with a mean that is on average more accurate than any particular ensemble member; it also provides a range of plausible solutions including those that give rise to extreme scenarios. Integration of multiple climate and multiple hydrological models increases the pool of streamflow forecast ensemble members. Frameworks have been developed to make this integration of models which have been used in river basin for demonstration purposes.

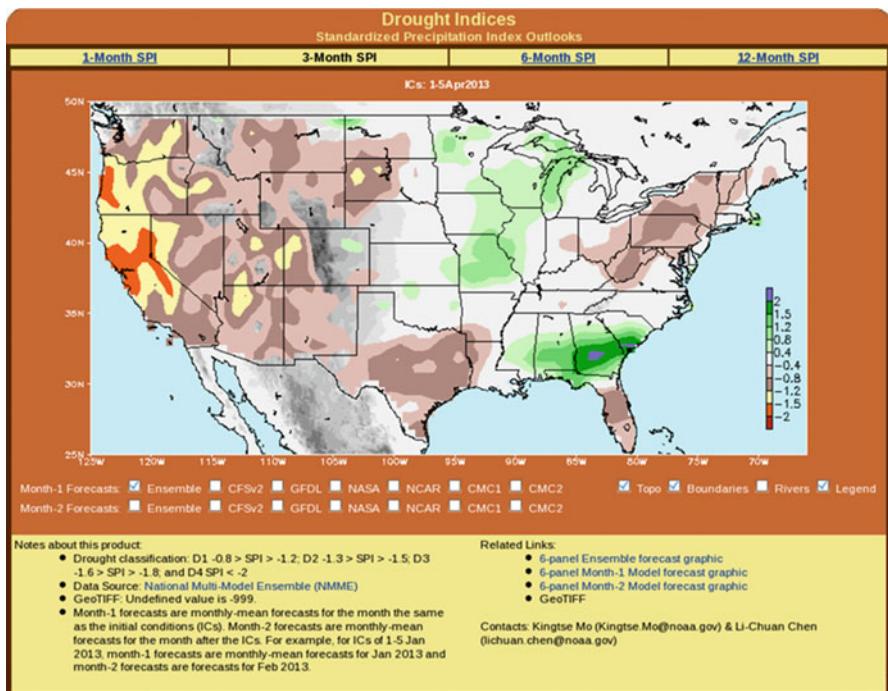


Fig. 8 Snapshot of the 3-month SPI forecast with initial condition on 1–5 April 2013 over the contiguous US

An established example is the NCEP/Princeton University hydroclimate forecast system (Luo and Wood 2008) that uses climate forecasts from multiple models merged with observed climatology in a Bayesian framework to drive a macroscale hydrologic model (VIC) with initial conditions derived from NLDAS. Simultaneously, climate forecasts are downscaled to an appropriate spatial scale for hydrologic predictions. When generating daily meteorological forcing, the system uses the rank structures of selected historical forcing records to ensure reasonable weather patterns in space and time. Seasonal forecasts from this system show promising skill in soil moisture and streamflow, comparing to forecasts produced with the traditional Ensemble Streamflow Prediction (ESP) approach used in operational seasonal streamflow predictions. Based on this framework, several systems have been developed to provide real-time forecasts over the contiguous USA, such as the Princeton Seasonal Hydrological Forecast System at <http://hydrology.princeton.edu/forecast/current.php>, NLDAS Drought Forecast Analysis at <http://www.emc.ncep.noaa.gov/mmb/nldas/forecast/BASE/perc>, Michigan State University Hydrological Monitoring and Seasonal Forecasting at <http://drought.geo.msu.edu/research/forecast/drought.php>.

7 Summary

This chapter is a short account of short-term climate variability and prediction. Predicting future states of climate from intraseasonal to interannual has grown out of necessity to help make decisions on a broad range of human activities including agriculture, energy, and hydrology. The most common approach to make predictions of some aspects of the climate say next month's near-surface temperature over North America now relies on numerical guidances. How these guidances are generated, applied, and interpreted has improved throughout the years thanks to scientific and technologic advancements. In this chapter, some of the elements to create the numerical guidances were presented. The description attempts to follow an easy path to the topics starting with an overview of the science and rationale behind ISI predictions, particularly coupled model predictions. The main message is that sources of predictability abound but must be identified through the large data set of observations and the interrelation of variables. The more comprehensive global observations and reanalysis archives have helped identify preferable modes of ISI variability. The identification of the sources and phenomena in the data alone does not warrant successful predictions, and the quest is how to model that variability.

References

- M.A. Balmaseda, M.K. Davey, D.L.T. Anderson, Decadal and seasonal dependence of ENSO prediction skill. *J. Clim.* **8**, 2705–2715 (1995). [https://doi.org/10.1175/1520-0442\(1995\)008<2705:DASDOE>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<2705:DASDOE>2.0.CO;2)
- M.A. Balmaseda, D. Dee, A. Vidard, D.L.T. Anderson, A multivariate treatment of bias for sequential data assimilation: Application to the tropical oceans. *Q. J. R. Meteorol. Soc.* **133**, 167–179 (2007)
- T.P. Barnett, R. Preisendorfer, Origins and levels of monthly and seasonal forecast skill for United States surface air temperatures determined by canonical correlation analysis. *Mon. Weather Rev.* **115**, 1825–1850 (1987). [https://doi.org/10.1175/1520-0493\(1987\)115<1825:OALOMA>2.0.CO;2](https://doi.org/10.1175/1520-0493(1987)115<1825:OALOMA>2.0.CO;2)
- A.G. Barnston, Linear statistical short-term climate predictive skill in the Northern Hemisphere. *J. Clim.* **7**, 1513–1564 (1994). [https://doi.org/10.1175/1520-0442\(1994\)007<1513:LSSTCP>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<1513:LSSTCP>2.0.CO;2)
- P. Bauer et al., The quiet revolution of numerical weather prediction. *Nature* **525**, 47–55 (2015)
- L. Bengtsson, U. Schleese, E. Roeckner, A two tiered approach to climate forecasting. *Science* **261**, 1026–1029 (1993)
- M.B. Blumenthal, Predictability of a coupled ocean-atmosphere model. *J. Clim.* **4**, 766–784 (1991)
- G. Burgers et al., The simplest ENSO recharge oscillator. *Geophys. Res. Lett.* **32**, L13706–L13709 (2005)
- J. Blunden et al., State of the Climate in 2010. *Bull. Amer. Soc.* **92**, S1–S236, (2011)
- M. Cai, E. Kalnay, Z. Toth, Bred vectors of the Zebiak–Cane model and their application to ENSO predictions. *J. Clim.* **16**, 40–55 (2003)
- Y.-S. Chang, S. Zhang, A. Rosati, T.L. Delworth, W.F. Stern, An assessment of oceanic variability for 1960–2010 from the GFDL ensemble coupled data assimilation. *Clim. Dyn.* **40**(3–4), 775 (2013)

- M. Chen, W. Shi, P. Xie, V.B.S. Silva, V.E. Kousky, R.W. Higgins, J.E. Janowiak, Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res.* **113**, D04110 (2008). <https://doi.org/10.1029/2007JD009132>
- L.-C. Chen, K.C. Mo, Q. Zhang, J. Huang, Meteorological drought prediction using a multi-model ensemble approach, in *38th NOAA Climate Diagnostics & Prediction Workshop Special Issue*, Climate Prediction S&T Digest, 2013, pp. 48–50
- L.G. Chen, H. van den Dool, Combination of Multimodel Probabilistic Forecasts Using an Optimal Weighting System. *Wea. Forecasting*, **32**, 1967–1987 (2017). <https://doi.org/10.1175/WAF-D-17-0074.1>
- L. Chen, H. van den Dool, E. Becker, Q. Zhang, ENSO Precipitation and Temperature Forecasts in the North American Multimodel Ensemble: Composite Analysis and Validation. *J. Climate*, **30**, 1103–1125 (2017). <https://doi.org/10.1175/JCLI-D-15-0903.1>
- P.A. Dirmeyer, The land surface contribution to the potential predictability of boreal summer season climate. *J. Hydrometeorol.* **6**, 618–632 (2005)
- C. Frankignoul, N. Chouaib, Z. Liu, Estimating the observed atmospheric response to SST anomalies: Maximum covariance analysis, generalized equilibrium feedback assessment, and maximum response estimation. *J. Clim.* **24**, 2523–2539 (2011). <https://doi.org/10.1175/2010JCLI3696.1>
- R. Hagedorn, The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concepts. *Tellus* **57A**, 219–233 (2005)
- B.J. Hoskins, D.J. Karoly, The steady linear response of a spherical atmosphere to thermal and orographic forcing. *J. Atmos. Sci.* **38**, 1179–1196 (1981). [https://doi.org/10.1175/1520-0469\(1981\)038<1179:TSLROA>2.0.CO;2](https://doi.org/10.1175/1520-0469(1981)038<1179:TSLROA>2.0.CO;2)
- B.P. Kirtman et al., The north American multimodel ensemble: Phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**, 585–601 (2014). <https://doi.org/10.1175/BAMS-D-12-00050.1>
- B.P. Kirtman, K. Pegion, S. Kinter, Internal atmospheric dynamics and climate variability. *J. Atmos. Sci.* **62**, 2220–2233 (2005)
- W.H. Klein, J.E. Walsh, A comparison of pointwise screening and empirical orthogonal functions in specifying monthly surface temperatures from 700 mb data. *Mon. Weather Rev.* **111**, 669–673 (1983)
- R.D. Koster, M.J. Suarez, M. Heiser, Variance and Predictability of Precipitation at Seasonal-to-Interannual Timescales. *J. Hydrometeorol.* **1**, 26–46, (2000)
- R.D. Koster, S.P. Mahanama, Land surface controls on hydroclimatic means and variability. *J. Hydrometeorol.* **13**(5), 1604–1620 (2012)
- M. Latif et al., Climate variability in a coupled GCSM, I, the tropical Pacific. *J. Clim.* **6**, 5–21 (1993)
- M. Latif, D. Anderson, T. Barnett, M. Cane, R. Kleeman, A. Leetmaa, J. O'Brien, A. Rosati, E. Schneider, A review of the predictability and prediction of ENSO. *J. Geophys. Res.* **103**(C7), 14375–14393 (1998). <https://doi.org/10.1029/97JC03413>
- N.C. Lau, Modeling the seasonal dependence of the atmospheric response to observed El Niños 1962–1976. *Mon. Weather Rev.* **113**, 1970–1996 (1985)
- R.S. Lindzen, S. Nigam, On the role of the sea surface temperature gradients in forcing low-level winds and convergence in the tropics. *J. Atmos. Sci.* **44**, 2440–2458 (1987). [https://doi.org/10.1175/1520-0469\(1987\)044<2440:OTROSS>2.0.CO;2](https://doi.org/10.1175/1520-0469(1987)044<2440:OTROSS>2.0.CO;2)
- E. Lorenz, *Empirical Orthogonal Functions and Statistical Weather Prediction*. Sci. Rep. No. 1, Statistical Forecasting Project, M.I.T., Cambridge, MA, 1956, 48 pp.
- L. Luo, E.F. Wood, Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the Eastern United States. *J. Hydrometeorol.* **9**, 866–884 (2008). <https://doi.org/10.1175/2008JHM980.1>
- J.J. Luo, S. Masson, E. Roeckner, G. Madec, T. Yamagata, Reducing climatology bias in an ocean-atmosphere CGCM with improved coupling physics. *J. Clim.* **18**, 2344–2360 (2005)
- R. Madden, P. Julian, Detection of a 40–50 day oscillation in the zonal wind in the tropical Pacific. *J. Atmos. Sci.* **28**, 702–708 (1971)

- S. Manabe, K. Bryan, Climate calculations with a combined ocean-atmosphere model. *J. Atmos. Sci.* **26**, 786–789 (1969)
- M.J. McPhaden et al., The tropical ocean-global atmosphere observing system: A decade of progress. *J. Geophys. Res.* **103**(C7), 14169–14240 (1998). <https://doi.org/10.1029/97JC02906>
- F. Mesinger, G. DiMego, E. Kalnay, K. Mitchell, P.C. Shafran, W. Ebisuzaki, D. Jović, J. Woollen, E. Rogers, E.H. Berbery, M.B. Ek, Y. Fan, R. Grumbine, W. Higgins, H. Li, Y. Lin, G. Manikin, D. Parrish, W. Shi, North American regional reanalysis. *Bull. Am. Meteorol. Soc.* **87**, 343–360 (2006)
- K.E. Mitchell et al., The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.* **109**, D07S90 (2004). <https://doi.org/10.1029/2003JD003823>
- K.C. Mo, Model-based drought indices over the United States. *J. Hydrometeorol.* **9**, 1212–1230 (2008)
- J. Namias, Synoptic and climatological problems associated with the general circulation of the Arctic. *Trans. Am. Geophys. Union* **39**(1), 45 (1958)
- M. Peña, E. Kalnay, Separating fast and slow modes in coupled chaotic system. *Nonlinear Process. Geophys.* **11**, 319–327 (2004)
- M. Peña, H. van den Dool, Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Clim.* **21**, 6521–6538 (2008). <https://doi.org/10.1175/2008JCLI2226.1>
- M. Peña, E. Kalnay, M. Cai, Statistics of locally coupled ocean and atmosphere intraseasonal anomalies in reanalysis and AMIP data. *Nonlinear Process. Geophys.* **10**, 245–251 (2003). <https://doi.org/10.5194/npg-10-245-2003>
- C. Penland, P.D. Sardeshmukh, The optimal growth of tropical sea surface temperature anomalies. *J. Clim.* **8**, 1999–2024 (1995). [https://doi.org/10.1175/1520-0442\(1995\)008<1999:TOGOTS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1995)008<1999:TOGOTS>2.0.CO;2)
- E.M. Rasmusson, T.H. Carpenter, Variations in tropical sea surface temperature and surface wind fields associated with the southern oscillation/El Niño. *Mon. Weather Rev.* **110**, 354–384 (1982)
- S. Saha et al., The NCEP climate forecast system reanalysis. *Bull. Am. Meteorol. Soc.* **91**, 1015–1057 (2010). <https://doi.org/10.1175/2010BAMS3001.1>
- S. Saha et al., The NCEP climate forecast system version 2. *J. Clim.* **27**, 2185–2208 (2014)
- L.C. Shaffrey et al., UK HiGEM: The new UK high-resolution global environment model – Model description and basic evaluation. *J. Clim.* **22**, 1861–1896 (2009). <https://doi.org/10.1175/2008JCLI2508.1>
- J. Shukla, Dynamical predictability of monthly means. *J. Atmos. Sci.* **38**, 2547–2572 (1981). [https://doi.org/10.1175/1520-0469\(1981\)038<2547:DPOMM>2.0.CO;2](https://doi.org/10.1175/1520-0469(1981)038<2547:DPOMM>2.0.CO;2)
- M. Tippett, T. DelSole, Constructed analogs and linear regression. *Mon. Weather Rev.* **141**, 2519–2525 (2013)
- H. Van den Dool, Searching for analogs, how long must we wait? *Tellus* **46A**, 314–324 (1994)
- H. Van den Dool, *Empirical Methods in Short-Term Climate Prediction* (Oxford University, Oxford, 2007)
- H.M. Van den Dool, A.G. Barnston, 1995: Forecasts of global sea surface temperature out to a year using the constructed analogue method, in *Proceedings of the 19th Annual Climate Diagnostics Workshop*, 14–18 Nov 1994, College Park, pp. 416–419
- H. Van den Dool, J. Huang, Y. Fan, Performance and analysis of the constructed analogue method applied to U.S. soil moisture over 1981–2001. *J. Geophys. Res.* **108**(D16), 8617 (2003)
- J. Vialard et al., An ensemble generation method for seasonal forecasting with an ocean–atmosphere coupled model. *Mon. Weather Rev.* **133**, 441–453 (2005)
- J.M. Wallace, D.S. Gutzler, Teleconnections in the geopotential height field during the Northern Hemisphere winter. *Mon. Weather Rev.* **109**, 784–812 (1981). [https://doi.org/10.1175/1520-0493\(1981\)109<0784:TITGHF>2.0.CO;2](https://doi.org/10.1175/1520-0493(1981)109<0784:TITGHF>2.0.CO;2)

-
- A.W. Wood, E.P. Maurer, A. Kumar, D. Lettenmaier, Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.* **107**(D20), 4429 (2002). <https://doi.org/10.1029/2001JD000659>
- Y. Xia et al., Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.* **117**, D03109 (2012a). <https://doi.org/10.1029/2011JD016048>
- Y. Xia et al., Continental-scale water and energy flux analysis and validation for North American Land Data Assimilation System project phase 2 (NLDAS-2): 2. Validation of model-simulated streamflow. *J. Geophys. Res.* **117**, D03110 (2012b). <https://doi.org/10.1029/2011JD016051>

Part III

Post-processing of Meteorological Ensemble Forecasting for Hydrological Applications



Hydrological Challenges in Meteorological Post-processing

Fredrik Wetterhall and Paul Smith

Contents

1	Introduction	240
2	What the Hydrologist Needs from Post-processing	242
2.1	Consistency with the Calibration Fields	243
2.2	Temporal Relevance	244
2.3	Spatial Relevance	245
2.4	Event Magnitude	246
2.5	Consistency Between Variables	246
3	How Meteorological Post-processing Affects Hydrological Prediction	247
3.1	Spatial and Temporal Interpolation	247
3.2	The Analogue Approach and Poor Man's Ensemble	248
3.3	Model Output Statistics	248
3.4	Bayesian Post-processing Methods	249
4	Future Challenges in Post-processing	249
4.1	Extreme Values	249
4.2	Combining with Hydrological Post-processing	250
4.3	Toward a Seamless Forecasting System	250
5	Summary	251
	References	251

Abstract

Uncertainties in the hydrometeorological forecasting chain derive from a large number of sources and are inherent to any system. One source of uncertainty is the discrepancy between the meteorological forecasts and the weather which subsequently occurs. Post-processing meteorological forecasts can reduce this discrepancy by removing systematic errors and produce more reliable, corrected forecasts. However, when the corrected NWP output is used in hydrological

F. Wetterhall (✉) · P. Smith

Forecast Department, European Centre for Medium-Range Weather Forecasts, Reading, UK
e-mail: fredrik.wetterhall@ecmwf.int

applications, problems may occur where consistency and correlation between meteorological variables have not been maintained. Therefore a correction that improves the forecast performance of one or more NWP outputs does not necessarily have a positive influence on the hydrological model forecasts. In this chapter the most important needs of the hydrological community in terms of meteorological post-processing are presented. The most commonly used techniques for post-processing are presented along with the pros, cons, and pitfalls in terms of their usage in hydrological applications. Finally, a few important areas of future research are identified.

Keywords

Post-processing · Statistics · Model output · Forecasting · Meteorology · Hydrology · User needs · Uncertainty · Interpolation · Statistics · Stationarity · Correlation

1 Introduction

Output from numerical weather prediction (NWP) models is routinely used in operational flood forecasting to drive hydrological models which in the next link of the forecast chain predict future river discharge or inundation (Fig. 1). Within this context, the use of ensemble rather than deterministic forecasts is becoming

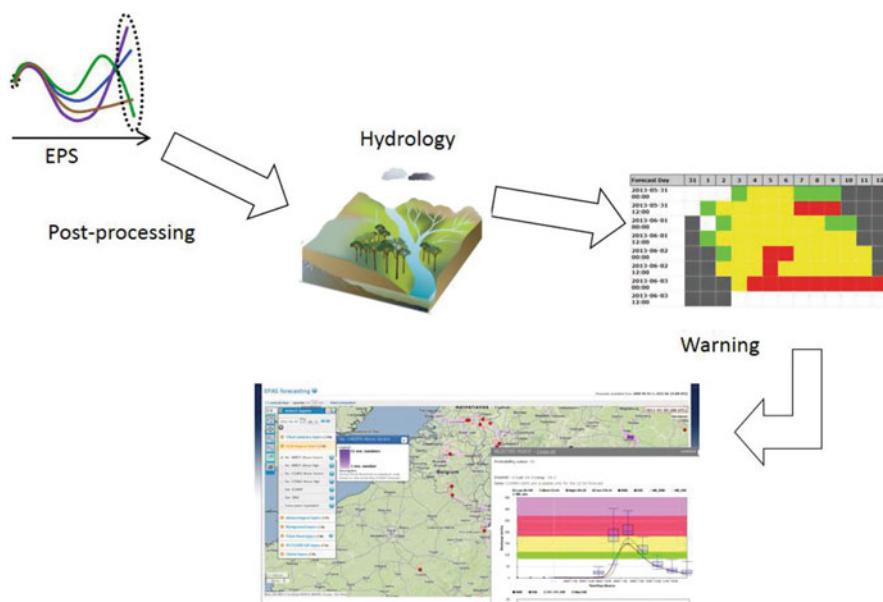


Fig. 1 Schematic view of the European Flood Awareness System (EFAS) from the NWP ensembles to the final warning

increasingly common (for an overview, see Cloke and Pappenberger 2009). In setting up the forecast chain, the hydrological models are usually calibrated and validated using observed meteorological forcing data, the so-called *perfect prognosis*.

However, in operational practice, the performance of these models will differ from that seen in calibration or validation due, in part, to the inability of the NWP models to perfectly predict the future values of the observed input variables such as precipitation or temperature.

As with all model-based forecasts of the natural environmental predictions made by NWP models are inherently uncertain. This uncertainty arises from many sources such as errors in the observations assimilated, model parameterization and structure, estimation of parameter values, initial conditions, numerical deficiencies, etc. which are not easy to untangle (Salamon and Feyen 2010).

The uncertainty is typically classified into two types *epistemic* and *aleatoric* (Beven and Smith 2015). The epistemic uncertainty is derived from inadequate knowledge and interpretation of both the physical phenomenon and modelling process (Klemeš 1986). This manifests itself through, for example, physical processes that are not fully understood being unsuccessfully translated into numerical solutions within the model. The aleatoric uncertainties derive, as the name suggests, from seemingly stochastic processes, perhaps the error in the observations or the chaotic behavior of the environment. In theory, uncertainties are reducible to only those that are only truly aleatoric (random). In practice it is unlikely that, in the foreseeable future, complete knowledge of all physical phenomena (or a close approximation to it) will be achieved. Given this, any separation of epistemic and aleatory uncertainty is itself subject to epistemic uncertainty, hence highly subjective.

Regardless of the source of the uncertainty, the ultimate result is NWP forecasts having errors in comparison with both the true state of the modelled systems and the observations of it. While the improvement in the forecasts that are relevant for hydrological applications, e.g., precipitation, temperature, and evapotranspiration, which has already been seen is likely to continue as NWP models develop (Haiden et al. 2014), it is unlikely that in the medium term the forecast error will become negligible (at least from the perspective of hydrological modelling) and that post-processing is often necessary (Madadgar et al. 2014).

For the purposes of this chapter, post-processing is considered to be “one or more procedures that statistically correct the NWP output with the aim of improving the accuracy, sharpness, and reliability of the forecast variables of interest when compared to observations.” The statistical models are themselves only approximations of the forecast error which is often complex having nonlinear spatial correlations. Within the meteorological literature, post-processing has also been referred to with a variety of other names such as calibration or model output statistics (MOS; Glahn and Lowry 1972).

The post-processing of NWP forecasts is not the only means of improving the resulting hydrological forecasts. Studies have shown that post-processing of the forecasted discharge is more efficient in terms of improving scores than post-

processing of the driving meteorological variables (Arheimer et al. 2011; Roulin and Vannitsem 2015). Similarly the assimilation of data into the hydrological model to improve the initial conditions of the forecast may be beneficial. Liu et al. (2012) give a broad overview of the different approaches and uses of data assimilation in hydrological forecasting as well as how these approaches can be transformed to an operational setting.

Which, if any of these methods, to use alongside or instead of the post-processing of the NWP forecasts depends on a number of factors including the timely availability of good quality observed data. It is also worth noting that the post-processing of hydrological forecasts is typically carried out to improve the quality of the forecasts at observed locations. Recently there have been developments to regionalize post-processing to also improve forecasts of ungauged basins (Skøien et al. 2016).

The HEPEX community has identified post-processing as one of the key areas of interest for future research (HEPEX 2015). At a workshop on post-processing in Toulouse in 2009, the following processes and techniques that needed to be addressed were identified (Schaake et al. 2010):

- Spatial and temporal consistency between hydrological and meteorological variables (*interpolation*)
- Removal of epistemic biases (*model output statistics*)
- Climatological consistency between meteorological observations and forcing (*reliability and stationarity*)
- Preservation of the *correlation* between the driving variables
- Improvement of forecast *sharpness*
- Reliable *uncertainty* estimates
- Extraction of relevant *information* from meteorological forecasts for end users

These points focus the discussion in the remainder of this chapter which, while leaving detailed descriptions to the references, discusses the pitfalls and advantages of existing post-processing methodologies as well as important areas of future research.

2 What the Hydrologist Needs from Post-processing

The hydrological forecasting community is very diverse and uses forecasts from meteorological forecasts, typically from NWP models, very differently depending on the specific needs of a particular application. For example, a forecaster who is more interested in short-range flash floods would require a forecast product with an extremely high spatial and temporal resolution which is frequently updated in real time (Alfieri et al. 2012; Liechti et al. 2013), whereas for a forecaster who is managing a dam or water supply infrastructure, it might be more important to correctly forecast the expected inflow (water balance) over longer time scales (Yuan et al. 2015). These differences may result in post-processing methodologies

that target improvements in different aspects of the NWP output. Further, hydrological models are applied in different climatic regions across the globe, with a varying degree of complexity, time step and availability of data for training and validation, etc. This leads to a variety of inputs which are generated from the NWP forecasts, for example, precipitation may be required on a gridded basis or as a catchment average value, which in turn may influence the chosen post-processing. In this section, some of the most important hydrological needs from NWP output are discussed.

2.1 Consistency with the Calibration Fields

In hydrometeorological forecasting, the hydrological model is driven by meteorological forecasts; however, it is in most cases calibrated using observations. The hydrological model is often assumed to behave similarly (i.e., error structure will be similar) regardless whether it uses weather observations or NWP output as forcing data. This assumption is violated if the operational forcing and calibration data come from different data sources or when the data used for calibration is not exactly the same data as used in validation (Bárdossy and Das 2008). If the meteorological forecast and observations have a different climatology (which often is the case), then the resulting forecast will possibly behave differently than the simulations driven by observations, thus resulting in biases in the forecasts. These differences can have a different error structure, a different climatology, etc. Post-processing of meteorological forecasts to better represent the observed fields before using them with a hydrological model is one way of reducing this bias.

While post-processing of meteorological variables usually improves the sharpness and removes biases from the forecasts, it can also introduce unwanted features such as inflated variability. This when combined with uncertainty processors of other error sources that play an important role in the total predictive uncertainty may give unreasonable forecast bounds, reflective in an inadequate (perhaps too simplistic) representation of uncertainty sources.

Post-processing to better match future observations is dependent on observational data of good quality. However, it is seldom straightforward to access data of good quality in real time. An alternative is to calibrate the hydrological model driven not by observed data but by a model climatology, generated using either hindcasts or the most recent model runs.

An example of this is the extreme forecast indices (EFI; Lalaurette 2003), which compare the forecasts to a reforecast of model climates to create an index of severity of the studied variable. This strategy still needs observations for validation, but these can be secondary or proxy variables, such as reported floods or discharge levels at downstream locations. If the forecast is used to assess the probability of an event with a certain return period to occur rather than forecasting of real thresholds, this method is useful. The method has also been developed to include an index of the severity of the event, so-called shift of tails (SOT; Zsoter 2006), which is a measure of the magnitude of the severity.

Here, one assumes that the weather forecast model used during the calibration (i.e., in hindcast mode) will be the same to what is used in real-time forecasting, which may not always be the case given that operational centers release new model versions on a regular basis and these are not always accompanied by a hindcast or reforecast. As an example, ECMWF routinely produces hindcasts (currently 20-year reforecasts run twice a week for the same day of year with a limited number of members) that are issued alongside new model cycles. This however means that the number of available hindcasts is not as extensive as for the operational forecasts.

2.2 Temporal Relevance

Since both meteorological and hydrological forecasts are issued for different lead times and different objectives, consideration should be given to which aspects of the forecasts post-processing should target. Post-processing is needed for forecasts at all time scales, from nowcasting (a few minutes to 12 h) short range (>12 h to a few days) for flash floods and medium range (3–15 days) for flooding to monthly to seasonal for groundwater and water management predictions.

Alfieri et al. (2012) listed the operational short- to medium-range meteorological forecasting methods in increasing temporal (and spatial) resolution as radar and ensemble radar nowcasting, radar-NWP blending, deterministic and ensemble-limited area NWP, and deterministic and ensemble global NWP (Fig. 2).

Radar precipitation forecasts can be very useful as an estimate of high-intensity precipitation, as was showed in the project HAREN (<http://haren-project.eu/>) and

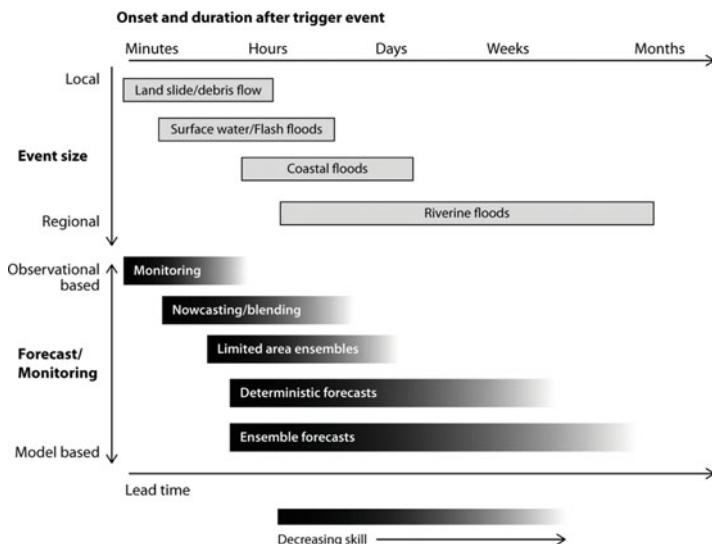


Fig. 2 Typical lead times for hazardous events and the lead time and skill horizons of operational forecast systems (Alfieri et al. 2012)

follow-up project EDHIT (<http://edhit.eu/>), and in combination with ensemble forecasts, it can be a powerful nowcasting tool, given that the limitations and uncertainties are sufficiently taken into account (Smith et al. 2014). The radar data also has to be available in real time and easily accessible for systems.

Flood forecasting of rivers, lakes, and reservoirs on longer lead times, ranging from 1 day up to weeks in advance, depends upon NWP models. The sensitivity of the hydrological forecasts to the NWP output depends on the physical properties of the watershed. The world-leading NWPs are typically skilful up to 15 days. However, the need for post-processing is still substantial due to the shortcomings of model output in capturing all aspects of the most important weather variables, such as precipitation, temperature, and evapotranspiration. The most important aspect for flood forecasting is the prediction of the magnitude, timing, and location of extreme precipitation events. Similar requirements, but potentially with less focus on the extreme, may be needed for the real-time management of water infrastructure, for example, in the dam management involved in the generation of hydroelectric power.

On the sub-seasonal to seasonal scale, it becomes more important to correctly model the water balance as these forecasts are potentially useful to assess, for example, the inflow into reservoirs from snowmelt or the risk of droughts or low flows. These forecasts are useful for seasonal planning and the post-processing needs to be able to correct systematic errors rather than extreme values. On the longer forecast time scales, there is also the possibility to mix dynamical climate models with statistical tools, for example, using known correlations of large-scale phenomenon such as NAO with local predictors to correct the model output.

2.3 Spatial Relevance

As already noted in hydrological applications, it is important that the NWP exhibits the same spatial correlation and variability as the observational data; otherwise, the assumptions that were used during the calibration are violated, for example, the correct representation of orographic lifting which will have large impact on both the location and the magnitude of precipitation, which in turn will affect the modelled flow (He et al. 2009). The spatial correlation between modelled NWP output and observations is inherently linked to the spatial resolution of the forecast. Computational constraints limit the current resolution in operational NWPs which in turn affect the output in terms of spatial smoothing. Further, the “effective resolution,” which is the resolution at which the NWP produces useful information, is often coarser than the formal output resolution.

It is to be hoped that these issues will become less prominent as models progress with higher resolution and improved description of the sub-grid physical processes (Haiden et al. 2014). Currently however post-processing is required to relate the meteorological forecasts on the “effective resolution” to the observed variability. This may involve elements of downscaling, the degree of which is dependent upon the hydrological model, to capture features at a scale smaller than the “effective resolution.” One aspect of this is discussed in the following section.

However post-processing could also be used to account for uncertainty in the spatial displacement of, for example, convective cells which may be significant if it results in precipitation falling in the wrong catchment.

2.4 Event Magnitude

Even a timely spatially correct forecast may not be of use in forecasting if the event magnitude is poorly predicted. The event magnitude is important in this aspect, pluvial flood events are caused entirely by intense precipitation.

The difficulties in predicting event magnitude are exemplified by two typical problems with general circulation models (GCMs), regardless whether they are applied as NWP or climate models. These are the overestimation of the number of rainy days (the “drizzle effect”; Murphy 1999) and the underestimation of extreme precipitation. These two contradicting errors (overestimation of low-intensity events and underestimation of the magnitude of high-intensity events) can result in a good estimation of the mean precipitation over time (Wetterhall et al. 2012). The “drizzle” problem attributes itself as too few days with zero precipitation and will cause problems for applications that are dependent on the succession of dry days to define, for example, dry spells (Wetterhall et al. 2015). The inability to correctly model heavy precipitation events limits the model’s usability in flood forecasting since intensity in precipitation is a flood-generating mechanism.

Both these deficiencies are to some extent related to the resolution, since the model describes the weather over each grid area. Sub-grid extremes will therefore not be represented correctly in the model. But there are other factors that contribute to the underestimation of both extremes. Precipitation is in itself an intermittent and nonlinear process, which makes it very difficult to model correctly. Important processes are parameterized and simplified which will have a negative effect on representing the observed precipitation distribution.

The correction of such deficiencies is one of the most common uses of statistical post-processing. Care though should be taken that the post-processing can deal with the event sizes that are most crucial for a particular location (Wetterhall et al. 2012).

2.5 Consistency Between Variables

Hydrological models are sensitive to preserving the consistency between the driving variables such as precipitation, temperature, and evapotranspiration in order to correctly model important process. This is, for example, particularly important during snowmelt periods which are to a large extent driven by temperature. The diurnal cycle of temperature will heavily affect the melting; a cold night will slow down any melting, whereas a warm cloudy night during winter will heavily accelerate the melting. Likewise, the evapotranspiration during warm, dry summer days may be misrepresented if the variables are individually post-processed.

Post-processing needs to preserve the consistency between variables to prevent unphysical behavior in the hydrological modelling. Roulin and Vannitsem (2015) show that only post-processing of precipitation did not improve the resolution of the hydrological ensembles.

3 How Meteorological Post-processing Affects Hydrological Prediction

There are a number of post-processing methods used to calibrate meteorological fields in hydrological applications, and for an overview of the used methods, we refer to other sections in this handbook or the vast literature on this topic (Glahn and Lowry 1972; Wilks 2011; Gneiting 2014). This section describes these methods, including the advantages and disadvantages from a hydrological perspective.

3.1 Spatial and Temporal Interpolation

In most forecasting chains, there is a mismatch of the resolution of the driving NWP and the hydrological model, and the driving data consequently has to be interpolated. The choice of interpolation method will depend on the variable, for example, can bilinear be useful for temperature and evapotranspiration, whereas a mass-conservative method might be more useful for precipitation. The interpolation can include a spatial correction, for example, using a DEM with high resolution to take into account local effects. He et al. (2009) used the previous months' precipitation to estimate a mass-conservative spatial correction of over a small catchment. The method was basically a reshuffling of the precipitation over the catchment without preserving the predominating pattern of the observed precipitation. Balsamo et al. (2010) used the opposite approach by preserving the spatial fields generated by the model, in that case ERA-Interim (Dee et al. 2011), and correcting the monthly mean to correspond to the global GPCC dataset. This method has the advantage that the modelled consistency between the output parameters is preserved and that the water balance over time is comparable to observations.

Correlation between variables needs to be addressed in any application, especially regarding spatially varying variables such as precipitation, and a rule of thumb is to coarsen the output spatially by taking the mean from the surrounding grid points rather than using individual points. Another technique is spatial pooling, where the neighboring points are included in the analysis. However, often the model output is taken as face value without consideration as to which spatial scale you would expect predictability. This creates a false sense of accuracy of the model system. Reducing the spatial resolution, either through averaging or pooling, can have large impacts in areas where there are sharp borders in the landscape, such as coastal zones and mountainous regions.

3.2 The Analogue Approach and Poor Man's Ensemble

Instead of using modelled output, another approach is to use historical observations as forecasts. In its simplest form, previous years are randomly selected and used as an ensemble forecast, or the “poor man’s ensemble.” The nice feature of this approach is that it is perfectly reliable. The analogue approach also reuses historical observations, but it makes the selection based on similarities between observed weather and the historical catalogue (Obled et al. 2002; Radanovics et al. 2013; Wetterhall 2005). The analogues are usually selected based on large-scale parameters, such as pressure fields, winds, and atmospheric humidity. Analogue methods work over smaller areas where there is a clear connection between the large-scale circulation and local variables, but they have also been used in country-wide applications. The main drawback is the stationarity of the relationship between predictor and predictand, and the methods need long time series of calibration.

3.3 Model Output Statistics

Model output statistics (MOS; Wilks 2011) has in NWP traditionally been applied through regression techniques in order to correct systematic errors in the output. Methods have been developed over the years and have gone from simpler mean bias correction techniques to full-blown probabilistic post-processing techniques (Gneiting 2014). One limitation of MOS is that it requires long time series of consistent model output. This is often not the case for hydrometeorological forecasts where the model development is constant, which has the effect that the operational output is nonstationary (i.e., not constant over time). These developments are, for example, through upgrades in the model physics, in the resolution, or in the initial conditions. This can be overcome if the NWP produces hindcasts which can be used for calibration (Hagedorn et al. 2008; Hamill et al. 2008) or by using the latest operational runs. The latter approach has the disadvantage of using only a small subsample of the historical period for the calibration, whereas hindcasts are usually run over a larger sample of the historical period. In the case of ECMWF, the hindcasts are run twice a week for the forecasts starting on a particular date over the previous 20 years.

The most common correction is to remove the mean bias from the forecast in comparison with the observations. This is straightforward if variables can be assumed to be normally distributed and the model error can be assumed linear. *Quantile mapping (QM)* is a technique where the cumulative distributions of the forecasted variables are adjusted to match those of the observed counterparts, for example (Wood and Lettenmaier 2006; Wetterhall et al. 2012; Yang et al. 2010; Themeßl et al. 2011). A disadvantage of QM for hydrological application is that in a multivariate system that the pairing between the variables is disrupted (Madadgar et al. 2014).

3.4 Bayesian Post-processing Methods

The most used methods in hydrological uncertainty analysis are Bayesian frameworks, which are used to derive the joint distribution of forecasts and observations. New data is used to update prior knowledge and likelihoods to provide conditional posterior distributions which estimate the predictive uncertainty of the predictands. Hydrological forecasting has used Bayesian uncertainty processors for a long time, and the methods cover a wide range of applications, such as Bayesian model averaging (BMA; Raftery et al. 2005), ensemble dressing technique (Roulston and Smith 2003; Wang and Bishop 2005), and ensemble kernel density MOS. Many BMA methods post-process one variable at one location for a specific time. However, as discussed earlier, consistency between variables are important, and it is necessary to account for the joint probability distributions corrected variables. There are methods for addressing this, for example, the “Schaake shuffle” which uses the rank order structure from past observations (Clark et al. 2004). Another approach is using ensemble copula coupling (Schefzik et al. 2013), which uses the multivariate rank dependence structure from the raw ensemble to produce corrected ensembles without destroying the correlation between the variables. Madadgar et al. (2014) suggested using a multivariate copula approach to account for the joint behavior of the included variables.

4 Future Challenges in Post-processing

Post-processing of NWP has come a long way in theory and praxis, and they are now routinely used in many hydrological and meteorological institutes (Gneiting 2014). However, there are areas that still need to be addressed, and in this section the most important of them are highlighted:

In the case of flood forecasting, predictive uncertainty can be defined as the uncertainty that a decision maker has on the future evolution of a predictand that he uses to make a specific decision. (Coccia and Todini 2011)

4.1 Extreme Values

One of the most important aspects of hydrological impact modelling is extreme events, both in terms of flash floods and riverine and groundwater floods. These events occur on different time scales and pose therefore naturally different demands on the post-processing. Flash floods caused by intense precipitation are mainly event driven, and the magnitude and location of convective storms are particularly challenging, both because of the NWP’s inability to model convection and due to the rapid developments of these thunderstorms. In this context, blending forecast model

output with, for example, remote sensing techniques (satellite, weather radar) has become increasingly used (Tafferner et al. 2008; Velasco-Forero et al. 2009). Modelling of extreme precipitation will most certainly improve in the future since processes that previously were parameterized will with higher resolution and increasing computer power be resolved explicitly, thus enabling a more realistic model output (Haiden et al. 2014). This is however not trivial, and it will still take many years before this is possible on a global scale. Statistical and dynamical downscaling are techniques that can bridge this gap in the meantime (Maraun et al. 2010).

4.2 Combining with Hydrological Post-processing

Studies have shown that post-processing of the forecasted discharge is more efficient in terms of improving scores than post-processing of the driving meteorological variables (Arheimer et al. 2011; Roulin and Vannitsem 2015). However, Roulin and Vannitsem (2015) did find that only using pre-processing of precipitation did not improve the resolution of the hydrological ensembles. Good initial conditions were necessary to improve resolution, but post-processing could compensate for this. Post-processing of meteorological forecasts can be combined with hydrologic post-processing through data assimilation techniques applied for model state variable updating or output error correction. It is important in those situations to separate the improvements from the pre- and post-processing to rightly account for the individual treatments, for example, through multivariate analysis.

4.3 Toward a Seamless Forecasting System

The holy grail of meteo-hydrological forecasting is the idea of seamless predictions with the basic concept of using one model system over all forecast horizon, in theory from nowcasting to multi-decadal forecasting (Pappenberger et al. 2013). The term “seamless” is misleading, since it assumes a forecast without “seams” between the modelling components. This idea is appealing but unrealistic and unnecessary since the demands on complexity, dissemination, and temporal and spatial resolution on the forecast will change with lead time. Therefore, “seemingly seamless” which implies an unnoticeable transition between forecasts over time more accurately describes the desired attributes of the forecast.

The simplest form of seamless forecasts concatenates different models of different spatial and temporal resolution. The resulting forecast will not be homogenous and biases will not be constant in time which puts hard demands on the post-processing to account for this. MOS for different lead times (i.e., model versions) might overcome some of these problems, such as distribution mapping (Wetterhall et al. 2012; Yang et al. 2010) and quantile transformation which are methods to address the biases. Di Giuseppe et al. (2013) proposed a mapping of the observed principal components on the modelled eigenvectors which results in spatially

unbiased forecasts. Temporal inconsistencies are more difficult to address in such a system, especially on the shorter lead times. The forecast models will inevitably have different initial conditions and weather developments which potentially can lead to forecast jumps in the transition from one model to the other.

5 Summary

This paper summarizes the hydrological needs of meteorological post-processing in order to make NWP model output useful for hydrological applications. It is important that post-processing not only removes the structural errors but also preserves the temporal and spatial correlation between variables and produces reliable forecasts. The most important techniques in post-processing and how they affect the NWP output are discussed. Finally, the most interesting future research directions are discussed as research areas where more research is needed.

References

- L. Alfieri, P. Salamon, F. Pappenberger, F. Wetterhall, J. Thielen, Operational early warning systems for water-related hazards in Europe. *Environ. Sci. Policy* **21**, 35–49 (2012)
- B. Arheimer, G. Lindström, J. Olsson, A systematic review of sensitivities in the Swedish flood-forecasting system. *Atmos. Res.* **100**, 275–284 (2011)
- G. Balsamo, S. Boussetta, P. Lopez, L. Ferranti, *Evaluation of ERA-Interim and ERA-Interim-GPCP-rescaled Precipitation Over the U.S.A.* ERA Report Series (European Centre for Medium Range Weather Forecasts, Reading, 2010)
- A. Bárdossy, T. Das, Influence of rainfall observation network on model calibration and application. *Hydrol. Earth Syst. Sci.* **12**, 77–89 (2008)
- K. Beven, P. Smith, Concepts of information content and likelihood in parameter calibration for hydrological simulation models. *J. Hydrol. Eng.* **20**(1) (2015). doi:10.1061/(ASCE)HE.1943-5584.0000991
- M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The Schaake Shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262 (2004)
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009)
- G. Coccia, E. Todini, Recent developments in predictive uncertainty assessment based on the model conditional processor approach. *Hydrol. Earth Syst. Sci.* **15**, 3253–3274 (2011)
- D.P. Dee, S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. Van De Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Källberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.J. Morcrette, B.K. Park, C. Peubey, P. De Rosnay, C. Tavolato, J.N. Thépaut, F. Vitart, The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. Roy. Meteorol. Soc.* **137**, 553–597 (2011)
- F. Di Giuseppe, F. Molteni, A.M. Tompkins, A rainfall calibration methodology for impacts modelling based on spatial mapping. *Q. J. Roy. Meteorol. Soc.* **139**, 1389–1401 (2013)
- H.R. Glahn, D.A. Lowry, The Use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**, 1203–1211 (1972)
- T. Gneiting, *Calibration of Medium-Range Weather Forecasts. Technicla Memorandum* (ECMWF, Reading, 2014)

- R. Hagedorn, T.M. Hamill, J.S. Whitaker, Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: two-meter temperatures. *Mon. Weather Rev.* **136**, 2608–2619 (2008)
- T. Haiden, L. Magnusson, I. Tsonevsky, F. Wetterhall, L. Alfieri, F. Pappenberger, P. De Rosnay, J. Muñoz-Sabater, G. Balsamo, C. Albergel, R. Forbes, T. Hewson, S. Malardel, D. Richardson, *ECMWF Forecast Performance During the June 2013 Flood in Central Europe* (European Centre for Medium-Range Weather Forecasts, Reading, 2014)
- T.M. Hamill, R. Hagedorn, J.S. Whitaker, Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Weather Rev.* **136**, 2620–2632 (2008)
- Y. He, F. Wetterhall, H.L. Cloke, F. Pappenberger, M. Wilson, J. Freer, G. McGregor, Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorol. Appl.* **16**, 91–101 (2009)
- HEPEX, *HEPEX-SIP Topic: Post-processing (1/3)* (2015) [Online]. Available: <http://hepex.irstea.fr/hepex-sip-topic-post-processing-13>. Accessed 10 June 2015
- V. Kleméš, Operational testing of hydrological simulation models. *Hydrol. Sci. J.* **31**, 13–24 (1986)
- F. Lalaurette, Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Q. J. Roy. Meteorol. Soc.* **129**, 3037–3057 (2003)
- K. Liechti, L. Panziera, U. Germann, M. Zappa, The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrol. Earth Syst. Sci.* **17**, 3853–3869 (2013)
- Y. Liu, A.H. Weerts, M. Clark, H.J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.J. Seo, D. Schwanenberg, P. Smith, A.I.J.M. Van Dijk, N. Van Velzen, M. He, H. Lee, S.J. Noh, O. Rakovec, P. Restrepo, Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrol. Earth Syst. Sci.* **16**, 3863–3887 (2012)
- S. Madadgar, H. Moradkhani, D. Garen, Towards improved post-processing of hydrologic forecast ensembles. *Hydrol. Process.* **28**, 104–122 (2014)
- D. Maraun, F. Wetterhall, A.M. Ireson, R.E. Chandler, E.J. Kendon, M. Widmann, S. Brienen, H.W. Rust, T. Sauter, M. Themeßl, V.K.C. Venema, K.P. Chun, C.M. Goodess, R.G. Jones, C. Onof, M. Vrac, I. Thiele-Eich, Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Rev. Geophys.* **48**, RG3003 (2010)
- J. Murphy, An evaluation of statistical and dynamical techniques for downscaling local climate. *J. Climate* **12**, 2256–2284 (1999)
- C. Obled, G. Bontron, R. Garçon, Quantitative precipitation forecasts: a statistical adaptation of model outputs through an analogues sorting approach. *Atmos. Res.* **63**, 303–324 (2002)
- F. Pappenberger, F. Wetterhall, E. Dutra, F. Di Giuseppe, K. Bogner, L. Alfieri, H.L. Cloke, *Seamless Forecasting of Extreme Events on a Global Scale* (IAHS-IAPSO-IASPEI Assembly, Gothenburg, 2013)
- S. Radanovics, J.P. Vidal, E. Sauquet, A. Ben Daoud, G. Bontron, Optimising predictor domains for spatially coherent precipitation downscaling. *Hydrol. Earth Syst. Sci.* **17**, 4189–4208 (2013)
- A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian Model Averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174 (2005)
- E. Roulin, S. Vannitsem, Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrol. Process.* **29**, 1434–1449 (2015)
- M.S. Roulston, L.A. Smith, Combining dynamical and statistical ensembles. *Tellus A* **55**, 16–30 (2003)
- P. Salamon, L. Feyen, Disentangling uncertainties in distributed hydrological modeling using multiplicative error models and sequential data assimilation. *Water Resources Research* **46**, n/a–n/a (2010)
- J. Schaake, J. Pailleux, J. Thielen, R. Arritt, T. Hamill, L. Luo, E. Martin, D. Mccollor, F. Pappenberger, Summary of recommendations of the first workshop on postprocessing and

- downscaling Atmospheric Forecasts for Hydrologic Applications held at Météo-France, Toulouse, France, 15–18 June 2009. *Atmos. Sci. Lett.* **11**, 59–63 (2010)
- R. Scheffzik, T.L. Thorarinsdottir, T. Gneiting, Uncertainty quantification in complex simulation models using ensemble copula coupling. **48**, 616–640 (2013)
- J.O. Skøien, K. Bogner, P. Salamon, P. Smith, F. Pappenberger, Regionalization of post-processed ensemble runoff forecasts. *Proc IAHS* **373**, 109–114 (2016)
- P.J. Smith, L. Panziera, K.J. Beven, Forecasting flash floods using data-based mechanistic models and NORA radar rainfall forecasts. *Hydrol. Sci. J.* **59**, 1403–1417 (2014)
- A. Tafferner, C. Forster, M. Hagen, C. Keil, T. Zinner, H. Volkert, Development and propagation of severe thunderstorms in the upper Danube catchment area: towards an integrated nowcasting and forecasting system using real-time data and high-resolution simulations. *Meteorol. Atmos. Phys.* **101**, 211–227 (2008)
- M.J. Themeßl, A. Gobiet, A. Leuprecht, Empirical-statistical downscaling and error correction of daily precipitation from regional climate models. *Int. J. Climatol.* **31**, 1530–1544 (2011)
- C.A. Velasco-Forero, D. Sempere-Torres, E.F. Cassiraga, J.J.A.I.M.E. Gómez-Hernández, A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. *Adv. Water Resour.* **32**, 986–1002 (2009)
- X. Wang, C.H. Bishop, Improvement of ensemble reliability with a new dressing kernel. *Q. J. Roy. Meteorol. Soc.* **131**, 965–986 (2005)
- F. Wetterhall, *Statistical Downscaling of Precipitation from Large-scale Atmospheric Circulation* (Philosophical Doctor/Uppsala University, Uppsala, 2005)
- F. Wetterhall, F. Pappenberger, Y. He, J. Freer, H.L. Cloke, Conditioning model output statistics of regional climate model precipitation on circulation patterns. *Nonlinear Processes Geophys.* **19**, 623–633 (2012)
- F. Wetterhall, H.C. Winsemius, E. Dutra, M. Werner, E. Pappenberger, Seasonal predictions of agrometeorological drought indicators for the Limpopo basin. *Hydrol. Earth Syst. Sci.* **19**, 2577–2586 (2015)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd edn. (Elsevier, Burlington, 2011)
- A.W. Wood, D.P. Lettenmaier, A test bed for new seasonal hydrologic forecasting approaches in the Western United States. *Bull. Am. Meteorol. Soc.* **87**, 1699–1712 (2006)
- W. Yang, J. Andreasson, P. Graham, J. Olsson, J. Rosberg, F. Wetterhall, Distribution-based scaling to improve usability of regional climate model projections for hydrological climate change impacts studies. *Hydrol. Res.* **41**, 211–229 (2010)
- X. Yuan, E.F. Wood, Z. Ma, A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *Wiley Interdiscip Rev Water* **2**, 523–536 (2015)
- E. Zsoter, Recent developments in extreme weather forecasting. *ECMWF Newsletter* (Reading, 2006)



Application to Post-processing of Meteorological Seasonal Forecasting

Andrew Schepen, Q. J. Wang, and David E. Robertson

Contents

1	Introduction	256
2	Methods	258
2.1	An Overview of CBaM	258
2.2	The Bayesian Joint Probability Modeling Approach	259
2.3	Model Parameter Inference	261
2.4	Forecasting	265
2.5	Forecast Merging Through Bayesian Model Averaging	266
2.6	Connecting Ensemble Members	268
3	Application of CBaM for Forecasting Catchment Rainfall	269
3.1	Overview	269
3.2	The GCM	270
3.3	The Catchment	270
3.4	Predictor Variables	270
3.5	Predictand Variable	272
3.6	Extending Forecasts to Longer Lead Times	273
3.7	Verification	273
4	Application Results	275
4.1	Overview	275
4.2	Overall Bias	276
4.3	Skill of Monthly Forecasts	276
4.4	Skill of Three-Month Accumulation Forecasts	278
4.5	Statistical Reliability	279
4.6	Discussion and Conclusion	280
	References	281

A. Schepen
CSIRO Land and Water, Dutton Park, QLD, Australia

Q. J. Wang (✉) · D. E. Robertson
CSIRO Land and Water, Clayton, VIC, Australia
e-mail: qj.wang@csiro.au

Abstract

Seasonal hydrological forecasting relies on accurate and reliable ensemble climate forecasts. A calibration, bridging, and merging (CBaM) method has been developed to statistically postprocess seasonal climate forecasts from general circulation models (GCMs). Postprocessing corrects conditional biases in raw GCM outputs and produces forecasts that are reliable in ensemble spread. The CBaM method is designed to extract as much skill as possible from the GCM. This is achieved by firstly producing multiple forecasts using different GCM output fields, such as rainfall, temperature, and sea surface temperatures, as predictors. These forecasts are then combined based on evidence of skill in hindcasts. Calibration refers to direct postprocessing of the target variable – rainfall for example. Bridging refers to indirect forecasting of the target variable – forecasting rainfall with the GCM's Nino3.4 forecast for example. Merging is designed to optimally combine calibration and bridging forecasts. Merging includes connecting forecast ensemble members across forecast time periods by using the "Schaake Shuffle," which creates time series forecasts with appropriate temporal correlation structure. CBaM incorporates parameter and model uncertainty, leading to reliable forecasts in most applications. Here, CBaM is applied to produce monthly catchment rainfall forecasts out to 12 months for a catchment in northeastern Australia. Bridging is shown to improve forecast skill in several seasons, and the ensemble time series forecasts are shown to be reliable for both monthly and seasonal totals.

Keywords

Seasonal forecasting · Post-processing · Bayesian joint probability · Bayesian model averaging · Precipitation · Temperature · Forecast verification

1 Introduction

Seasonal climate forecasting centers around the world now routinely run coupled ocean–atmosphere general circulation models (GCMs). GCMs are similar to numerical weather prediction (NWP) models, in that they solve physical equations numerically to produce forecasts of meteorological variables. GCMs have simplified or parameterized physics and are run on coarse grids but include coupling of ocean and atmosphere modules. The main aim of running GCMs is to produce intra- to inter-seasonal forecasts driven by the slowly evolving boundary conditions such as sea surface temperatures.

The development of GCMs has interested many in the streamflow forecasting community. A number of organizations around the world now officially produce seasonal streamflow forecasts. For example, the Australia Bureau of Meteorology releases probabilistic forecasts of the next 3 months streamflow at the beginning of each month (Tuteja et al. 2012). The forecasts released in Australia are based on a statistical model; however plans are in place to also deliver dynamic forecasts based on hydrological models and recent advances in GCMs.

In an ideal world, it would be convenient to take forecast variables from GCMs and input them directly into hydrological models. A problem with GCM forecast variables is that they are often biased relative to observations. This is largely due to simplifications and parameterizations that are unavoidable elements of a GCM. It is also due to the coarse scale of the GCMs. For example, GCMs can sometimes have difficulty simulating rainfall effects associated with locally complex or steep terrain.

Furthermore, beyond about 10 days from initialization, deterministic forecasts from a GCM are not necessarily realistic, due to the chaotic nature of the climate system. GCMs are therefore almost always run in ensemble mode. That is, multiple simulations are obtained by perturbing the initial conditions, using modified physics, or using initial conditions from different points in time. Often, the spread in ensemble forecasts produced using any of these methods is incorrectly dispersed and therefore incorrectly conveys forecast uncertainty. For an example see Lim et al. (2011). Most frequently, the ensembles are underdispersed, although it is quite possible that the ensembles are correctly dispersed or even overdispersed. Incorrectly dispersed ensembles fail to give reasonable estimates of forecast uncertainty.

If GCM ensembles are to be used in hydrological models, it is necessary to post-process the output to overcome biases and incorrectly dispersed ensembles. The severity of the problems is often a function of forecast lead time. For example, the longer a GCM run is, the more likely the ensemble mean a forecast variable is to drift from the mean of observed values. There are many approaches to post-processing in the literature. One such example is an analog downscaling method (Shao and Li 2013; Timbal and Jones 2008) that analyzes spatial patterns of the GCM forecasts of meteorological variables and then resamples the variable of interest, daily, from the historical record, based on a Euclidean distance metric.

However, seasonal forecasting is usually concerned with time scales of weeks to months. It is therefore worthwhile to ask the question – is it better to post-process daily or monthly variables? There are a number of reasons in favor of post-processing monthly variables. Predictability at the seasonal time scale arises due to the slowly evolving (low-frequency) climate signals. Monthly variables will therefore be less influenced by high-frequency weather noise and exhibit a stronger relationship with the source of predictability. After capturing the low-frequency climate signals in monthly post-processing, variables can be disaggregated to daily by, for example, using stochastic weather generators. In addition, monthly post-processing requires significantly fewer computational resources. Monthly approaches to post-processing GCM outputs have been demonstrated by Schepen and Wang (2014) and Hawthorne et al. (2013).

From a streamflow forecasting point of view, monthly rainfall forecasts can be directly used as input to a monthly hydrological model. Wang et al. (2011b) found that simulations of monthly streamflow volumes from the monthly *water partition and balance* (WAPABA) model are as skillful as those produced by two of the most widely used daily hydrological models in Australia for intra- to inter-seasonal forecasting of streamflows; it may therefore be sufficient to model at the monthly time step.

In this section, the Bayesian method of Schepen and Wang (2014) for post-processing monthly GCM outputs will be presented. The method is referred to as CBaM, which stands for calibration, bridging, and merging. Calibration is focused on correcting GCM forecasts of variables that may be biased and unreliable in conveying uncertainty through ensemble spread. Bridging forecasts are produced by using GCM forecasts of climate indices, such as sea surface temperature anomalies, to indirectly forecast variables. Sometimes, the GCM is unable to translate skillful forecasts of climate indices into skillful forecasts of regional variables. Bridging can overcome this problem. Merging is designed to optimally combine calibration and bridging forecasts. Merging includes connecting forecast ensemble members across forecast time periods by using the “Schaake Shuffle,” which creates time series forecasts with appropriate temporal correlation structure (Clark et al. 2004).

The method builds on the earlier Bayesian joint probability modeling approach of Wang et al. (2009) and Wang and Robertson (2011), and the Bayesian model averaging (BMA) approach of Wang et al. (2012b). The method will be outlined in detail to support implementation. An example application of CBaM to post-process catchment rainfall in northeastern Australia will highlight the suitability of the approach for producing time series forecasts in the form of ensembles. The example will focus on 12-month forecasts of monthly rainfall and seasonal accumulations.

2 Methods

2.1 An Overview of CBaM

The calibration, bridging, and merging method is underpinned by a number of separate steps.

The first step in CBaM is to establish multiple statistical models that relate predictor variables derived from coupled GCM output fields to observed variables. A Bayesian joint probability modeling approach (Wang and Robertson 2011; Wang et al. 2009) is applied for this purpose. The model is referred to as either a calibration model or a bridging model depending on the predictor variable in the model. This distinction will be expanded upon in the next section.

The second step in CBaM is to reforecast the variable of interest (the predictand) using each of the calibration and bridging models. This is done by casting the calibration and bridging models as forecasting models by conditioning the joint probability models on predictor values.

The third step in CBaM is to infer a weight for each calibration and bridging model based on historical performance. Historically better performing models receive higher weight. Bayesian model averaging (Wang et al. 2012b) is applied for this purpose.

The fourth step in CBaM is to merge the forecasts from multiple calibration and bridging models based on the weights. Bayesian model averaging is applied to merge the ensemble forecasts.

The fifth and final step in CBaM is to connect the ensemble members of forecasts for individual months to form ensemble time series forecasts. The Schaeke Shuffle (Clark et al. 2004) is applied so that the ensemble time series have a realistic temporal correlation structure. This final step is important for variables that have a strong temporal correlation structure.

2.2 The Bayesian Joint Probability Modeling Approach

2.2.1 Overview

Wang et al. (2009) and Wang and Robertson (2011) developed a sophisticated Bayesian joint probability (BJP) modeling approach to address many of the vexatious problems that arise in the statistical modeling of environmental data. The BJP modeling approach is able to handle data with missing records, data records of different lengths, skewed data, heterogeneous error structures, and bounded variables. It is thus ideal for post-processing hydroclimatological data. At the core of the BJP modeling approach is a multivariate normal distribution. Use of a symmetric continuous multivariate distribution eases the statistical modeling and overcomes the problems listed above, as will be described below.

To present a formulation here that is as simple as possible, the formulation of BJP models presented here is for a single predictor and a single predictand, i.e., a bivariate formulation. The full multivariate formulation of BJP models is given by Wang et al. (2009) and Wang and Robertson (2011).

2.2.2 Model Setup

Consider a predictor variable y_1 and a predictand variable y_2 such that

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix}. \quad (1)$$

The relationship of the variables can be modeled as a bivariate distribution. Meteorological and hydrological variables can sometimes have a skewed distribution. For example, while temperature typically follows a normal distribution, rainfall and streamflow often have right-skewed distributions. The BJP modeling approach allows for this by incorporating data transformations. It is thus the distribution of transformed variables that is assumed to follow a continuous bivariate normal distribution. The transformed predictor and predictand variables are \hat{y}_1 and \hat{y}_2 , such that

$$\hat{\mathbf{y}} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad (2)$$

where $\boldsymbol{\mu}$ is the vector of variable means and $\boldsymbol{\Sigma}$ is the covariance matrix. That is,

$$\boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \quad (3)$$

and

$$\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & r\sigma_1\sigma_2 \\ r\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \quad (4)$$

where σ_i is a standard deviation and r is the correlation coefficient.

In CBaM, if y_1 and y_2 refer to the same variable (e.g., y_1 is a temperature forecast and y_2 is observed temperature), then the model is a calibration model. If y_1 and y_2 refer to different variables (e.g., y_1 is a sea surface temperature index forecast and y_2 is observed temperature), then the model is a bridging model. Apart from the conceptual difference, the distinction is merely for convenience, as the BJP models are otherwise formulated identically.

There are two transformations commonly used in the BJP modeling approach. For variables like rainfall and streamflow that can be quite skewed and which are bounded below by zero, a two-parameter log-sinh transform is preferred (Wang et al. 2012a). More specifically,

$$\hat{y} = \frac{1}{\beta} \ln(\sinh(\alpha + \beta y)). \quad (5)$$

For other variables the log-sinh transformation is not suitable. For example, the log-sinh transformation cannot be used with sea surface temperature (SST) climate indices due to the presence of negative values in the data. A single-parameter Yeo–Johnson transform (Yeo and Johnson 2000) can be used instead. In this case,

$$\hat{y} = \begin{cases} \left((y+1)^\lambda - 1 \right) / \lambda & \lambda \neq 0, y \geq 0 \\ \log(y+1) & \lambda = 0, y \geq 0 \\ -\left((-y+1)^{2-\lambda} - 1 \right) / 2 - \lambda & \lambda \neq 2, y < 0 \\ -\log(-y+1) & \lambda = 2, y < 0 \end{cases}. \quad (6)$$

In the BJP modeling approach, the predictor and predictand variables do not have to follow the same transformation. In CBaM this most commonly occurs in bridging models. For example, an SST predictor variable will follow a Yeo–Johnson transformation, whereas a rainfall predictand variable will follow a log-sinh transformation.

Meteorological and hydrological variables sometimes have a lower bound (e.g., rainfall and streamflow have a natural lower bound of zero). This is problematic, as a probability mass at a lower bound implies a mixed discrete–continuous distribution, which is not compatible with the use of the bivariate normal distribution. The solution to this problem is to treat the occurrences of values equal to the lower bound as censored data. The lower bounds corresponding to the variables y_1 and y_2 are denoted c_1 and c_2 , respectively, such that

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix}. \quad (7)$$

When a value of y_i is c_i , the assignment $y_i \leq c_i$ is made. In other words, the variable is hypothetically allowed to be below an actual lower bound, but its precise value is unknown.

2.3 Model Parameter Inference

2.3.1 Posterior Distribution of the Parameters and Numerical Sampling

The model parameters, including the bivariate normal distribution parameters and the transformation parameters for each variable, are collected in $\boldsymbol{\theta}$. A Bayesian inference of all the parameters in $\boldsymbol{\theta}$ is made with historical data records for events in $\mathbf{y}_D = [\mathbf{y}_D^1, \mathbf{y}_D^2, \dots, \mathbf{y}_D^t, \dots, \mathbf{y}_D^T]$. This inference accounts for parameter uncertainty. In application of CBaM to the problem of monthly and seasonal forecasting, parameter uncertainty is considerable due to the limited data available for use in model parameter inference. A typical number for T is about 30.

According to Bayes' theorem, the posterior distribution of the model parameters is defined by

$$p(\boldsymbol{\theta}|\mathbf{y}_D) \propto p(\boldsymbol{\theta})p(\mathbf{y}_D|\boldsymbol{\theta}) = p(\boldsymbol{\theta}) \prod_{t=1}^T p(\mathbf{y}_D^t|\boldsymbol{\theta}) \quad (8)$$

where $p(\boldsymbol{\theta})$ is a prior distribution of the parameters and $\prod_{t=1}^T p(\mathbf{y}_D^t|\boldsymbol{\theta})$ is the likelihood function defining the probability of observing the historical data \mathbf{y}_D given the model and its parameter set.

A numerical solution to Eq. 8 is required. A full Bayesian inference of the joint distribution of the model parameters is performed using Markov chain Monte Carlo (MCMC) sampling. More specifically, the Metropolis algorithm with a symmetric, Gaussian proposal distribution is applied to draw samples from the posterior distribution. In CBaM it is typical to retain 1000 sets of parameters from MCMC sampling. Inferences about the parameters can then be made from the drawn samples. Further details of the MCMC sampling used in CBaM are given in the subsequent sections.

2.3.2 Prior Distribution of the Parameters

The prior distribution of the parameters is given by

$$p(\boldsymbol{\theta}) = p(r) \prod_{i=1}^2 p(\Delta_i) p(\mu_i, \sigma_i^2) \quad (9)$$

where Δ_i is the set of model transformation parameters.

If a variable is under a log-sinh transformation, then the set of transformation parameters is

$$\Delta_i = \{\alpha_i, \beta_i\}. \quad (10)$$

If a variable is under a Yeo–Johnson transformation, then the set of transformation parameters is

$$\Delta_i = \{\lambda_i\}. \quad (11)$$

A uniform prior is specified for each of the transformation parameters. That is,

$$p(\Delta_i) \propto 1 \quad (12)$$

The prior for the model parameters μ_i and σ_i^2 takes the simplest form commonly used for mean and variance (Gelman et al. 1995):

$$p(\mu_i, \sigma_i^2) = 1/\sigma_i^2. \quad (13)$$

The prior for r is the marginally uniform prior (Barnard et al. 2000). In the bivariate case, it reduces simply to

$$p(r) \propto 1. \quad (14)$$

2.3.3 Likelihood Function

The calculation of the likelihood function differs for combinations of variables above and below their respective censor thresholds. For each event time period t , the calculation of the likelihood function is given by (the superscript t is dropped for simplicity)

$$p(y|\boldsymbol{\theta}) = \begin{cases} p(y_1, y_2|\boldsymbol{\theta}) & = J_{\hat{y}_1 \rightarrow y_1} J_{\hat{y}_2 \rightarrow y_2} p(\hat{y}_1, \hat{y}_2|\boldsymbol{\theta}) \quad y_1 > c_1, y_2 > c_2 \\ p(y_1 \leq c_1, y_2|\boldsymbol{\theta}) & = J_{\hat{y}_2 \rightarrow y_2} p(\hat{y}_1 \leq \hat{c}_1, \hat{y}_2|\boldsymbol{\theta}) \quad y_1 \leq c_1, y_2 > c_2 \\ p(y_1, y_2 \leq c_2|\boldsymbol{\theta}) & = J_{\hat{y}_1 \rightarrow y_1} p(\hat{y}_1, \hat{y}_2 \leq \hat{c}_2|\boldsymbol{\theta}) \quad y_1 > c_1, y_2 \leq c_2 \\ p(y_1 \leq c_1, y_2 \leq c_2|\boldsymbol{\theta}) & = p(\hat{y}_1 \leq \hat{c}_1, \hat{y}_2 \leq \hat{c}_2|\boldsymbol{\theta}) \quad y_1 \leq c_1, y_2 \leq c_2 \end{cases} \quad (15)$$

where $J_{\hat{y}_i \rightarrow y_i}$ is the Jacobian determinant for the transformation from \hat{y}_i to y_i .

If the log-sinh transformation is applied to y_i , then

$$J_{\hat{y}_i \rightarrow y_i} = \coth(\alpha_i + \beta_i y_i). \quad (16)$$

If the Yeo–Johnson transformation is applied to y_i , then

$$J_{\hat{y}_i \rightarrow y_i} = \begin{cases} (y_i + 1)^{\lambda_i - 1} & y_i \geq 0 \\ (-y_i - 1)^{1 - \lambda_i} & y_i < 0. \end{cases} \quad (17)$$

The remaining terms in Eq. 15 are given by

$$p(\hat{y}_1 \leq \hat{c}_1, \hat{y}_2 | \boldsymbol{\theta}) = p(\hat{y}_2 | \boldsymbol{\theta}) \int_{-\infty}^{\hat{c}_1} p(\hat{y}_1 | \hat{y}_2, \boldsymbol{\theta}) d\hat{y}_1, \quad (18)$$

$$p(\hat{y}_1, \hat{y}_2 \leq \hat{c}_2 | \boldsymbol{\theta}) = p(\hat{y}_1 | \boldsymbol{\theta}) \int_{-\infty}^{\hat{c}_2} p(\hat{y}_2 | \hat{y}_1, \boldsymbol{\theta}) d\hat{y}_2, \quad (19)$$

and

$$p(\hat{y}_1 \leq \hat{c}_1, \hat{y}_2 \leq \hat{c}_2 | \boldsymbol{\theta}) \int_{-\infty}^{\hat{c}_1} \int_{-\infty}^{\hat{c}_2} p(\hat{y}_1, \hat{y}_2 | \boldsymbol{\theta}) d\hat{y}_1 d\hat{y}_2. \quad (20)$$

The standard bivariate conditional distribution is applied to obtain $p(\hat{y}_1 | \hat{y}_2, \boldsymbol{\theta})$ and $p(\hat{y}_2 | \hat{y}_1, \boldsymbol{\theta})$. For example,

$$p(\hat{y}_1 | \hat{y}_2, \boldsymbol{\theta}) \sim N\left(\mu_{\hat{y}_1} + r \frac{\sigma_{\hat{y}_1}}{\sigma_{\hat{y}_2}} (\hat{y}_2 - \mu_{\hat{y}_2}), (1 - r^2) \sigma_{\hat{y}_1}\right)^{(1-r^2)}. \quad (21)$$

2.3.4 Reparameterization of Model Parameters

It is well documented that attempts to sample model parameters using the Metropolis algorithm with a Gaussian proposal can be problematic if some of the parameters are nonlinearly related. The problem manifests as low Metropolis acceptance rates (Thyer et al. 2002; Wang and Robertson 2011) and thus inefficient sampling. Therefore, a number of parameters are reparameterized to ease the MCMC inference and make for more efficient sampling.

The parameters μ and σ are reparameterized to m and s through first-order Taylor series approximations where m and s are approximations of the mean and standard deviation of y , respectively.

Where the log-sinh transformation is applied to a variable, the reparameterization of μ and σ is as follows:

$$\mu = \frac{1}{\beta} \ln(\sinh(\alpha + \beta m)) \quad (22)$$

and

$$\sigma = \coth(\alpha + \beta y)s. \quad (23)$$

The Jacobian of the reparameterization is

$$J_{\mu, \sigma^2 \rightarrow m, s^2} = [\coth(\alpha + \beta y)]^3. \quad (24)$$

Where the Yeo–Johnson transformation is applied to a variable, the reparameterization of μ and σ is as follows:

$$\mu = \begin{cases} ((m+1)^\lambda - 1)/\lambda & \lambda \neq 0, m \geq 0 \\ \log(m+1) & \lambda = 0, m \geq 0 \\ -((-m+1)^{2-\lambda} - 1)/(2-\lambda) & \lambda \neq 2, m < 0 \\ -\log(-m+1) & \lambda = 2, m < 0 \end{cases} \quad (25)$$

and

$$\sigma = \begin{cases} (m+1)^{\lambda-1}s & m \geq 0 \\ (-m-1)^{1-\lambda}s & m < 0. \end{cases} \quad (26)$$

The Jacobian of the reparameterization is

$$J_{\mu, \sigma^2 \rightarrow m, s^2} = \begin{cases} (m+1)^{3\lambda-3} & m \geq 0 \\ (-m-1)^{3-3\lambda} & m < 0 \end{cases}. \quad (27)$$

Further reparameterization of s^2 , α , and β to $\ln(s^2)$, $\ln(\alpha)$, and $\ln(\beta)$ allows for parameter inference on the entire real space and an approximately linear dependence between the estimated parameters, yielding more efficient MCMC sampling (Robertson et al. 2013). The Jacobians for the further reparameterizations are

$$J_{\alpha \rightarrow \ln(\alpha)} = \alpha, \quad (28)$$

$$J_{\beta \rightarrow \ln(\beta)} = \beta, \quad (29)$$

and

$$J_{s^2 \rightarrow \ln(s^2)} = s^2. \quad (30)$$

The correlation coefficient is reparameterized through a Fisher Z transformation to give

$$\varphi = \tanh^{-1}(r). \quad (31)$$

The reparameterized correlation coefficient is asymptotically normal (Zhu and Hero 2007). The associated Jacobian is

$$J_{r \rightarrow \varphi} p(r) = [\cosh(\varphi)]^{-2}. \quad (32)$$

After reparameterization, the prior distribution of the parameters is given by

$$p(\boldsymbol{\theta}) = J_{r \rightarrow \varphi} p(r) \prod_{i=1}^2 J_{\Delta_i \rightarrow \hat{\Delta}_i} p(\Delta_i) J_{s_i^2 \rightarrow \ln(s_i^2)} J_{\mu_i, \sigma_i^2 \rightarrow m_i, s_i^2} p(\mu_i, \sigma_i^2) \quad (33)$$

where $J_{\Delta_i \rightarrow \hat{\Delta}_i} = J_{\alpha \rightarrow \ln(\alpha)} J_{\beta \rightarrow \ln(\beta)}$ if y_i is under a log-sinh transformation and $J_{\Delta_i \rightarrow \hat{\Delta}_i} = J_{\lambda_i \rightarrow \hat{\lambda}_i} = 1$ if y_i is under a Yeo–Johnson transformation.

2.4 Forecasting

A probabilistic forecast of y_2 is produced using a new value of y_1 . The following sections describe the forecast procedure: firstly, for a single set of parameters and, secondly, for multiple sets of parameters as derived from the full Bayesian inference.

2.4.1 Forecast Procedure for a Single Set of Parameters

The conditional (forecast) distribution for a single set of parameters, $\boldsymbol{\theta}$, is univariate normal and given by

$$p(y_2|y_1, \boldsymbol{\theta}) = J_{\hat{y}_2 \rightarrow y_2} p(\hat{y}_2|\hat{y}_1, \boldsymbol{\theta}) \quad (34)$$

where

$$p(\hat{y}_2|\hat{y}_1, \boldsymbol{\theta}) \sim N\left(\mu_{\hat{y}_2} + r \frac{\sigma_{\hat{y}_2}}{\sigma_{\hat{y}_1}} (\hat{y}_1 - \mu_{\hat{y}_1}), (1 - r^2)\sigma_{\hat{y}_1}^2\right) \quad (35)$$

and $J_{\hat{y}_2 \rightarrow y_2}$ is given by Eq. 18 or Eq. 19 depending on the transformation applied to y_2 .

A forecast ensemble member is obtained by making a draw of \hat{y}_2 from the distribution given by Eq. 35 and back-transforming to y_2 using the inverse of the transformation given in Eqs. 6 or 7.

Equation 35 is not used precisely as written when $\hat{y}_1 \leq \hat{c}_1$. Instead, data augmentation is used to draw a random value, $\hat{y}_{1,AUG}$, from the distribution $p(\hat{y}_1|\boldsymbol{\theta})$ that satisfies $\hat{y}_{1,AUG} < \hat{c}_1$, which is then used in Eq. 35 in place of \hat{y}_1 . That is, values of \hat{y}_{12} are drawn from the following distribution:

$$p(\hat{y}_1|\hat{y}_{1,AUG}, \boldsymbol{\theta}) \sim N\left(\mu_{\hat{y}_2} + r \frac{\sigma_{\hat{y}_2}}{\sigma_{\hat{y}_1}} (\hat{y}_{1,AUG} - \mu_{\hat{y}_1}), (1 - r^2)\sigma_{\hat{y}_1}^2\right). \quad (36)$$

An ensemble of the desired size is produced by repeatedly and randomly sampling ensemble members.

2.4.2 Forecast for Multiple Sets of Parameters

A forecast incorporating parameter uncertainty corresponds to the posterior predictive density:

$$\begin{aligned} f(y_2) &= \int p(y_2|y_1, \boldsymbol{\theta}) p(\boldsymbol{\theta}|y_D) d\boldsymbol{\theta} \\ &\approx \frac{1}{N} \sum_{n=1}^N p(y_2|y_1, \boldsymbol{\theta}_n) \end{aligned} \quad (37)$$

where N is the number of parameter sets retained from MCMC sampling.

To obtain a forecast ensemble that is representative of Eq. 37, a small number of ensemble members are drawn for each parameter set $\boldsymbol{\theta}_n$ ($n = 1, 2, \dots, N$), using the procedure outlined in Sect. 2.4.1. The resulting ensemble members are pooled.

As indicated in Sect. 2.3.1, a typical number for N is 1000. If one sample is drawn for each parameter set, the size of the forecast ensemble will also be 1000.

2.5 Forecast Merging Through Bayesian Model Averaging

Forecast merging in CBaM is accomplished through Bayesian model averaging (BMA). BMA merges forecast predictive densities and can therefore be applied to achieve weighted ensemble forecasts of K different models. The specific version of Bayesian model averaging (BMA) implemented in CBaM was developed by Wang et al. (2012b).

A merged forecast is the BMA predictive density, which is a weighted average of the individual model predictive densities. It follows from Eq. 37 that

$$f_{BMA}(y_2) = \sum_{k=1}^K w_k f_k(y_2) \quad (38)$$

and, similarly,

$$F_{BMA}(y_2) = \sum_{k=1}^K w_k F_k(y_2) \quad (39)$$

where F_{BMA} is the cumulative distribution function of the merged forecast.

The weights w_k can be specified in many ways. For example, equal weights could be specified. However, equal weights are unlikely to be optimal. In CBaM, optimal BMA weights are sought. For models $k = 1, 2, \dots, K$ and event time periods $t = 1, 2, \dots, T$, a Bayesian inference is made for the weights using forecasts $f_k(y_2)$ and corresponding events $y_{D,2}^t$. The posterior distribution of the weights is

$$\begin{aligned} & p(w_k | y'_{D,2}, f'_k(y_2), k = 1, 2, \dots, K, t = 1, 2, \dots, T) \\ & \propto p(w_k, k = 1, 2, \dots, K) p(y'_{D,2}, t = 1, 2, \dots, T | w_k, f'_k(y_2), k = 1, 2, \dots, K, t = 1, 2, \dots, T). \end{aligned} \quad (40)$$

The RHS of Eq. 40 is the product of the likelihood of the weights, given a set of forecast and corresponding observations, and the prior distribution of weights.

The prior distribution of the weights is specified as a symmetric Dirichlet distribution,

$$p(w_k, k = 1, \dots, K) \propto \prod_{k=1}^K (w_k)^{\alpha-1} \quad (41)$$

where α is the concentration parameter. More evenly distributed weights among the models are encouraged when $\alpha > 1$. In CBaM α is a function of the number of models; more specifically $\alpha = 1 + a_0/K$. A typical value for a_0 is 0.5 or 1.0. Such a prior helps stabilizing the weights in presence of significant sampling variability.

In the calculation of the likelihood function, the set of time periods $\mathbb{T} = \{1, 2, \dots, t, \dots, T\}$ is partitioned into two sets. One set, \mathbb{T}_1 , contains time periods for which $y'_{D,2} > c_2$. The second set, \mathbb{T}_2 , contains time periods for which $y'_{D,2} \leq c_2$.

The calculation of the likelihood function is given by

$$\begin{aligned} & p(y'_{D,2}, t = 1, 2, \dots, T | w_k, f'_k(y_2), k = 1, 2, \dots, K, t = 1, 2, \dots, T) \\ & = \prod_{t \in \mathbb{T}_1} f'_{BMA}(y'_{D,2}) \prod_{t \in \mathbb{T}_2} F'_{BMA}(y'_{D,2} = c_2) \end{aligned} \quad (42)$$

where

$$f'_{BMA}(y'_{D,2}) = \sum_{k=1}^K w_k f'_k(y'_{D,2}) \quad (43)$$

and

$$F'_{BMA}(y'_{D,2} = c_2) = \sum_{k=1}^K w_k F'_k(y'_{D,2} = c_2) = \sum_{k=1}^K w_k \int_{-\infty}^{c_2} f'_k(y_2) dy_2. \quad (44)$$

A point estimate of the weights is obtained by maximizing the posterior distribution of the weights. This is the maximum a posteriori (MAP) solution, which can be found by using an iterative expectation–maximization (EM) algorithm (Cheng et al. 2006; Zivkovic and van der Heijden 2004). The algorithm is as follows.

Given weights $w_k^{(j)}$, $k = 1, \dots, K$ at iteration j , first, $O_k^{t,(j+1)}$ is calculated for all t and k , where

$$O_k^{t,(j+1)} = \begin{cases} \frac{w_k^{(j)} f_k^t(y_{D,2}^t)}{\sum_{m=1}^K w_m^{(j)} f_m^t(y_{D,2}^t)} & y_{D,2}^t > c_2 \\ \frac{w_k^{(j)} F_k^t(y_{D,2}^t = c_2)}{\sum_{m=1}^K w_m^{(j)} F_m^t(y_{D,2}^t = c_2)} & y_{D,2}^t \leq c_2 \end{cases}. \quad (45)$$

Then, new weights are calculated by

$$w_k^{j+1} = \frac{\left(\frac{1}{T}\right) O_k^{t,(j+1)} + (a-1)/T}{1 + K(a-1)/T}. \quad (46)$$

Equations 44 and 45 are iteratively applied until the posterior probability of the weights converges. The EM algorithm is only guaranteed to converge to a local minimum. It is therefore wise to choose reasonable starting values for the weights $w_k^{(0)}$. A conservative choice in this regard is to use equal weights to start the first iteration.

An alternative to using the model fitted predictive densities $f_k^t(y_2|y_1)$ is to replace them with cross validation predictive densities $f_k^{(t)}(y_2|y_1)$. This is the approach of Shinozaki et al. (2010). The likelihood function in Eq. 41 is subsequently replaced with a cross validation likelihood function (Rust and Schmittlein 1985; Shinozaki et al. 2010; Smyth 1996, 2000; Stone 1977). By using a cross validation likelihood function instead of the classical likelihood function, the weights are assigned according to the model predictive abilities rather than fitting abilities. Indeed, there is much literature in support of using predictive performance measures for model choice and combination based on the idea that a model is only as good as its predictions (e.g., Eklund and Karlsson 2007; Geweke and Whiteman 2006; Jackson et al. 2009).

BMA applies to averaging to forecast probability densities (and thus cumulative probabilities) for a given forecast variable value. An alternative is to average variable values (quantiles) for a given cumulative probability (quantile fraction). This is the method of quantile model averaging (QMA; Schepen and Wang 2015). Merging unimodal forecasts such as those from CBaM with QMA yields a unimodal merged forecast. BMA can sometimes produce multimodal forecasts with excessively wide uncertainty bands. QMA may therefore be more practical for some applications; however it is not expanded upon here.

2.6 Connecting Ensemble Members

In CBaM, ensemble forecasts for different time periods and lead times are post-processed independently. Due to the inherent randomness in sampling the forecasts,

the forecasts do not automatically have an appropriate temporal correlation structure. If the forecasts are to be used as time series, for example, in hydrological modeling, it is necessary to establish ensembles with realistic temporal correlation structure.

By way of example, if forecasting monthly temperature, a time series constructed by taking equally positioned ensemble members may indiscriminately swing between relatively hot and cold states. This behavior does not seem reasonable, as an ensemble member forecasting hot for one time period would be expected to forecast warm or hot for the next time period (or more) and so on. A variable can have weak or strong correlation from one forecast period to the next. One pragmatic way to achieve the appropriate behavior is to reorder ensemble members by using a procedure known as the Schaake Shuffle (Clark et al. 2004). The Schaake Shuffle is performed as follows:

- Obtain historical data for a similar time period to that being forecast (e.g., month).
- For each historical event time period $t = 1, 2, \dots, T$, assign a rank to each corresponding observation by magnitude.
- For each forecast ensemble member in an ensemble of size T , also assign a rank by magnitude.
- Reorder the forecast ensemble so that the pairs of forecast ensemble members and historical observations have equal ranks. For example, if the 5th year in the historical record has the largest magnitude, the forecast ensemble member with the largest magnitude will be put in 5th position.

The Schaake Shuffle is performed separately on ensemble forecasts for different time periods and lead times. However, once it is done, the temporal rank-correlation structure across forecast lead times will implicitly be established. Ensemble members with the same position can therefore be used as time series forecasts. More detailed descriptions of the Schaake Shuffle can be found elsewhere in this handbook, and tutorial-style examples are given by Clarke et al. (2004). It is noted here, however, that the Schaake Shuffle is also useful in restoring the appropriate spatial correlation structure among ensemble forecasts for different locations.

One restriction of the Schaake Shuffle is that the number of ensemble members that can be reordered at once has to be equal to the number of historical observations. In CBaM, the number of forecast ensemble members greatly exceeds the number of historical observations. The Schaake Shuffle is performed in batches by breaking the forecast into chunks of size T and connecting each chunk separately.

3 Application of CBaM for Forecasting Catchment Rainfall

3.1 Overview

CBaM is applied to post-process GCM rainfall forecasts for the Barron river catchment in the northeast of Australia. The predictors are derived from outputs from a coupled GCM with atmospheric horizontal resolution of 250 km. The

predictands are derived from 5 km gridded observed rainfall. This application of CBaM involves establishing multiple Bayesian joint probability (BJP) models, each with a single predictor and a single predictand. The forecasts from calibration and bridging models are weighted and merged with BMA. Ensemble members are connected by using the Schaake Shuffle to create time series ensemble forecasts. Finally, the monthly forecasts are aggregated to seasonal (3-month) forecasts.

3.2 The GCM

The GCM adopted for this application is the Predictive Ocean Atmosphere Model for Australia (POAMA) version M2.4 (Wang et al. 2011a). POAMA is a coupled ocean–atmosphere GCM operated by the Australian Bureau of Meteorology specifically for intra- to inter-seasonal climate forecasting. In hindcast mode, M2.4 is initialized on day 1 of each month and run for 9 months. M2.4 is set up to provide daily and monthly forecasts of many atmospheric and oceanic climate variables, including pressure, temperature, and rainfall. The horizontal resolution of the atmospheric grids is coarse at approximately 250 km. There are three variants of the model (simply named M24A, M24B, and M24C), with only minor differences in the model structure. All three variants are initialized with identical, jointly perturbed atmospheric and ocean initial conditions (Marshall et al. 2012). Each variant outputs an 11-member ensemble. In this application, all ensemble members are treated equally and pooled together to form a 33-member ensemble. Monthly data are used in this application.

3.3 The Catchment

The upper Barron river catchment is located in a tropical climatic zone. A map displaying the location of the catchment together with the boundary of M2.4 grid cells is shown in Fig. 1. The catchment experiences heavy rainfall during the so-called wet season from October through to April. The monthly and seasonal rainfall totals can be dominated by several rainfall events associated with intense active phases of the monsoon and the passing of tropical cyclones. Throughout the remainder of the year, rainfall is relatively low and infrequent. The distributions of observed monthly and seasonal (3-month accumulation) rainfall are plotted later in the results section (Figs. 2 and 3).

3.4 Predictor Variables

In a calibration model, y_1 is the GCM forecast of catchment rainfall. Since rainfall has a lower bound of zero, the censor threshold of zero is imposed and the log-sinh transformation is applied. If a catchment sits within one POAMA grid cell, the calibration model predictor variable is simply taken as the ensemble mean for that

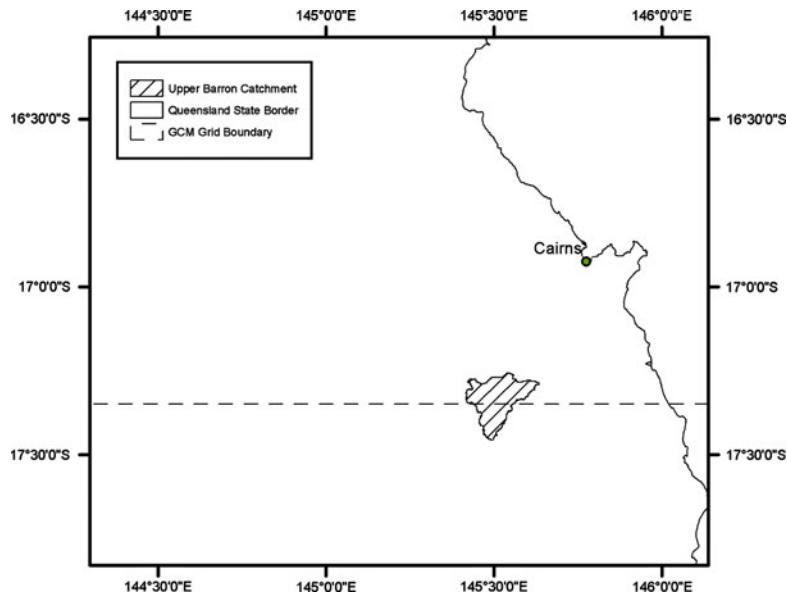


Fig. 1 Map showing the location of the upper Barron catchment in Queensland, Australia. The boundaries of POAMA M2.4 grid cells are also plotted

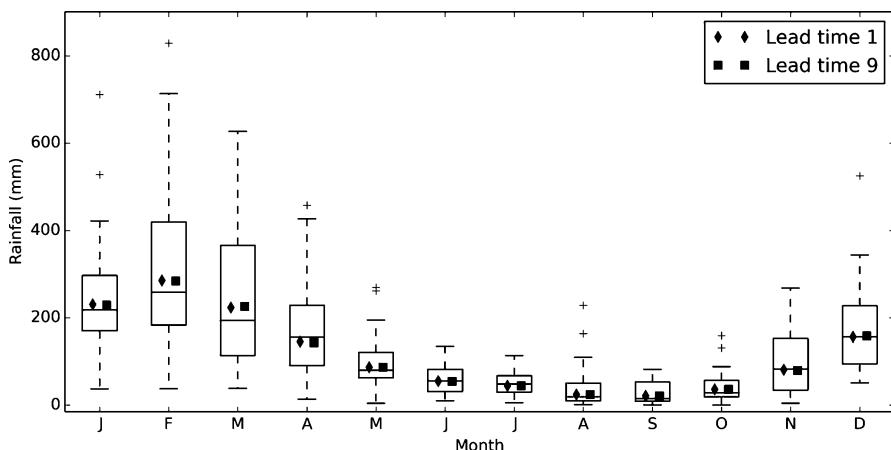


Fig. 2 Box plots describing the distribution of observed monthly rainfall for the upper Barron catchment. Observed data for 1982–2010 is used. The markers show the overall median of merged (CBaM) forecasts for the same period

grid cell. However, as shown by Fig. 1, the upper Barron river catchment boundary intersects with two or more GCM grid cells. A straightforward way to derive the calibration predictor variable is to take an area-weighted average of the ensemble means of the grid cells.

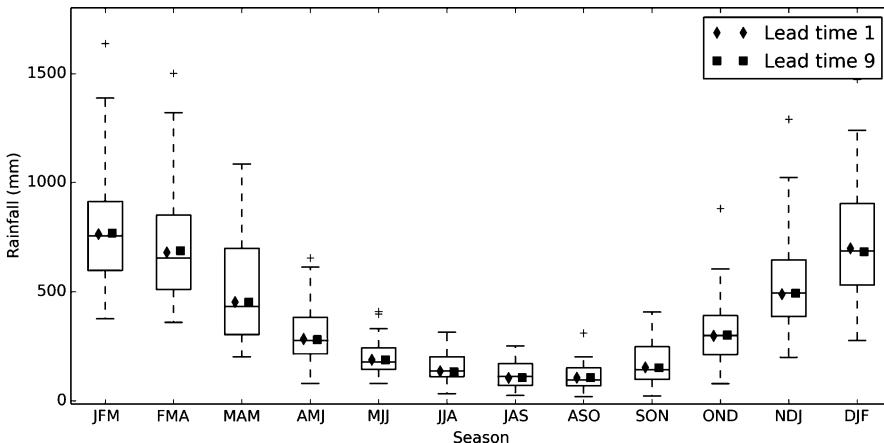


Fig. 3 As Fig. 2, but for 3-month accumulation (seasonal) rainfall

In a bridging model, y_1 is the GCM forecast of a different variable. In this application, the bridging variables are derived from sea surface temperature forecasts. More specifically, the bridging variables are six indices that capture fluctuations in tropical Indian and Pacific Ocean sea surface temperatures (SSTs). The indices calculated from the M2.4 ensemble mean SST anomaly forecast. The SST indices are NINO3, NINO3.4, NINO4, El Niño Modoki (Ashok et al. 2007), the Indonesian Index (Verdon and Franks 2005), and the Indian Ocean Dipole (Saji et al. 1999). Since the variables take negative values, the variables are transformed using Yeo–Johnson transformations. The variables are not censored. In other applications, bridging variables could be derived from, for example, atmospheric pressure, winds, or subsurface sea temperatures.

These indices are suitable for application of CBaM to Australia as it has been shown that fluctuations in the El Niño–Southern Oscillation (ENSO) and Indian Ocean SSTs are linked to fluctuations in Australian rainfall (Risbey et al. 2009; Schepen et al. 2012). Refer to Table 1 for locations of the SST regions used in the calculation of the indices.

3.5 Predictand Variable

In a CBaM model, y_2 is an observed variable. In this application, the observed variable is average catchment rainfall. Similarly to the rainfall predictor variable, the rainfall predictand variable is transformed using a log-sinh transformation, and a lower bound of zero is imposed. Observed rainfall data is derived from the Australian Water Availability Project (AWAP) 5 km gridded dataset of monthly rainfall (Jones et al. 2009).

Table 1 Definitions of climate indices based on sea surface temperature anomalies

Index	Definition
NINO3	Average over 150°W–90°W, 5°S–5°N
NINO3.4	Average over 120°W–170°W, 5°S–5°N
NINO4	Average over 160°E–150°W, 5°S–5°N
EMI (El Niño Modoki)	C – 0.5(E + W), where C = average over 165°E–140°W, 10°S–10°N E = average over 110°W–70°W, 15°S–5°N W = average over 125°E–145°E, 10°S–20°N
DMI (Indian Ocean Dipole)	WPI – EPI, where EPI = average over 90°E–110°E, 10°S–0°S WPI = average over 50°E–70°E, 10°S–10°N
II (Indonesian Index)	Average over 120°E–130°E, 10°S–0°S

3.6 Extending Forecasts to Longer Lead Times

Ensemble rainfall forecasts produced using CBaM are intended for use in hydrological models to produce streamflow forecasts. For water management planning, streamflow forecasts are desirable for up to 12 months in advance. POAMA only produces 9 months of forecasts; however, CBaM can be applied to produce forecasts for a number of months beyond the end of the GCM run, thus enabling longer lead-time streamflow forecasts.

CBaM is applied here to produce forecasts for 12 months. The forecasts are referred to as having lead times from 0 to 11 months. The lead time refers to the number of months between the initialization day of the GCM and the first day of the forecast period. For example, a 0-month lead-time forecast for January is made using forecast data from a GCM model run that is initialized on the 1st of January.

Forecasts are able to be made out to 12 months by establishing the calibration and bridging models with lagged predictor–predictand relationships. That is, the predictor values from the 9th month of the POAMA run are used to forecast the 10th–12th months.

3.7 Verification

3.7.1 Cross Validation

In this application, forecasts for the period 1982–2010 are verified. That is, forecasts are produced for each month in 1982–2010 with 0–11 months lead time. Although very long records (100 years+) of observed rainfall data are available for the upper Barron catchment, M2.4 hindcasts are first initialized for the year 1981. Separate CBaM models are established for each month and lead time independently.

With the short period of available data, it is necessary to establish CBaM models using data for the same events that are being forecast. It is therefore necessary to

assess forecasting performance through cross validation to obtain realistic estimates of forecast quality. Cross validation involves hiding predictor and predictand data that is deemed to not be independent of the verifying event in the establishment of CBaM models. This is repeated for each event in the dataset. In this application, leave-3-years-out cross validation is used. For example, to establish the calibration and bridging models for March 1983, all predictor and predictand data for March 1982–1984 is removed. Cross validation predictive densities are used to develop weights for the merging BMA models. Continuing the previous example, for a March 1983 forecast, the cross validation predictive densities for the March 1982–1984 events are omitted in the calculation of the weights. For the ensemble connection step (the Schaake Shuffle), only the data for the year of event to be forecast is left out. To again continue the previous example for a March 1983 forecast, observed data for 1983 is removed for the Schaake Shuffle.

3.7.2 Tools for Assessing Skill and Reliability

As a first check of the forecasts, a visual assessment is made of the overall bias of the forecasts. More detailed checks are then undertaken. Forecast accuracy and sharpness are two desirable attributes of probabilistic forecasts. If forecasts are sharp, then there is a strong tendency for forecast probabilities to deviate from climatological probabilities. A score that is commonly applied to jointly assess the accuracy and sharpness of probabilistic forecasts is the continuous ranked probability score (CRPS; Matheson and Winkler 1976). The average CRPS for event time periods $t = 1, 2, \dots, T$ is given by

$$\text{CRPS} = \frac{1}{T} \sum_{t=1}^T \int [F^t(y) - H(y - y_D^t)]^2 dy \quad (47)$$

where $F^t(y)$ is the forecast cumulative distribution function and H is the Heaviside step function defined as

$$H(y - y_D^t) = \begin{cases} 0 & y < y_D^t \\ 1 & y \geq y_D^t \end{cases} \quad (48)$$

CRPS skill scores are calculated to measure the relative improvement in the average CRPS of the CBaM forecasts over average CRPS for simple reference forecasts. The formulation for CRPS skill scores therefore follows the formulation for a generalized skill score:

$$\text{CRPS}_{SS} = \frac{\text{CRPS}_{\text{ref}} - \text{CRPS}}{\text{CRPS}_{\text{ref}}} \times 100 \quad (49)$$

A CRPS skill score of 100 is the maximum score and implies forecasts that are both perfectly sharp and accurate. A skill score of 0 implies that model forecasts are on average no better than the reference forecasts. The generalized skill score is negatively unbounded. A negative skill score implies that the model forecasts are on

average worse than the reference forecasts. In this application, the reference forecasts are taken as cross validation climatology.

Reliability of a set of probabilistic forecasts refers to the statistical consistency of forecast probabilities with observed values. For instance, an outcome with forecast probability of 20% should be observed in 20% of cases. Statistical reliability of the probabilistic forecasts is assessed by plotting observed values transformed through probability integral transforms (PITs) in a uniform probability plot. The PIT $F^t(y)$ for a probabilistic forecast for event at time period t when applied to the observed value for the same event gives

$$\pi^t = F^t(y_D^t). \quad (50)$$

If the forecasts are statistically reliable, then the PIT values should theoretically follow a standard uniform distribution. One way to check for the expected uniformity is to pool together all the values of π^t for $t = 1, 2, \dots, T$ and plot the ranked values in a uniform probability plot. This PIT uniform probability plot (termed predictive QQ plot by Thyre et al. (2009)) is useful for indicating whether the forecasts are systematically biased and whether the forecast uncertainty is too narrow or too wide (Laio and Tamea 2007). The plots can include the 95% Kolmogorov significance bands for a more formal test of uniformity. The width of the band is calculated as $1.358/\sqrt{T}$ (Laio and Tamea 2007). The forecasts are deemed to be statistically reliable if all the PIT values fall within the significance bands.

It is noted that the PIT histogram is an alternative method for displaying PIT uniformity but more suited for large samples. An alternative graphical tool for assessing probabilistic reliability is the attributes or reliability diagram for binary expressions of the probabilistic forecasts (Hsu and Murphy 1986). In addition to reliability assessment, attributes diagrams permit visual checking of forecast sharpness and resolution. The attributes diagram is more suited to large sample sizes and is therefore not considered here.

4 Application Results

4.1 Overview

For the upper Barron catchment, we assess the skill and reliability of the CBaM rainfall forecasts for the period 1982–2010. Monthly forecasts are assessed for lead times of 0–11 months, and 3-month accumulation forecasts are assessed for lead times of 0–9 months. The CRPS skill scores of the calibration, bridging, and merged forecasts are compared. As with any post-processing method, the skill that is evident in the CBaM forecasts stems predominantly from the intrinsic skill in the GCM. CBaM is designed to capitalize on the intrinsic skill in the GCM by correcting biases, adjusting forecast uncertainty, and merging forecasts based separately on the atmospheric and oceanic modules.

Although the focus is on monthly rainfall forecasts, reliability of the 3-month accumulation forecasts is also a critical concern. Failure to properly connect the ensemble members using the Schaake Shuffle can theoretically lead to accumulated forecasts that are too narrow in ensemble spread and therefore not reliable. This is unlikely to be a major issue with precipitation forecasts as monthly rainfall exhibits low autocorrelation.

4.2 Overall Bias

One of the aims of CBaM is to produce forecasts that are overall unbiased relative to observed catchment rainfall. It is important that the bias be minimized so that the ensemble forecasts are suitable as inputs to hydrological models. To check that the CBaM forecasts do not suffer from overall biases, the overall medians of the CBaM forecasts are plotted against the distributions of observed data for each month and season.

The plot for monthly rainfall is Fig. 2 and the plot for 3-month accumulation rainfall is Fig. 3. The distribution of observed catchment rainfall is summarized by a box plot. The box spans the interquartile range (IQR) with a line at the median. The whiskers extend to the most extreme point within $1.5 \times \text{IQR}$, and the plus symbols represent the remaining data points. It is clear from Figs. 2 and 3 that the observed rainfall data follows a skewed distribution. The diamond markers are the overall medians of the 1-month lead-time merged forecasts considering all forecasts for that period in 1982–2010. Similarly, the square markers are the overall medians of the 9-month lead-time merged forecasts. The medians of the merged forecasts show little deviation from the observed medians, confirming that CBaM generally produces monthly forecasts that are overall unbiased, and 3-month accumulation forecasts are also overall unbiased. The monthly merged forecasts do show a slight positive bias in the median for February and March, the months with the most intense rainfall.

4.3 Skill of Monthly Forecasts

The CRPS skill scores for each month and lead time are shown in Fig. 4. There are 12 bars plotted for each month corresponding to lead times of 0–11 months. The top panel shows skill scores for the calibration forecasts, the middle panel shows skill scores for the bridging forecasts, and the bottom panel shows skill scores for the merged forecasts. It is evident that both calibration and bridging forecasts for the upper Barron catchment are skillful for some months and lead times. However, it is also evident that the forecasts are not skillful for many other months and lead times. The monthly calibration forecasts are mostly skillful for November for lead times of up to 8 months. The monthly bridging forecasts exhibit similar patterns of skill to the monthly calibration forecasts, except skill scores for October and March are higher

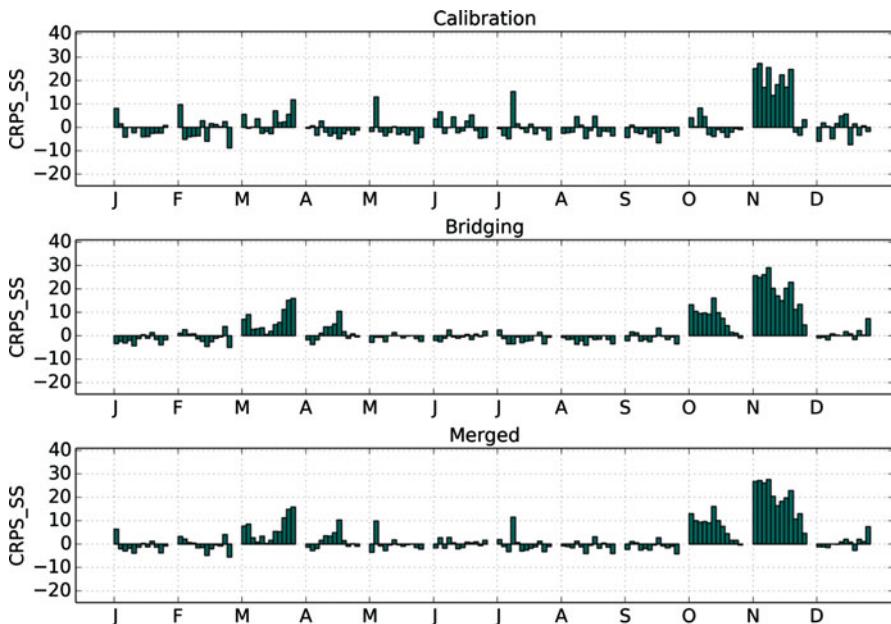


Fig. 4 Comparison of CRPS skill scores for monthly calibration, bridging, and merged forecasts. Forecasts are for the period 1982–2010 and based on leave-3-years-out cross validation. For each month, there is one bar for each forecast lead time of 0–11 months, in order

for the bridging forecasts. In this application, the bridging forecast skill improves upon the calibration forecast skill, and it is evident that merging the calibration and bridging forecasts takes advantage of the skill of both calibration and bridging forecasts.

Because the skill scores for monthly rainfall forecasting are typically low, it is difficult to discern real positive skill scores from random positive skill scores related to sampling variability. If in a particular application of CBaM it is deemed necessary to quantify the significance of skill scores, then a pragmatic bootstrap procedure could be used to determine a skill score threshold above which positive skill is judged to be discernible. Such a procedure is described by Schepen and Wang (2014). In their study, which used the root mean square error in probability (RMSEP) score, a skill score of 5 was established as bare minimum threshold above which skill becomes increasingly discernible and significant (i.e., skill scores less than 5 were to be disregarded). Higher skill scores and/or a pattern of skill across multiple target months and/or lead times are seen to increase the credibility of positive skill score results. In this application the aim is simply to extract as much skill as possible from the forecasts through the use of calibration, bridging, and merging.

4.4 Skill of Three-Month Accumulation Forecasts

The ensemble connection step of CBaM connects ensemble members of forecasts for different lead times to form time series forecasts that have a realistic correlation structure across lead times. This allows aggregation of rainfall ensemble members to obtain accumulation forecasts. Any number of months can be accumulated. The most common number of months to accumulate is three, corresponding to seasonal forecasts. Hence 3-month accumulations are the focus here.

The CRPS skill scores for each 3-month accumulation period and lead time are shown in Fig. 5. The top panel shows skill scores for the calibration forecasts, the middle panel shows skill scores for the bridging forecasts, and the bottom panel shows skill scores for the merged forecasts. It is evident that the 3-month accumulation forecasts are skillful for more periods and lead times than the monthly forecasts. The calibration forecasts are predominantly skillful from SON to NDJ. The bridging forecasts show a similar pattern of skill across lead times; however the skill scores are generally higher than those for the calibration forecasts. Skill is evident for the bridging forecasts from SON to FMA. Although skill is seen to mostly decrease with lead time, the bridging forecasts are evidently skillful in some cases for very long lead times. For example, the bridging forecasts for OND are skillful for all 12 lead times. Similar to the monthly forecasts, the merged 3-month accumulation forecasts take advantage of the skill in both the calibration and bridging forecasts.

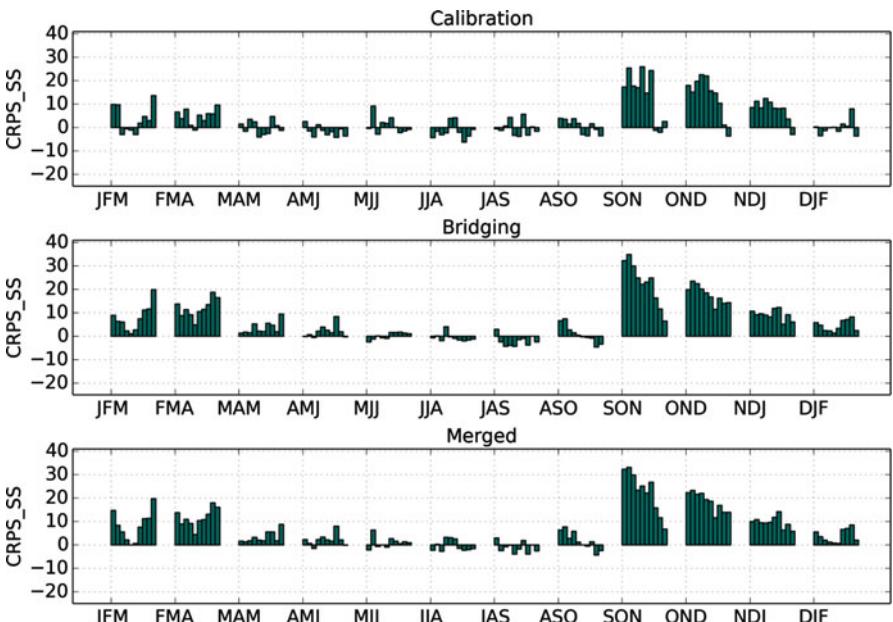


Fig. 5 As Fig. 4, but for 3-month accumulation forecasts

4.5 Statistical Reliability

The statistical reliability of the merged forecasts is visually assessed using PIT uniform probability plots. If the forecasts are found to be statistically reliable, it means that the forecast ensembles typically have appropriate spread and therefore correctly characterize forecast uncertainty. PIT uniform probability plots are presented for monthly and 3-month accumulation forecasts, and for lead times of 1 month and 9 months (Figs. 6 and 7). There are 12 panels in each figure, each corresponding to a month or 3-month period. For each month and 3-month period, and for both lead times, every PIT value falls within the 95% Kolmogorov significance bands. It can be concluded that the merged forecasts are statistically reliable. That the 3-month accumulation forecasts are reliable implies that when the ensemble members are taken as time series, the ensembles have an appropriate temporal correlation structure.

It is noted that the PIT uniform probability plot will not reveal some types of systematic problems with the forecasts. For example, Wang et al. (2009) also plot the PIT values against the event magnitude as well as the forecast year. If such plots reveal patterns or trends in the PIT values, then the forecasts could be conditionally biased.

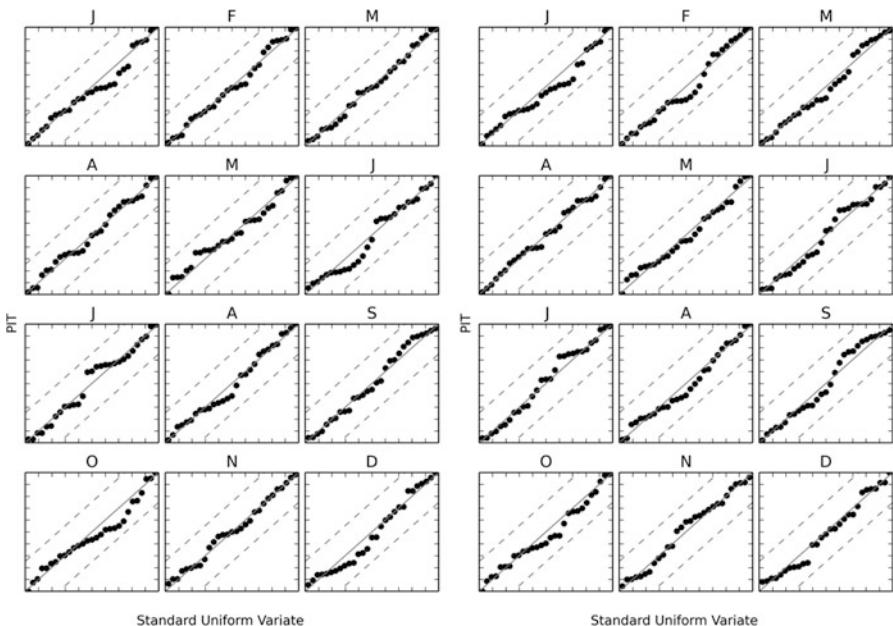


Fig. 6 PIT uniform probability plots for the monthly merged forecasts for the period 1982–2010. Left panel is for forecasts at a lead time of 1 month, and right panel is for forecasts at a lead time of 9 months

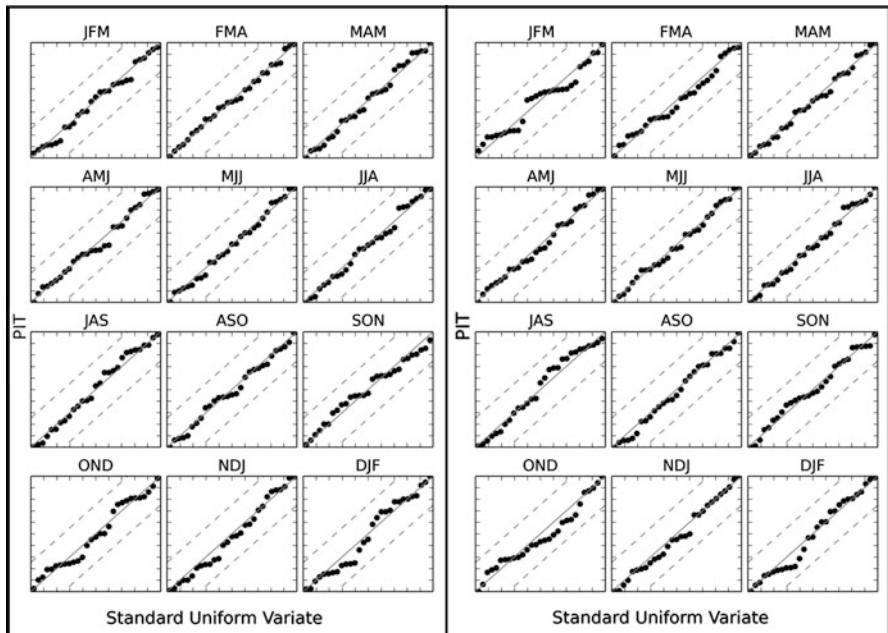


Fig. 7 As Fig. 6 but for 3-month accumulation forecasts

4.6 Discussion and Conclusion

CBaM is designed to harness the skill in the raw CGCM forecasts and produce reliable ensemble time series of monthly meteorological variables. The application demonstrates that CBaM produces useful ensembles of rainfall that can be used to force a hydrological model calibrated against observed data. CBaM forecasts exhibit low overall bias and extract skill from both the atmospheric and oceanic modules of the GCM.

The variation in skill between catchments and months and seasons is related to a combination of the natural predictability of regional and remote influences on the catchment climate and the quality of the raw CGCM forecasts. In some months or 3-month periods, it is clear that the raw CGCM forecasts cannot provide particularly skillful forecasts of inter-annual rainfall variability.

The skill of monthly rainfall forecasts is generally much lower than the skill of 3-month accumulation forecasts. Monthly rainfall is more influenced by climate features that occur on shorter time scales and that are inherently less predictable. For example, monthly rainfall in the upper Barron during the skillful period (November–March) is likely to be more strongly influenced by phases of the monsoon than is the 3-month accumulated rainfall. The 3-month accumulated rainfall in the upper Barron is also likely to be influenced by the phases of the monsoon, but is additionally likely to be influenced by ENSO, which is inherently more predictable.

The application here did not consider a catchment with zero rainfall. Monthly or seasonal rainfall totals of zero are not common. The CBaM methodology is applicable at shorter time scales (e.g., weeks), and in such applications the requirement to properly handle zero values would become more evident.

In the merging step of CBaM, it is not assumed that the predictors provide independent information. Some approaches (e.g., Luo et al. 2007) assume independent information in the predictors. As a result, merging identical forecasts reduces the forecast uncertainty. In contrast, if identical forecasts are merged using Bayesian model averaging in CBaM, the forecast uncertainty will remain unchanged. Explicit model selection is not a part of CBaM. The weight given to each model is determined in the merging step.

For streamflow forecasting, the dominant factors in determining forecast skill are the initial catchment conditions and accurate representation of the soil moisture states in the model. Even if the CBaM forecasts are climatological in nature, the incorporation of forecast uncertainty through MCMC sampling may mean the ensemble members are more suitable for forcing the hydrological model than are historical observations.

The results show there is a strong need to improve the skill of monthly and 3-monthly accumulated rainfall forecasts. Such improvements are seen as coming through advances in GCM modeling rather than in statistical post-processing techniques. It is noted that other GCMs may produce different results, although the broad patterns of predictability are expected to be similar given that it is large-scale climate features that give rise to monthly and 3-monthly forecasting skill.

References

- K. Ashok, S. Behera, S. Rao, H. Weng, T. Yamagata, El Niño Modoki and its possible teleconnection. *J. Geophys. Res.* **112**, C11007 (2007)
- J. Barnard, R. McCulloch, X.-L. Meng, Modeling covariance matrices in terms of standard deviations and correlations, with application to shrinkage. *Stat. Sin.* **10**, 1281–1312 (2000)
- J. Cheng, J. Yang, Y. Zhou, Y. Cui, Flexible background mixture models for foreground segmentation. *Image Vis. Comput.* **24**, 473–482 (2006)
- M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The Schaeake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262 (2004)
- J. Eklund, S. Karlsson, Forecast combination and model averaging using predictive measures. *Econ. Rev.* **26**, 329–363 (2007)
- A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, *Bayesian data analysis* (Chapman and Hall, New York, 1995). 526 pp
- J. Geweke, C. Whiteman, Chapter 1, Bayesian forecasting, in *Handbook of Economic Forecasting*, ed. by G. Elliott, C.W.J. Granger, A. Timmermann, vol. 1 (Elsevier B.V., Amsterdam, 2006), pp. 3–80. ISSN 1574-0706, ISBN 9780444513953, [https://doi.org/10.1016/S1574-0706\(05\)01001-3](https://doi.org/10.1016/S1574-0706(05)01001-3)
- S. Hawthrone, Q.J. Wang, A. Schepen, D. Robertson, Effective use of general circulation model outputs for forecasting monthly rainfalls to long lead times. *Water Resour. Res.* **49**(9), 5427–5436 (2013)

- W.-R. Hsu, A.H. Murphy, The attributes diagram. A geometrical framework for assessing the quality of probability forecasts. *Int. J. Forecast.* **2**, 285–293 (1986)
- C.H. Jackson, S.G. Thompson, L.D. Sharples, Accounting for uncertainty in health economic decision models by using model averaging. *J. R. Stat. Soc. A. Stat. Soc.* **172**, 383–404 (2009)
- D.A. Jones, W. Wang, R. Fawcett, High-quality spatial climate data-sets for Australia. *Aust. Meteorol. Oceanogr. J.* **58**, 233–248 (2009)
- F. Laio, S. Tamea, Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.* **11**, 1267–1277 (2007)
- E.-P. Lim, H.H. Hendon, D.L.T. Anderson, A. Charles, O. Alves, Dynamical, statistical-dynamical, and multimodel ensemble forecasts of Australian spring season rainfall. *Mon. Weather Rev.* **139**, 958–975 (2011)
- L. Luo, E.F. Wood, M. Pan, Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.* **112**, D10102 (2007)
- A.G. Marshall, D. Hudson, M.C. Wheeler, H.H. Hendon, O. Alves, Evaluating key drivers of Australian intra-seasonal climate variability in POAMA-2: a progress report. *CAWCR Res. Lett.* **7**, 10–16 (2012)
- J.E. Matheson, R.L. Winkler, Scoring rules for continuous probability distributions. *Manag. Sci.* **22**, 1087–1096 (1976)
- J.S. Risbey, M.J. Pook, P.C. McIntosh, M.C. Wheeler, H.H. Hendon, On the remote drivers of rainfall variability in Australia. *Mon. Weather Rev.* **137**, 3233–3253 (2009)
- D. Robertson, D. Shrestha, Q. Wang, Post processing rainfall forecasts from numerical weather prediction models for short term streamflow forecasting. *Hydrol. Earth Syst. Sci. Discuss.* **10**, 6765–6806 (2013)
- R.T. Rust, D.C. Schmittlein, A Bayesian cross-validated likelihood method for comparing alternative specifications of quantitative models. *Market. Sci.* **4**(1), 20–40 (1985)
- N.H. Saji, B.N. Goswami, P.N. Vinayachandran, T. Yamagata, A dipole mode in the tropical Indian Ocean. *Nature* **401**, 360–363 (1999)
- A. Schepen, Q.J. Wang, Ensemble forecasts of monthly catchment rainfall out to long lead times by post-processing coupled general circulation model output. *J. Hydrol.* **519**, 2920–2931 (2014)
- A. Schepen, Q.J. Wang, Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia. *Water Resour. Res.* **51**(3), 1797–1812 (2015)
- A. Schepen, Q.J. Wang, D. Robertson, Evidence for using lagged climate indices to forecast Australian seasonal rainfall. *J. Climate* **25**, 1230–1246 (2012)
- Q. Shao, M. Li, An improved statistical analogue downscaling procedure for seasonal precipitation forecast. *Stoch. Env. Res. Risk Assess.* **27**(4), 819–830 (2013)
- T. Shinohaki, S. Furui, T. Kawahara, Gaussian mixture optimization based on efficient cross-validation. *IEEE J. Sel. Top. Sign. Process* **4**, 540–547 (2010)
- P. Smyth, Clustering using Monte Carlo cross-validation. KDD, 126–133 (1996). <http://www.aaai.org/Papers/KDD/1996/KDD96-021.pdf>
- P. Smyth, Model selection for probabilistic clustering using cross-validated likelihood. *Stat. Comput.* **10**(1), 63–72 (2000)
- M. Stone, An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *J. R. Stat. Soc. Ser. B Methodol.* **44**–47 (1977)
- M. Thyer, G. Kuczera, Q.J. Wang, Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation. *J. Hydrol.* **265**, 246–257 (2002)
- M. Thyer, B. Renard, D. Kavetski, G. Kuczera, S.W. Franks, S. Srikanthan, Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. *Water Resour. Res.* **45**, W00B14 (2009)
- B. Timbal, D. Jones, Future projections of winter rainfall in southeast Australia using a statistical downscaling technique. *Clim. Change* **86**, 165–187 (2008)
- N.K. Tuteja, D. Shin, R. Laugesen, U. Khan, Q. Shao, E. Wang, M. Li et al., Experimental evaluation of the dynamic seasonal streamflow forecasting approach (2012). Report published

- by the Bureau of Meteorology, Melbourne. Available online http://www.bom.gov.au/water/about/publications/document/dynamic_seasonal_streamflow_forecasting.pdf
- D.C. Verdon, S.W. Franks, Indian Ocean sea surface temperature variability and winter rainfall: Eastern Australia. *Water Resour. Res.* **41**, W09413 (2005). doi:10.1029/2004WR003845.
- Q.J. Wang, D.E. Robertson, Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.* **47**, W02546 (2011)
- Q.J. Wang, D.E. Robertson, F.H.S. Chiew, A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.* **45**, W05407 (2009), <https://doi.org/10.1029/2008WR007355>
- G. Wang et al., POAMA-2 SST skill assessment and beyond. *CAWCR Res. Lett.* **6**, 40–46 (2011a)
- Q. Wang, T. Pagano, S. Zhou, H. Hapuarachchi, L. Zhang, D. Robertson, Monthly versus daily water balance models in simulating monthly runoff. *J. Hydrol.* **404**, 166–175 (2011b)
- Q. Wang, D. Shrestha, D. Robertson, P. Pokhrel, A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.* **48**, W05514 (2012a)
- Q.J. Wang, A. Schepen, D.E. Robertson, Merging seasonal rainfall forecasts from multiple statistical models through Bayesian model averaging. *J. Climate* **25**, 5524–5537 (2012b)
- I.K. Yeo, R.A. Johnson, A new family of power transformations to improve normality or symmetry. *Biometrika* **87**, 954–959 (2000)
- D. Zhu, A.O. Hero, Bayesian hierarchical model for large-scale covariance matrix estimation. *J. Comput. Biol.* **14**, 1311–1326 (2007)
- Z. Zivkovic, F. van der Heijden, Recursive unsupervised learning of finite mixture models. *IEEE Trans. Pattern Anal. Mach. Intell.* **26**, 651–656 (2004)



Multi-model Combination and Seamless Prediction

Stephan Hemri

Contents

1	Introduction	286
2	Raw and Post-processed Ensemble Forecasts	288
2.1	Raw Ensemble Forecasts	288
2.2	Post-processing of Multi-model Ensemble Forecasts	290
3	Seamless Prediction Methods	294
3.1	Generation of Seamless Hydrologic Forecasts	294
3.2	How to Account for Spatiotemporal Dependence Structures?	295
4	Probabilistic Single-Model Forecasts	298
4.1	Ensemble Kalman Filter	299
4.2	Bayesian Forecasting System	299
5	Verification	301
5.1	Univariate Verification	301
5.2	Multivariate Verification	302
6	Summary	303
7	Conclusion	304
	References	304

Abstract

(Hydro-) Meteorological predictions are inherently uncertain. Forecasters are trying to estimate and to ultimately also reduce predictive uncertainty. Atmospheric ensemble prediction systems (EPS) provide forecast ensembles that give a first idea of forecast uncertainty. Combining EPS forecasts, issued by different weather services, to multi-model ensembles gives an even better understanding of forecast uncertainty. This article reviews state of the art statistical post-processing methods that allow for sound combinations of multi-model ensemble forecasts. The aim of statistical post-processing is to maximize the sharpness of the

S. Hemri (✉)

Department of Computational Statistics, HITS gGmbH, Heidelberg, Germany
e-mail: stephan.hemri@h-its.org

predictive distribution subject to calibration. This article focuses on the well-established parametric approaches: Bayesian model averaging (BMA) and ensemble model output statistics (EMOS). Both are readily available and can be used for straightforward implementation of methods for multi-model ensemble combination. Furthermore, methods to ensure seamless predictions in the context of statistical post-processing are summarized. These methods are mainly based on different types of copula approaches. Since skill of (statistically post-processed) ensemble forecasts is generally assessed using particular verification methods, an overview over such methods to verify probabilistic forecasts is provided as well.

Keywords

Bayesian forecasting system (BFS) · Bayesian model averaging · Bayesian model averaging (BMA) · Box-Cox transformation · Brier score (BS) · Consortium for Small-Scale Modeling (COSMO) ensemble · Development of the European Multi-model Ensemble system for seasonal-to-interannual prediction (DEMIETER) · DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm · Ensemble Kalman filter (EnKF) · Ensemble model output statistics (EMOS) · Ensemble model output statistics (EMOS) method · Ensemble prediction systems (EPS) · European Centre for Medium-Range Weather Forecasts (ECMWF) · European Flood Alert System (EFAS) · Multivariate verification · Numerical weather prediction (NWP) model · Schaake shuffle · Seamless prediction methods · Spatiotemporal dependence structures · U.S. National Centers for Environmental Prediction (NCEP) · Univariate verification · Verification

1 Introduction

Even though (hydro-) meteorological forecasts are inherently uncertain, deterministic forecasts, which completely neglect uncertainty, have been the state of the art over many decades. However, it is crucial to assess predictive uncertainty, i.e., the uncertainty conditional on the available information set and the forecaster's expertise (Krzysztofowicz 1999; Todini 2008). Accordingly, over the last 20 years the paradigm in weather forecasting has shifted to probabilistic forecasting (see e.g., Palmer 2000; Hamill et al. 2000). This led to the development of ensemble prediction systems (EPS) in the early 1990s. An EPS consists of multiple runs of the same numerical weather prediction (NWP) model with different initial conditions and/or model variants. An EPS forecast provides an estimate of the predictive distribution of atmospheric variables like air pressure, temperature, precipitation, or wind speed. Such an ensemble forecast can ideally be interpreted as a random sample from this distribution. Based on the availability of meteorological EPSSs, hydrologists started to develop hydrologic ensemble forecasts (Cloke and Pappenberger 2009). Hydrologic ensemble forecasts are often driven by input from several atmospheric models,

issued by different weather centers, of which each is either deterministic or probabilistic (Thielen et al. 2009; Bartholmes et al. 2009).

EPS forecasts are affected by systematic biases and dispersion errors (i.e., wrong forecast spread). Hence, they typically benefit from (a) the combination of predictions from several independent weather centers as in the poor person's approach (Arribas et al. 2005; Atger 1999; Ebert 2001; Ziehmann 2000) and (b) from statistical post-processing (Glahn and Lowry 1972; Gneiting et al. 2005). The former approach explicitly accounts for model uncertainty, while the latter infers predictive uncertainty using statistical methods. The goal of statistical post-processing is to obtain well calibrated and yet sharp forecasts (Gneiting et al. 2007). Calibration is a joint property of the forecasts and the observations. In a nutshell, the observations look like random samples from the predictive distributions in case of a well-calibrated forecast. Sharpness refers to the spread of the forecasts. Subject to calibration, minimization of forecast spread, i.e., maximizing sharpness, maximizes forecast skill.

This chapter discusses methods for (hydro-) meteorological multi-model combination with a particular focus on statistical post-processing and seamless prediction. For a broader overview of the hydrological challenges in meteorological post-processing refer to the chapter on “► [Hydrological Challenges in Meteorological Post-processing](#)” by F. Wetterhall in this handbook.

The first goal of the chapter at hand is to summarize methods for multi-model combination. These methods comprise primarily poor person's approaches and statistical post-processing methods based on parametric distributions like ensemble model output statistics (EMOS, Gneiting et al. 2005) and Bayesian model averaging (BMA, Raftery et al. 2005). As stated above, poor person's approaches refer to the combination of NWP forecasts from several operational centers in order to obtain an ensemble forecast. Since multi-model combination aims at quantifying the uncertainty of the forecasts, alternative reference methods for uncertainty quantification in hydrology like the ensemble Kalman filter (EnKF, Evensen 1994) and the Bayesian forecasting system (BFS) are summarized as well.

The second goal of this chapter is to summarize statistical methods that ensure seamless predictions. Given that the different forecast models cover the same range of lead times, discontinuities in the marginal predictive distributions can be reduced by parameter smoothing. The more challenging task is to ensure that correlations among different lead times are well represented by the forecasts. Several different methods have been proposed for this purpose. Most of them are related to the concept of copulas (Nelsen 2006; Sklar 1959). The most prominent method to incorporate correlation structures into (hydro-) meteorological forecasts is the Schaake shuffle (Clark et al. 2004).

The remainder of this chapter is organized as follows. In Sect. 2 a summary on uncertainty quantification by ensemble forecasting is given. Within that context, both the combination of multi-model raw ensembles and post-processing of ensemble forecasts are discussed. Here, a raw ensemble denotes the forecast ensemble prior to any statistical post-processing. Section 3 focuses on seamless prediction

with main focus on spatiotemporal dependence structures. In Sect. 4 alternative uncertainty quantification methods based on deterministic forecasts are discussed. As probabilistic forecasting is inherently connected to probabilistic verification, some verification methods are summarized in Sect. 5. This is followed by a summary in Sect. 6.

2 Raw and Post-processed Ensemble Forecasts

2.1 Raw Ensemble Forecasts

2.1.1 Early Techniques

In the 1990s, EPS methods like the ones implemented by the US National Centers for Environmental Prediction (NCEP) and the European Centre for Medium-Range Weather Forecasts (ECMWF) took only the uncertainty about initial states into consideration, while neglecting model uncertainty (Ziehmann 2000). However, the poor person's ensemble implicitly takes model uncertainty into account. It consists of a set of deterministic NWP forecasts from different forecast centers. In this virtually cost-free approach, the ensemble is constructed from independent operational deterministic forecasts. As there are only a few operational centers that run global atmospheric models, the ensemble size of poor person's ensembles constructed from deterministic NWP forecasts is usually limited to a few members. The combination of EPS forecasts from independent operation centers, which leads to very large ensemble sizes, will be discussed in Sect. 2.1.2. Poor person's ensemble approaches proved to be competitive compared to the NCEP and ECMWF EPS ensemble forecasts (Arribas et al. 2005; Atger 1999; Ebert 2001; Ziehmann 2000). While poor person's ensembles often lack a correct representation of spread, they perform particularly well in terms of resolution. Resolution is a measure of the ability to issue case-dependent forecasts that differ from the climatological forecasts. For details on resolution refer to Hersbach (2000).

The multi-model superensemble (Krishnamurti et al. 1999, 2000) is closely related to the poor person's ensemble. In a nutshell, the superensemble technique fits a multiple regression model with the members of a poor person's ensemble as predictors and the observations as dependent variable. Since this is a statistical approach, parameters have to be estimated over a training period. The coefficients of the model, which can also be understood as weights assigned to the ensemble members, are estimated for each location (geographical and vertical) and each variable separately. In Krishnamurti et al. (1999, 2000) predictions from such a model proved to outperform any of the poor person's ensemble members in terms of root mean squared error. They performed even better than the ensemble mean and the mean of an ensemble consisting of individually bias-corrected ensemble members. The superensemble method is a deterministic approach in that it uses the output from an ensemble forecast, predominantly from a poor person's ensemble, to generate a deterministic forecast with increased skill.

2.1.2 More Recent Developments

As already stated above, the most commonly used technique to account for uncertainty in meteorological forecasting is ensemble prediction. Such ensembles of parallel deterministic forecast runs may ideally be understood as probabilistic forecasts through their empirical cumulative distribution function. The combination of different model runs that have been generated by different uncertainty algorithms (i.e., perturbations in initial states and/or model parameters) typically accounts for the corresponding sources of uncertainty. Additionally, ensemble forecasts issued by different weather services are combined to multi-model ensembles to account for model formulation uncertainty. This is a further development of the poor person's ensemble approach. In other words, a multi-model raw ensemble is a physically based approach to quantify multifaceted sources of uncertainty. Hereby, the uncertainties in initial conditions, parameterizations, model structure, and data assimilation methods of the different meteorological ensembles are considered as well. The most prominent example of such a multi-model ensemble is the THORPEX Interactive Grand Global Ensemble (TIGGE) project ensemble (Bougeault et al. 2010; Park et al. 2008; Richardson et al. 2005). It currently comprises the EPS forecasts from ten different operational centers. The TIGGE ensemble exhibits high forecast skill. Hagedorn et al. (2012) analyzed its performance for 850 hPa temperature and 2-m temperature. A reduced TIGGE ensemble consisting only of the four best EPSSs (provided by Canada, the US, the UK, and ECMWF) showed an improved performance on the global domain compared to the best single-model EPS, the ECMWF EPS. In case of precipitation, the multi-model ensemble consisting of the four best TIGGE EPSSs performed even better compared to the reforecast calibrated ECMWF EPS (Hamill 2012). Furthermore, the results indicated that statistical post-processing of the multi-model ensemble did not provide as much improvement as post-processing of the ECMWF EPS did. From this, Hamill (2012) concluded that "all operational centers, even ECMWF, would benefit from the open, real-time sharing of precipitation data and the use of reforecasts." Multi-model approaches are not restricted to the global domain. For instance, the Grand Limited Area Model Ensemble Prediction System (GLAMEPS) combines four regional EPS forecasts over the European domain (Iversen et al. 2011).

Turning now to river runoff, Cloke and Pappenberger (2009) provide a review on ensemble flood forecasting. For a forecast horizon up to 2–3 days, most forecast uncertainty is related to uncertainty in the meteorological inputs. This source of uncertainty can be addressed in a straightforward manner by using a meteorological ensemble forecast as input to the hydrologic model. This approach can be improved by using a multi-model ensemble of meteorological EPS forecasts from different forecast centers. However, (multi-model) meteorological raw ensembles are prone to systematic biases and spread errors (Park et al. 2008). Furthermore, catchment related hydrologic sources of uncertainty affect river runoff forecasts as well. Palmer et al. (2004) summarize the development of the European Multi-model Ensemble system for seasonal-to-interannual prediction (DEMETER). As part of the DEMETER project, a multi-model ensemble seasonal forecasting system based on seven global atmospheric models has been tested. The results indicate that such a

multi-model combination approach leads to more reliable seasonal-to-interannual predictions. Detailed analyses on the performance of the DEMETER multi-model ensemble can be found in Hagedorn et al. (2005), Doblas-Reyes et al. (2005), and Weinheimer et al. (2005).

Such multi-model raw ensemble forecasts are used operationally by different flood warning services. The European Flood Alert System (EFAS) incorporates two deterministic models, i.e., the high-resolution run of the ECMWF and the deterministic model of the German weather service (DWD), and two ensemble models, i.e., the 51 member ECMWF ensemble and the 16 member Consortium for Small-Scale Modeling (COSMO) ensemble (Bartholmes et al. 2009; Thielen et al. 2009; European Flood Awareness System 2014). On a regional scale, there are a lot of similar flood alert systems. For instance, an operational hydrologic ensemble prediction system based on the COSMO ensemble and the COSMO-7 deterministic model by MeteoSwiss is run routinely for river Sihl (Addor et al. 2011).

The first study that quantifies the uncertainty in hydrologic model structure by multi-model combination has been performed by Georgakakos et al. (2004). They constructed a multi-model ensemble consisting of ensemble members stemming from both calibrated and uncalibrated deterministic hydrologic models. That ensemble performed quite well in terms of reliability and its mean performed better than the best single model in terms of quadratic error, which is in line with the results from the studies on the poor person's ensemble. Zappa et al. (2011) present a framework to investigate the relative contributions of meteorological inputs, initial conditions, and hydrologic model parameter estimates to the total predictive uncertainty. To this end, a large multi-model ensemble is constructed. It consists of any permutation of the meteorological raw ensemble members, of the weather radar precipitation field ensemble members, which account for uncertainty in initial conditions, and of an ensemble of equifinal parameter sets.

The above mentioned multi-model approaches constitute a major improvement in quantifying predictive uncertainty of hydrologic forecasts. Nevertheless, forecasts based on raw ensembles may still tend to be biased and exhibit spread errors (mostly underdispersion). These problems demand for more detailed assessments of reliability and, if indicated, statistical post-processing. Studies addressing reliability benefit strongly from the availability of reforecasts and adequate statistical post-processing methods (Schaake et al. 2010; Thielen et al. 2008). Such post-processing methods, which are applicable to ensemble forecasts, are discussed next.

2.2 Post-processing of Multi-model Ensemble Forecasts

State of the art techniques for univariate, i.e., each lead-time and each location is considered independently, probabilistic multi-model combination and simultaneous statistical post-processing include Bayesian model averaging (BMA) developed by Raftery et al. (2005), and the ensemble model output statistics (EMOS) method introduced by Gneiting et al. (2005). An illustrative example of both methods is given in Fig. 1. Alongside a general introduction to BMA and EMOS, some

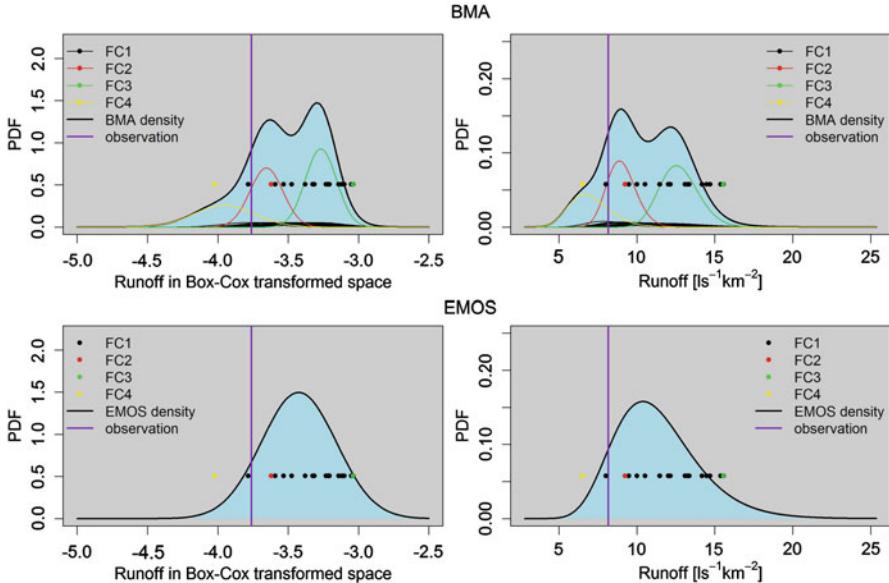


Fig. 1 Examples of BMA (upper panels) and EMOS (lower panels) predictive probability density functions (PDFs). The panels on the *left* show the PDFs in the Box-Cox transformed space, those in the panels on the *right* are in the original space. FC1 to FC4 refer to different forecast models, of which FC1 is an ensemble of size 16 and the others are deterministic. The *horizontally aligned dots* show the values of the raw ensemble members

specialties of hydrologic applications are discussed in the following. Note that the description of univariate post-processing methods closely follows Gneiting (2014).

2.2.1 Bayesian Model Averaging

For a variable of interest, y , the BMA approach employs a mixture distribution of the general form

$$y \mid x_1, \dots, x_M \sim \sum_{m=1}^M w_m g(y \mid x_m), \quad (1)$$

where $y \mid x_1, \dots, x_M$ denotes the distribution of the variable y conditional on the ensemble forecast members x_1, \dots, x_M . Here, $g(y \mid x_m)$ is a parametric density function that depends on the specific ensemble member forecast x_m in suitable ways, and the mixture weights w_1, \dots, w_M are nonnegative and sum to 1. The mixture weights w_1, \dots, w_M reflect the corresponding members' relative contributions to predictive skill over the training period.

As river runoff generally follows a highly skewed distribution, it is recommended to either transform the variables such that they are approximately normal or to model the distributions $g(\cdot)$ using more appropriate parametric (or nonparametric) density

functions. The former approach (see also Duan et al. 2007) may be employed using the Box-Cox transformation (Box and Cox 1964)

$$z = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0, \\ \log(x) & \text{if } \lambda = 0, \end{cases} \quad (2)$$

where z is the Box-Cox transformed value of the original value $x > 0$ and λ is the parameter of the Box-Cox transformation. This approach works well in practice, though it does not ensure normality. Alternatively, the quantile-quantile transformation

$$z = \Phi^{-1}(F(x)), \quad (3)$$

can be used to ensure normality (see e.g., Bogner et al. 2012). $F(x)$ is the value of an empirical cumulative distribution function (CDF) F at the point x and Φ^{-1} is the inverse CDF of the standard normal distribution.

Assuming that $g(y | x_m)$ is a normal density, where the mean is a bias-corrected affine function of x_m and the variance is fixed, is reasonable in the case of temperature and pressure (see e.g., Gneiting et al. 2005). For Box-Cox transformed runoff it is reasonable as well (Duan et al. 2007). For normally distributed variables, Raftery et al. (2005) propose the BMA specification

$$y | x_1, \dots, x_M \sim \sum_{m=1}^M w_m \mathcal{N}(a_{0m} + a_{1m}x_m, \sigma_m^2), \quad (4)$$

so that the kernel densities $g(y | x_m)$ are Gaussian with mean $a_{0m} + a_{1m}x_m$ and variance σ_m^2 . The BMA weights w_1, \dots, w_M , the bias parameters a_{01}, \dots, a_{0M} and $a_{11}, \dots, a_{1M} \geq 0$, and the variance parameters $\sigma_1^2, \dots, \sigma_M^2$ are fitted on training data in ways described by Raftery et al. (2005). This mainly involves maximum-likelihood estimation. Note that functions for the estimation of BMA models with Gaussian- and Gamma-distributed kernel distributions are readily available in the R package `ensembleBMA` (Fraley et al. 2015).

As stated in Sect. 2.1.2, probabilistic river runoff forecasts are usually based on multi-model atmospheric input ensembles. The members of a particular meteorological center's ensemble are in general exchangeable, that is, they are physically indistinguishable. Hence, the corresponding hydrologic ensemble members are exchangeable as well. Fraley et al. (2010) discuss the adaptation of the basic Gaussian BMA specification (Eq. 4) to ensembles with exchangeable members. Their model is given by

$$y | x_{11}, \dots, x_{1K_1}, \dots, x_{M1}, \dots, x_{MK_M} \sim \sum_{m=1}^M w_m \sum_{k=1}^{K_m} \mathcal{N}(a_{0m} + a_{1m}x_{mk}, \sigma_m^2), \quad (5)$$

where the members x_{m1}, \dots, x_{mK_m} are the exchangeable members of model m . K_m is the ensemble size of model m and equals one for a deterministic model, while, for instance, for the 50 member ECMWF ensemble, it would equal 50. In cases with only deterministic models, the model sizes are $K_m = 1$ for all $m = 1, \dots, M$, and, hence, K_m is omitted.

More recently, flexible BMA approaches that do not depend on parametric kernel distributions have been proposed. Parrish et al. (2012) introduced a BMA algorithm that is based on a particle filter approach. The kernel distributions are obtained by a sequential Monte Carlo simulation method. This allows for multimodal kernels, which may reflect the true uncertainty better than any parametric distribution. Rings et al. (2012) provide an alternative flexible BMA method that is also based on a Monte Carlo method, in which the kernel densities are estimated using the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al. 2008, 2009).

2.2.2 Ensemble Model Output Statistics

Gneiting et al. (2005) introduced ensemble model output statistics (EMOS), also named as nonhomogeneous regression (NR), which models the predictive distribution as a single parametric distribution of the general form:

$$y | x_1, \dots, x_M \sim g(y | x_1, \dots, x_M), \quad (6)$$

where the variables are defined as above for BMA. Assuming a Gaussian density to be appropriate for the variable to be forecast, the EMOS predictive distribution is

$$y | x_1, \dots, x_M \sim \mathcal{N}(\mu, \sigma^2), \quad (7)$$

where $\mu = a_0 + a_1 x_1 + \dots + a_M x_M$ and $\sigma^2 = b_0 + b_1 s^2$, with

$$s^2 = \frac{1}{M} \sum_{m=1}^M \left(x_m - \frac{1}{M} \sum_{m=1}^M x_m \right)^2 \quad (8)$$

denoting the variance of the raw ensemble.

In case of exchangeable members, the mean parameter μ is given by $\mu = a_0 + a_1 \bar{x}_1 + \dots + a_M \bar{x}_M$, where $\bar{x}_1, \dots, \bar{x}_M$ are the means of each set of exchangeable ensemble members (i.e., the members stemming from the same EPS) with $\bar{x}_m = \frac{1}{K_m} \sum_{k=1}^{K_m} x_{mk}$. The ensemble variance is then

$$s^2 = \frac{1}{\sum_{m=1}^M K_m} \sum_{m=1}^M \sum_{k=1}^{K_m} (x_{mk} - \bar{x}_m). \quad (9)$$

The coefficients $a_0, a_1, \dots, a_M, b_0, b_1$ are estimated by numerical optimization over a training period. That is, a scoring function, which depends on the model coefficients and the observations, is minimized. Usually, the continuous ranked probability score (CRPS, Matheson and Winkler 1976; Hersbach 2000) is well suited for that purpose. More details on the CRPS can be found in Sect. 5 about verification. For EMOS based on a Gaussian distribution, functions are available in the R package `ensembleMOS` (Yuen et al. 2013). Depending on the variable to be forecast, other, non-Gaussian distribution functions may be used instead.

The EMOS model can alternatively be formulated as an extended logistic regression (ExtLR) model (Wilks 2009). In order to obtain a complete predictive distribution, ExtLR uses the threshold to be forecast, y , as an additional predictor. The ExtLR model can then be written as

$$F(y) = \frac{\exp(a_0 + a_1 x_1^\alpha + \dots + a_K x_K^\alpha + b y^\beta)}{1 + \exp(a_0 + a_1 x_1^\alpha + \dots + a_K x_K^\alpha + b y^\beta)} \text{ for } y \geq 0, \quad (10)$$

where $\alpha > 0$ and $\beta > 0$ are fixed coefficients and $x \in \mathbb{R}^K$ is the vector of predictors. Fundel and Zappa (2011) apply ExtLR to hydrological reforecasts following Wilks (2011) who uses the ensemble mean \bar{x} and spread s as predictors. This leads to the ExtLR model

$$F(y) = \frac{\exp(a_0 + a_1 \bar{x}^{\alpha_1} + a_2 s^{\alpha_2} + b y^\beta)}{1 + \exp(a_0 + a_1 \bar{x}^{\alpha_1} + a_2 s^{\alpha_2} + b y^\beta)}, \quad (11)$$

where $\alpha_1, \alpha_2, \beta$ have to be determined based on the forecasters knowledge, and a_0, a_1, a_2, b are estimated, for instance, by maximum-likelihood estimation. A more recent development of ExtLR allows for interactions between the vector of predictors, x , and the threshold, y (Ben Bouallègue 2013).

3 Seamless Prediction Methods

3.1 Generation of Seamless Hydrologic Forecasts

Seamless predictions, i.e., consistent predictions over successive lead times, are of increasing importance in the field of (hydro-) meteorological forecasting. For instance, Palmer et al. (2008) motivate the use of seamless predictions by the verification of climate models. Based on the premise that the fundamental physical processes of seasonal forecasts and decadal climate projections are similar, probabilistic climate forecasts can be calibrated according to the validation results of the seasonal predictions of the corresponding models. In meteorology, a seamless prediction system is designed to cover the time span from weather to climate predictions. However, in hydrology, seamless predictions span a somewhat shorter

time horizon from nowcasting flash floods to seasonal drought predictions (Yuan et al. 2014). Short range hydrologic forecasts may benefit from a blending of precipitation nowcasts and forecasts. Kober et al. (2012, 2014) and Scheufele et al. (2014) propose and apply such a blending method using a weighting function that depends on lead time and the conditional square root of the ranked probability score. Hydrologic model runs based on seamless meteorological predictions can be expected to be seamless as well.

If hydrologic ensemble forecast trajectories, which have been obtained by statistical post-processing, are used as inputs to a hydrodynamic model or for river routing, it is crucial to avoid discontinuities in the marginal predictive distributions. Naïve approaches smooth the parameter estimates of the univariate model fits. For instance, in case of EMOS, the estimates of the parameters ($a_0, a_1, \dots, a_M, b_0, b_1$) can be smoothed using cubic smoothing splines as implemented in the R function `smooth.spline`, where the smoothing parameter is estimated using leave-one-out cross-validation. For instance, Hemri et al. (2015) apply such a smoothing method. More sophisticated approaches would be based on simultaneous parameter estimation over the entire range of lead times. To the author's knowledge, there are no studies addressing this in the context of hydrological post-processing, though several methods used for spatially adapted post-processing of meteorological forecasts have been developed. Such methods can often be transferred to temporal problems. For instance, the locally adaptive EMOS method (Feldmann et al. 2015; Scheuerer and Büermann 2014) could probably be modified in such a way that simultaneous parameter estimation over the entire range of lead times becomes feasible.

3.2 How to Account for Spatiotemporal Dependence Structures?

Recently, different methods to incorporate multivariate dependence structures into the post-processing of ensemble forecasts have been proposed. Let us first have a look at nonparametric reordering approaches that comprise mainly the Schaake shuffle (Clark et al. 2004) and ensemble copula coupling (ECC, Schefzik et al. 2013). Both approaches implicitly rely on empirical copulas. The notion of a copula is defined by Sklar's theorem (Sklar 1959), which states that any multivariate CDF F with margins F_1, \dots, F_L can be represented by

$$F(y_1, \dots, y_L) = C(F_1(y_1), \dots, F_L(y_L)), \quad (12)$$

where $y_1, \dots, y_L \in \mathbb{R}$ and $C : [0, 1]^L \rightarrow [0, 1]$. The copula C can be understood as a multivariate CDF with standard uniform marginal distributions. In case of ensemble forecasts, the rank order structure of the ensemble members defines an empirical copula over the forecast margins (i.e., particular lead times and locations).

The Schaake shuffle transfers historical spatiotemporal dependence structures to the forecasts of interest. For simplicity, it is assumed here that only one variable is of

interest (e.g., runoff, precipitation, or temperature), though the method could easily be used to model dependence structures between different variables. Additionally, the difference between single model and multi-model ensembles is ignored here. Let $X_{m,t,s}$ be the ensemble forecast array at a specific day. The index $m = 1, \dots, M$ refers to the ensemble members, $t = 1, \dots, T$ to the lead times, and $s = 1, \dots, S$ to the locations. Then a corresponding observation array $Y_{m,t,s}$ of equal size is selected. $Y_{m,t,s}$ is constructed by selecting the same number of historical observations as there are ensemble members. Times of day (to reflect lead times correctly) and locations have to be equal in $X_{m,t,s}$ and in $Y_{m,t,s}$. In Clark et al. (2004) the dates of the historical observations are selected such that they match the calendar day of the forecast of interest by ± 7 days, regardless of the year. The multi-index $\ell = (s, t)$ defines the margins at which univariate, statistically post-processed ensemble forecasts are available. For a given location s and lead time t , i.e., margin ℓ , the Schaake shuffle can be summarized as follows:

1. Sort the forecast vector $X_\ell = (x_1^\ell, \dots, x_M^\ell)$ such that $\tilde{X}_\ell = (\tilde{x}_1^\ell, \dots, \tilde{x}_M^\ell) = (x_{(1)}^\ell, \dots, x_{(M)}^\ell)$, with $x_{(1)}^\ell \leq x_{(2)}^\ell \leq \dots \leq x_{(M)}^\ell$.
2. Sort the observation vector $Y_\ell = (y_1^\ell, \dots, y_M^\ell)$ such that $\tilde{Y}_\ell = (\tilde{y}_1^\ell, \dots, \tilde{y}_M^\ell) = (y_{(1)}^\ell, \dots, y_{(M)}^\ell)$, $y_{(1)}^\ell \leq y_{(2)}^\ell \leq \dots \leq y_{(M)}^\ell$, and denote the corresponding ranks by rk_m^ℓ .
3. Construct the reordered forecast vector $X^{\text{ss}} = (\tilde{x}_{\text{rk}_1^\ell}^\ell, \dots, \tilde{x}_{\text{rk}_M^\ell}^\ell)$.

The above reordering procedure is applied to all margins ℓ . With this, the Schaake shuffle preserves the Spearman rank correlation structure between the margins ℓ . For instance, two positively correlated stations have probably similar historical rank orders and hence also similar forecast rank orders after applying the Schaake shuffle. For hydrologic ensemble forecasts of runoff from a larger catchment, it is crucial that meteorological input variables show an appropriate correlation structure among its subcatchments, because this ensures that forecasts of extremes are not leveled out by averaging over the subcatchments.

The ECC method works similar to the Schaake shuffle, but the ranks for reordering of the forecasts are derived from the corresponding raw ensemble instead of historical observations. For ECC, one first determines the rank order structure of the raw ensemble of size M at each margin ℓ . Then one samples M times from each post-processed marginal predictive CDF \hat{F}_ℓ . These samples are then reordered using the rank order structure of the raw ensemble, which leads to reordered post-processed ensemble trajectories. ECC distinguishes between three different variants. The first variant, ECC-Q, uses equidistant quantiles as samples from \hat{F}_ℓ . The second variant, ECC-R, uses random samples from \hat{F}_ℓ . The third variant, ECC-T, is a quantile mapping or transformation approach, where the quantiles to be calculated from \hat{F}_ℓ are determined by first fitting a parametric

distribution to the raw ensemble and then examining to which quantiles of this distribution the raw ensemble members correspond. Thus, ECC ensures that the Spearman rank correlation coefficient of the raw ensemble is retained (Schefzik et al. 2013). Denoting the raw ensemble at margin ℓ by $R^\ell = (r_1^\ell, \dots, r_M^\ell)$, ECC can be summarized as follows:

1. Obtain a marginal forecast CDF \widehat{F}_ℓ by univariate post-processing.
2. Sort the raw ensemble members $r_{(1)}^\ell \leq \dots \leq r_{(M)}^\ell$ at each margin ℓ . Denote the ranks by rk_m^ℓ .
3. Select one of the following three methods, which have been proposed by Schefzik et al. (2013), to obtain the ECC ensemble members \widehat{y}_m^ℓ at margin ℓ :
 - ECC-Q: $\widehat{y}_m^\ell = \widehat{F}_\ell^{-1}(x_{\text{rk}_m^\ell})$, where $x_{\text{rk}_m^\ell}$ are elements of the vector of equidistant probabilities $\left(\frac{1}{M+1}, \frac{2}{M+1}, \dots, \frac{M}{M+1}\right)$ or $\left(\frac{1/2}{M}, \frac{3/2}{M}, \dots, \frac{M-1/2}{M}\right)$ reordered according to the ranks rk_m^ℓ .
 - ECC-R: $\widehat{y}_m^\ell = \widehat{F}_\ell^{-1}(u_{\text{rk}_m^\ell})$, where $u_{\text{rk}_m^\ell}$ are elements of the vector $(u_{(1)}, \dots, u_{(m)})$ of sorted random samples from a standard uniform random variable reordered according to the ranks rk_m^ℓ .
 - ECC-T: $\widehat{y}_m^\ell = \widehat{F}_\ell^{-1}(\widehat{S}_\ell(r_m^\ell))$, where \widehat{S}_ℓ is the fitted CDF of a suitable parametric distribution to the raw ensemble R^ℓ .

In contrast to the Schaake shuffle, ECC intrinsically adapts to different weather patterns by relying on the current raw ensemble forecast and not on historical data. Obviously, the performance of ECC depends on how well the true dependence structures are represented by the raw ensemble.

Spatiotemporal dependence structures can also be modeled using parametric copulas. Many different copula approaches have been proposed to model such dependence structures in a (hydro-) meteorological context. The review article by Schölzel and Friederichs (2008) provides an introduction to this topic. As most of the applied studies rely on Gaussian copulas, the idea of using parametric copulas is illustrated using Gaussian copulas in the following. Following Schölzel and Friederichs (2008), let $\mathbf{Y} = (Y_1, \dots, Y_L)$ denote the random vector of interest. As before, the marginal forecast CDFs, \widehat{F}_ℓ , are known from prior post-processing. Reconsidering Eq. 12 and assuming that \widehat{F}_ℓ is univariately calibrated, we set $U_\ell = \widehat{F}_\ell(y_\ell) \sim \mathcal{U}(0, 1)$ following a standard uniform distribution. By applying the inverse of the CDF of the standard normal distribution, Φ , one obtains

$$Z_\ell = \Phi^{-1}(U_\ell), \quad (13)$$

which follows a standard normal distribution. Additionally, it is assumed that $\mathbf{Z} = (Z_1, \dots, Z_L)$ follows a multivariate normal distribution with covariance matrix Σ , i.e.,

$\mathbf{Z} \sim \mathcal{N}(0, \Sigma)$, and let the CDF of \mathbf{Z} be denoted by $G(\cdot)$. Then, the CDF of \mathbf{Y} is given by

$$C(u_1, \dots, u_L) = G(Z_1, \dots, Z_L) = G(\Phi^{-1}(U_1), \dots, \Phi^{-1}(U_L)). \quad (14)$$

As an example of the usage of parametric copulas, the Gaussian copula approach by Pinson and Girard (2012) is mentioned here. For this example, only temporal correlations are considered, i.e., the margins ℓ correspond to the lead times $\ell = 1, \dots, L$. This approach follows the procedure described above with \hat{F}_ℓ being the predictive CDF obtained by univariate post-processing of the raw ensemble at lead time ℓ . The covariance matrix Σ , which is equal to a correlation matrix here, is then estimated by fitting a parametric correlation function to the correlations between different lead times in the training period. Using, for instance, an exponential correlation function, Σ is given by

$$\text{cov}(Y_{t,\ell_1}, Y_{t,\ell_2}) = \exp\left(-\frac{|\ell_1 - \ell_2|}{v}\right), \quad 0 < \ell_1, \ell_2 \leq L, \quad (15)$$

where $Y_{t,\ell}$ is the observation at training day t with lead time ℓ . The measurements are assumed to be taken at hourly intervals. The parameter v needs to be estimated from the training data.

An additional advantage of using Gaussian copulas is the straightforward calculation of conditional forecast densities. Note that the joint distribution of runoff values $y_{1:L}$ over the range of lead times $1, \dots, L$ can be written as

$$f(y_{1:L}) = f(y_{1:\ell_1}) \cdot f(y_{\ell_1+1:\ell_2}|y_{1:\ell_1}) \cdots f(y_{\ell_n+1:\ell_L}|y_{1:\ell_n}), \quad (16)$$

where $1 < \ell_1 < \ell_2 < \dots < \ell_n < L$. For instance, Hemri et al. (2013) use this approach to construct seamless hydrologic BMA predictive densities from a raw ensemble, of which ensemble members drop out at particular lead times. Engeland and Steinsland (2014) apply a similar method to model the dependence between neighboring catchments and lead times in the framework of an EMOS forecast based on a raw ensemble of size three, which is obtained by combining the predictions from a hydrologic model, a sliding window climatology, and the persistence forecast.

4 Probabilistic Single-Model Forecasts

In parallel to the development of ensemble forecasting and multi-model combination methods, several approaches that quantify uncertainty on the basis of deterministic hydrologic models have been proposed. These methods are appealing, in that they do not rely on computationally expensive ensemble forecasts.

4.1 Ensemble Kalman Filter

As stated by Vrugt et al. (2005), most classical hydrologic models ignore most of the different sources of uncertainty and represent them by the uncertainty in the parameter estimates only. However, the ensemble Kalman filter (EnKF) provides a framework for the treatment of the different sources of uncertainty in deterministic hydrological models (Evensen 1994). The EnKF is an extension of the Kalman filter (KF). Both are described in detail in Vrugt et al. (2005) and in Vrugt and Robinson (2007). Here, the basic principle of sequential data assimilation via Kalman filtering approaches is presented using the example of the KF. The evolution of the state vector $\Psi_t \in \mathbb{R}^m$ (the index t denotes time) of a nonlinear hydrologic model is given by

$$\Psi_{t+1} = \eta(\Psi_t, \tilde{X}_t, \theta) + q_t, \quad (17)$$

where \tilde{X}_t denotes the input vector, θ is the parameter vector, the function $\eta(\cdot)$ stands for the nonlinear hydrologic model, and q_t denotes a noise term that simulates the errors in the hydrological model formulation. The prediction of the model at time t is then mapped from the model state space to the model output space

$$y_t = H(\Psi_t) + \epsilon_t, \quad \epsilon_t \sim \mathcal{N}(0, \sigma_t^o), \quad (18)$$

where $H(\cdot)$ is the measurement operator, y_t denotes model forecast at time t , and ϵ_t represents the measurement error. In case of a scalar output, such as runoff, $H(\cdot)$ simplifies to the transpose of a vector of size m . At each time step, when an observation measurement \tilde{y}_t becomes available, the forecast states, Ψ_t^f , are updated using

$$\Psi_t^u = \Psi_t^f + K_t (\tilde{y}_t - H(\Psi_t^f)). \quad (19)$$

In Eq. 19, K_t denotes the Kalman gain matrix, which is computed by

$$K_t = \Sigma_t^m H^T [H \Sigma_t^m H^T + \Sigma_t^o]^{-1}, \quad (20)$$

where Σ_t^m is the covariance matrix of q_t and Σ_t^o is the covariance matrix of the observations (simplifies to σ_t^o in case of scalar observations).

The linear model structure of the KF cannot account for the nonlinear, mainly threshold triggered, hydrologic responses. The EnKF resolves this issue by propagating an ensemble of model states $A_t = (\Psi_t^1, \dots, \Psi_t^N) \in \mathbb{R}^{m \times N}$ through the KF. A detailed description can be found in Vrugt et al. (2005).

4.2 Bayesian Forecasting System

The Bayesian forecasting system (BFS) has been suggested as a method to obtain probabilistic forecasts from any deterministic hydrologic model (Krzysztofowicz

1999). It decomposes uncertainty into the dominant uncertainty (i.e., input uncertainty, IU) and all other sources of uncertainty (hydrologic uncertainty, HU). Subsequently, input and hydrologic uncertainty are integrated to a probabilistic forecast. Following the BFS approach, the forecast density of river stages $\mathbf{H} = [(H_{11}, \dots, H_{T1}), \dots, (H_{1L}, \dots, H_{TL})]^T$ (lower case \mathbf{h} denotes a realization of the random vector \mathbf{H} , $n = 1, \dots, T$ is an index for forecast time, and $\ell = 1, \dots, L$ for forecast location) can be written as:

$$\psi(\mathbf{h} | \mathbf{h}_0, \mathbf{y}, \mathbf{u}, \mathbf{v}) = \int_{-\infty}^{\infty} \underbrace{\phi(\mathbf{h} | \mathbf{s}, \mathbf{h}_0, \mathbf{y})}_{HU} \underbrace{\pi(\mathbf{s} | \mathbf{u}, \mathbf{v})}_{IU} d\mathbf{s}, \quad (21)$$

where \mathbf{s} denotes hydrologic model output, \mathbf{h}_0 is the series of observations up to the time of forecast issue t_0 , \mathbf{y} denotes a realization of a state vector \mathbf{Y} that is available at the time of forecast issue, \mathbf{V} is the random vector of precipitation inputs, and \mathbf{U} is the random vector of inputs, whose realization \mathbf{u} comprises all deterministic inputs to the hydrologic model (exogenous variables and internal states, but not model parameters, which are assumed to be fixed here). Here, ϕ and π denote probability density functions. Equation 21 can be rewritten as

$$\psi(\mathbf{h} | \mathbf{h}_0, \mathbf{y}, \mathbf{u}, \mathbf{v}) = \gamma(\mathbf{h}; \mathbf{h}_0, \mathbf{y}, \mathbf{u}, \mathbf{v})g(\mathbf{h} | \mathbf{h}_0), \quad (22)$$

where $g(\mathbf{h} | \mathbf{h}_0)$ denotes the prior density conditional on \mathbf{h}_0 , which is typically modeled by a time series model. The weighting function $\gamma(\mathbf{h}; \mathbf{h}_0, \mathbf{y}, \mathbf{u}, \mathbf{v})$ is given by

$$\gamma(\mathbf{h}; \mathbf{h}_0, \mathbf{y}, \mathbf{u}, \mathbf{v}) = \int_{-\infty}^{\infty} f(\mathbf{s} | \mathbf{h}, \mathbf{y}) \frac{\pi(\mathbf{s} | \mathbf{u}, \mathbf{v})}{\kappa(\mathbf{s} | \mathbf{h}_0, \mathbf{y})} d\mathbf{s}, \quad (23)$$

where $\kappa(\mathbf{s} | \mathbf{h}_0, \mathbf{y})$ denotes the expected density of model output \mathbf{S} for any fixed \mathbf{h}_0 and hydrologic model state \mathbf{y} , that is,

$$\kappa(\mathbf{s} | \mathbf{h}_0, \mathbf{y}) = \int_{-\infty}^{\infty} f(\mathbf{s} | \mathbf{h}, \mathbf{y})g(\mathbf{h} | \mathbf{h}_0)d\mathbf{h}, \quad (24)$$

and $f(\mathbf{s} | \mathbf{h}, \mathbf{y})$ is the likelihood of \mathbf{s} conditional on observation \mathbf{h} and state \mathbf{y} .

BFS provides a rather general Bayesian framework for handling uncertainty when using deterministic hydrologic models. For numerical applications, $\pi(\cdot | \mathbf{u}, \mathbf{v})$ has to be estimated by simulation, models for g and f have to be chosen, and the state vector \mathbf{y} has to be determined by experimentation. The vector \mathbf{y} may include some elements of \mathbf{u} and \mathbf{h}_0 . Besides amending deterministic forecasts by a statement on uncertainty and providing an uncertainty decomposition that is soundly based on Bayesian statistics, BFS has two additional favorable properties. Firstly, it has a self-calibration property. That is, given the inputs \mathbf{u} and \mathbf{v} are calibrated, the BFS outputs

are also calibrated. Secondly, the BFS predictive distribution converges automatically to the prior distribution if the hydrologic model has no predictive capability or the input density is noninformative with regard to the variable to be forecast. This constitutes an effective barrier to systematically poor predictions. A discussion of the usage of such Bayesian frameworks for hydrological uncertainty assessment can be found in the chapter on “*Hydrological Challenges in Meteorological Post-processing*” by F. Wetterhall in this handbook.

5 Verification

5.1 Univariate Verification

As probabilistic forecasting is inherently connected to forecast verification, a short summary on probabilistic verification methods is given here. As stated in Gneiting et al. (2007), probabilistic forecasts should be (statistically) well calibrated and yet sharp. A probabilistic forecast is well calibrated if the predicted probability of any forecast interval and the relative frequency of the realizations to fall into that interval are similar. More technically, a forecast is well calibrated if its probability integral transform (PIT, Rosenblatt 1952) follows a standard uniform distribution. The PIT, z_t , for time t is given by

$$z_t = \int_{-\infty}^{y_t} f(y|r_t) dy, \quad (25)$$

where y denotes the variable of interest (for instance, runoff or precipitation), f is the predictive density, r_t denotes here the information set available for time t , e.g., the raw ensemble forecast, and y_t is the value that materializes, i.e., the observation. Using the PIT values z_t for many realizations over a verification period, a histogram like PIT plot can be drawn (Hamill 2001). As shown in Fig. 2, underdispersion, overdispersion, and biases are detected by such PIT histograms.

Sharpness relates to the concentration of predictive distributions and is evaluated by assessing the distribution of the widths of particular prediction intervals, e.g., the

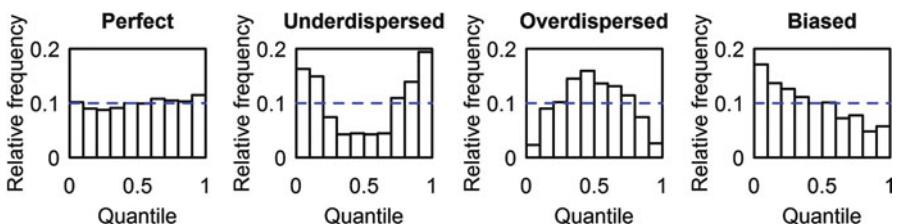


Fig. 2 Left to right: PIT histograms for a well calibrated, an underdispersed, an overdispersed, and a biased forecast (Figures are taken from Hemri et al. (2014))

centered 90% interval, of the forecasts over the verification period. The narrower those intervals the sharper is the forecast. An example of box-plot like diagrams that assess sharpness can be found in Gneiting et al. (2007).

For the verification of probabilistic forecasts, proper scoring rules should be applied as they “encourage the forecaster to make careful assessments and to be honest” (Gneiting and Raftery 2007). The continuous ranked probability score (CRPS, Matheson and Winkler 1976; Hersbach 2000) is such a proper score. It assesses both calibration and sharpness of probabilistic forecasts and hence gives an estimate of predictive skill. The CRPS for a single forecast-observation pair is defined as

$$\text{crps}(F, y) = \int_{-\infty}^{\infty} (F(u) - \mathbb{1}_{\{u \geq y\}})^2 du, \quad (26)$$

where F is the predictive CDF, y the verifying observation, and $\mathbb{1}_{\{A\}}$ the indicator function of the set A . The crps is negatively oriented, i.e., the smaller the score the more skillful is the forecast. In applications, one usually reports the average CRPS over a verification period T , i.e., $\text{CRPS} = \frac{1}{|T|} \sum_{t=1}^T \text{crps}(F_t, y_t)$, where $|T|$ denotes the length of the period T . The associated skill score can be written as

$$\text{CRPSS} = \frac{\text{CRPS}_{ref} - \text{CRPS}}{\text{CRPS}_{ref}}, \quad (27)$$

where CRPS_{ref} is the CRPS of a reference forecast, for instance, a climatological or a persistence forecast. The CRPSS attains values in $(-\infty, 1]$ and is positively oriented. Forecasts with skill equal to the skill of the reference forecast lead to a CRPSS of zero.

If the main interest is the forecast of a dichotomous event (e.g., rain/no rain or exceeding/not exceeding a threshold), the Brier score (BS) is an appropriate verification method (Brier 1950). The BS is given by

$$\text{BS} = \frac{1}{|T|} \sum_{t=1}^{|T|} (F_t(d) - \mathbb{1}_{\{y_t \leq d\}})^2, \quad (28)$$

where F_t is the predictive distribution at verification time t and d is the threshold of interest (e.g., an amount of precipitation or a flood alert level for river runoff). Note that the CRPS is the integral of the BS over the support of F_t .

5.2 Multivariate Verification

In case of issuing multivariate predictions, it is crucial to ensure a realistic correlation structure among the forecast margins (see also Sect. 3.2). To this end Gneiting et al. (2008) proposed the multivariate rank histogram, and Thorarinsdottir et al. (2015) introduced the average rank and the band depth rank histogram. For

forecast vectors of low dimension, the multivariate rank histogram works well, whereas it is recommended to use the average or the band depth rank histograms in case of forecasts of higher dimensions as it is typically the case in the field of (hydro-) meteorological forecasting. Given univariate calibration, these histograms can be used to detect unrealistic correlation structures among lead times and/or locations of the forecast distribution. Refer to the abovementioned literature for more details on these verification tools.

For multivariate assessment of forecast skill, Gneiting and Raftery (2007) proposed the energy score (ES) given by:

$$es(F, \mathbf{x}) = E_F \| \mathbf{X} - \mathbf{x} \| - \frac{1}{2} E_F \| \mathbf{X} - \mathbf{X}' \|, \quad (29)$$

where $\| \cdot \|$ denotes the Euclidian norm, E_F denotes expectation, \mathbf{X} and \mathbf{X}' are independent copies of a random vector following the distribution F , and \mathbf{x} is the corresponding vector of observations. The ES generalizes the CRPS to higher dimensions. Accordingly, ES and CRPS are equal for univariate forecasts. If the forecast is available as an ensemble of size M with ensemble member trajectories $\mathbf{f}_1, \dots, \mathbf{f}_M \in \mathbb{R}^L$, according to Gneiting et al. (2008) the ES can be calculated by:

$$es(F, \mathbf{x}) = \frac{1}{M} \sum_{j=1}^M \| \mathbf{f}_j - \mathbf{x} \| - \frac{1}{2M^2} \sum_{i=1}^M \sum_{j=1}^M \| \mathbf{f}_i - \mathbf{f}_j \| . \quad (30)$$

In case of a deterministic forecast trajectory \mathbf{f} , the ES reduces to the Euclidian norm:

$$es(\mathbf{f}, \mathbf{x}) = \| \mathbf{f} - \mathbf{x} \| . \quad (31)$$

Hence, the ES may be used to compare multivariate density forecasts, discrete ensemble forecasts, and deterministic forecasts (Gneiting et al. 2008).

6 Summary

The most natural combination of multi-model (hydro-) meteorological forecasts is to merge them to a multi-model ensemble consisting of all members of all models. Such multi-model ensembles generally outperform single model ensemble predictions. In order to benefit even more from the concept of multi-model combination, the models should be combined and weighted using state of the art statistical post-processing methods like BMA or EMOS. At present, forecasts for meteorological variables and river runoff benefit strongly from statistical post-processing. Keeping in mind that the methods presented in this chapter are rather simple, more sophisticated post-processing methods may even increase this benefit.

Seamless predictions in a (hydro-) meteorological context should, on the one hand, avoid discontinuities between the marginal forecast distributions from lead time to lead time and, on the other hand, consider the correlation structure among

lead times. While in general, univariate statistical post-processing can take account of the former by straightforward smoothing of parameters, the latter needs more sophisticated approaches. Methods based on empirical copulas like the Schaake shuffle or ECC are very flexible and thus well suited for a wealth of different forecasting problems. They can be applied to both spatial and temporal problems. Alternative methods like the Gaussian copula approach are less flexible, but they in turn provide complete multivariate predictive distributions. This allows, for instance, for straightforward calculation of conditional forecast distributions.

Furthermore, EnKF and BFS, which are alternative uncertainty quantification methods, have been discussed briefly in this chapter. As probabilistic forecasting and verification are inherently connected to each other, some methods for probabilistic verification have been presented from an applied perspective.

7 Conclusion

Probabilistic (hydro-) meteorological forecasting benefits from combining EPS forecasts to multi-model ensembles. In order to further improve forecast performance EPS forecasts can be combined and weighted using statistical post-processing methods. With some modifications and additions such statistical methods can be used to obtain seamless probabilistic predictions.

References

- N. Addor, S. Jaun, F. Fundel, M. Zappa, An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* **15**, 2327–2347 (2011)
- A. Arribas, K.B. Robertson, K.R. Mylne, Test of a poor man’s ensemble prediction system for short-range probability forecasting. *Mon. Weather Rev.* **133**, 1825–1839 (2005)
- F. Atger, The skill of ensemble prediction systems. *Mon. Weather Rev.* **127**, 1941–1953 (1999)
- J.C. Bartholmes, J. Thielen, M.-H. Ramos, S. Gentilini, The European Flood Alert System EFAS – part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.* **13**, 141–153 (2009)
- Z. Ben Bouallègue, Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather Forecast.* **28**, 515–524 (2013)
- K. Bogner, F. Pappenberger, H.L. Cloke, Technical note: the normal quantile transformation and its application in a flood forecasting system. *Hydrol. Earth Syst. Sci.* **16**, 1085–1094 (2012)
- P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D.H. Chen, B. Ebert, M. Fuentes, T.M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y.-Y. Park, D. Parsons, B. Raoult, D. Schuster, P.S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, L. Wilson, S. Worley, The THORPEX Interactive Grand Global Ensemble. *Bull. Am. Meteorol. Soc.* **91**, 1059–1072 (2010)
- G. Box, D. Cox, An analysis of transformations. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **26**, 211–252 (1964)
- G.W. Brier, Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950)

- M. Clark, S. Gangopadhyay, L. Rajagopal, R. Wilby, The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262 (2004)
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009)
- F.J. Doblas-Reyes, R. Hagedorn, T.N. Palmer, The rationale behind the success of multi-model ensembles in seasonal forecasting – II. Calibration and combination. *Tellus A* **57**, 234–252 (2005)
- Q. Duan, N.K. Ajami, X. Gao, S. Sorooshian, Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**(5), 1371–1386 (2007)
- E.E. Ebert, Ability of a poor man's ensemble to predict the probability and distribution of precipitation. *Mon. Weather Rev.* **129**, 2461–2480 (2001)
- K. Engeland, I. Steinsland, Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resour. Res.* **50**, 182–197 (2014)
- European Flood Awareness System, EFAS concepts and tools (2014), <https://www.efas.eu/about-efas.html>. Accessed 24 Apr 2015
- G. Evensen, Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J. Geophys. Res.* **99**(10), 143–162 (1994)
- K. Feldmann, M. Scheuerer, T.L. Thorarinsdottir, Spatial postprocessing of ensemble forecasts for temperature using nonhomogeneous Gaussian regression. *Mon. Weather Rev.* **143**, 955–971 (2015)
- C. Fraley, A.E. Raftery, T. Gneiting, Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Weather Rev.* **138**(1), 190–202 (2010)
- C. Fraley, A.E. Raftery, J.M. Sloughter, T. Gneiting, University of Washington, ensembleBMA: probabilistic forecasting using ensembles and Bayesian model averaging. R package version 5.1.1 (2015), <http://CRAN.R-project.org/package=ensembleBMA>. Accessed 23 Apr 2015
- F. Fundel, M. Zappa, Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. *Water Resour. Res.* **47**, W09520 (2011)
- K.P. Georgakakos, D.-J. Seo, H. Gupta, J. Schaake, M.B. Butts, Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.* **298**, 222–241 (2004)
- H.R. Glahn, D.A. Lowry, The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**, 1203–1211 (1972)
- T. Gneiting, Calibration of medium-range weather forecasts. ECMWF Technical Memorandum, No. 720 (2014), 28p
- T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007)
- T. Gneiting, A.E. Raftery, A.H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**(5), 1098–1118 (2005)
- T. Gneiting, F. Balabdui, A.E. Raftery, Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 243–268 (2007)
- T. Gneiting, L.I. Stanberry, E.P. Grimit, L. Held, N.A. Johnson, Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *Test* **7**(2), 211–235 (2008)
- R. Hagedorn, F.J. Doblas-Reyes, T.N. Palmer, The rationale behind the success of multi-model ensembles in seasonal forecasting – I. Basic concept. *Tellus A* **57**, 219–233 (2005)
- R. Hagedorn, R. Buizza, T.M. Hamill, M. Leutbecher, T.N. Palmer, Comparing TIGGE multimodel forecasts with reforecast-calibrated ECMWF ensemble forecasts. *Q. J. Roy. Meteorol. Soc.* **138** (668), 1814–1827 (2012)
- T.M. Hamill, Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129**, 550–560 (2001)
- T.M. Hamill, Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Weather Rev.* **140**, 2232–2252 (2012)

- T.M. Hamill, C. Snyder, R.E. Morss, A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Weather Rev.* **128**(6), 1835–1851 (2000)
- S. Hemri, F. Fundel, M. Zappa, Simultaneous calibration of ensemble river flow predictions over an entire range of lead-times. *Water Resour. Res.* **49** (2013). <https://doi.org/10.1002/wrcr.20542>
- S. Hemri, D. Lisniak, B. Klein, Ascertainment of probabilistic runoff forecasts considering censored data (in German). *Hydrol. Wasserbewirtsch.* **58**(2), 84–94 (2014). https://doi.org/10.5675/HyWa_2014_2_4
- S. Hemri, D. Lisniak, B. Klein, Multivariate post-processing techniques for probabilistic hydrological forecasting. *Water Resour. Res.* **51**(9), 7436–7451 (2015)
- H. Hersbach, Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**(5), 559–570 (2000)
- T. Iversen, A. Deckmyn, C. Santos, K. Sattler, J.B. Bremnes, H. Feddersen, I.-L. Frogner, Evaluation of ‘GLAMEPS’ – a proposed multimodel EPS for short range forecasting. *Tellus A* **63**, 513–530 (2011)
- K. Kober, G.C. Craig, C. Keil, A. Dörnbrack, Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Q. J. Roy. Meteorol. Soc.* **138**, 755–768 (2012)
- K. Kober, G.C. Craig, C. Keil, Aspects of short-term probabilistic blending in different weather regimes. *Q. J. Roy. Meteorol. Soc.* **140**, 1179–1188 (2014)
- T.N. Krishnamurti, C.M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, S. Surendran, Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**(5433), 1548–1550 (1999)
- T.N. Krishnamurti, C.M. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, S. Gadgil, S. Surendran, Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate* **13** (23), 4196–4216 (2000)
- R. Krzysztofowicz, Bayesian theory of probabilistic forecasting via deterministic hydrologic model. *Water Resour. Res.* **35**(9), 2739–2750 (1999)
- J.E. Matheson, R.L. Winkler, Scoring rules for continuous probability distributions. *Manag. Sci.* **22**, 1087–1096 (1976)
- R. Nelsen, *An Introduction to Copulas* (Springer, New York, 2006)
- T. Palmer, Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.* **63**, 71–116 (2000)
- T.N. Palmer, F.J. Doblas-Reyes, R. Hagedorn, A. Alessandri, S. Gualdi, U. Andersen, H. Feddersen, P. Cantelابе, J.-M. Terres, M. Davey, R. Graham, P. Délécluse, A. Lazar, M. Déqué, J.-F. Guérémy, E. Díez, B. Orfila, M. Hoshen, A.P. Morse, N. Keenlyside, M. Latif, E. Maisonnave, P. Rogel, V. Marletto, M.C. Thomson, Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMIETER). *Bull. Am. Meteorol. Soc.* **85**, 853–872 (2004)
- T.N. Palmer, F.J. Doblas-Reyes, A. Weisheimer, M.J. Rodwell, Toward seamless prediction: calibration of climate change projections using seasonal forecasts. *Bull. Am. Meteorol. Soc.* **89**, 459–470 (2008)
- Y.-Y. Park, R. Buizza, M. Leutbecher, TIGGE: preliminary results on comparing and combining ensembles. *Q. J. Roy. Meteorol. Soc.* **134**, 2029–2050 (2008)
- M.A. Parrish, H. Moradkhani, C.M. DeChant, Toward reduction of model uncertainty: integration of Bayesian model averaging and data assimilation. *Water Resour. Res.* **48**, 1–18 (2012)
- P. Pinson, R. Girard, Evaluating the quality of scenarios of short-term wind power generation. *Appl. Energy* **96**, 12–20 (2012)
- A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**(2), 1155–1174 (2005)
- D.S. Richardson, R. Buizza, R. Hagedorn, TIGGE – first final report. WMO/TD-No. 1273 (2005), http://www.wmo.int/pages/prog/arep/wWRP/new/thorpeX_publications.html. Accessed 24 Apr 2015.
- J. Rings, J.A. Vrugt, G. Schoups, J.A. Husman, H. Vereecken, Bayesian model averaging using particle filtering and Gaussian mixture modeling: theory, concepts, and simulation experiment. *Water Resour. Res.* **48**, W0552 (2012)

- M. Rosenblatt, Remarks on a multivariate transformation. *Ann. Math. Stat.* **23**, 470–472 (1952)
- J. Schaake, J. Pailleux, J. Thielen, R. Arritt, T. Hamill, L. Luo, E. Martin, D. McCollor, F. Pappenberger, Summary of recommendations of the first workshop on postprocessing and downscaling atmospheric forecasts for hydrologic applications held at Météo-France, Toulouse, France, 15–18 June 2009. *Atmos. Sci. Lett.* **11**, 59–63 (2010)
- R. Schefzik, T.L. Thorarinsdottir, T. Gneiting, Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.* **28**, 616–640 (2013)
- M. Scheuerer, L. Büermann, Spatially adaptive post-processing of ensemble forecasts for temperature. *J. R. Stat. Soc. Ser. C* **63**(3), 405–422 (2014)
- K. Scheufele, K. Kober, G.C. Craig, C. Keil, Combining probabilistic precipitation forecasts from a nowcasting technique with a time-lagged ensemble. *Meteorol. Appl.* **21**, 230–240 (2014)
- C. Schölzel, P. Friederichs, Multivariate non-normally distributed random variables in climate research – introduction to the copula approach. *Nonlinear Processes Geophys.* **15**, 761–772 (2008)
- A. Sklar, Fonctions de répartition à n dimensions et leurs marges. *Publ. Inst. Stat. Univ. Paris* **8**, 229–231 (1959)
- J. Thielen, J. Schaake, R. Hartman, R. Buizza, Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. *Atmos. Sci. Lett.* **9**, 29–35 (2008)
- J. Thielen, J.C. Bartholmes, M.-H. Ramos, A. de Roo, The European Flood Alert System – part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125–140 (2009)
- T.L. Thorarinsdottir, M. Scheuerer, C. Heinz, Assessing the calibration of high-dimensional ensemble forecasts using rank histograms. *J. Comput. Graph. Stat.* **25**, 105–122 (2016)
- E. Todini, A model conditional processor to assess predictive uncertainty in flood forecasting. *Int. J. River Basin Manag.* **6**, 123–137 (2008)
- J.A. Vrugt, B.A. Robinson, Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. *Water Resour. Res.* **43**, 1–18 (2007)
- J.A. Vrugt, C.G.H. Diks, H.V. Gupta, W. Bouten, J.M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. *Water Resour. Res.* **41**, W01017 (2005). <https://doi.org/10.1029/2004WR003059>
- J.A. Vrugt, C.J.F. ter Braak, M.P. Clark, J.M. Hyman, B.A. Robinson, Treatment of input uncertainty in hydrologic modeling: doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.* **44**, W00B09 (2008). <https://doi.org/10.1029/2007WR006720>
- J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, B.A. Robinson, J.M. Hyman, D. Higdon, Accelerating Markov chain Monte Carlo simulations by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlinear Sci. Numer. Simul.* **10**(3), 271–288 (2009)
- A. Weinheimer, L.A. Smith, K. Judd, A new view of seasonal forecast skill: bounding boxes from the DEMETER ensemble forecasts. *Tellus A* **57**, 265–279 (2005)
- D.S. Wilks, Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorol. Appl.* **16**(3), 361–368 (2009). <https://doi.org/10.1002/met.134>
- D. Wilks, *Statistical Methods in the Atmospheric Sciences* (Academic Press, Oxford, 2011)
- X. Yuan, E. Wood, M. Liang, Developing a seamless hydrologic forecast system: integrating weather and climate prediction. *Geophysical Research Abstracts* **16**(EGU2014-2268) (2014)
- R. Yuen, T. Gneiting, T. Thorarinsdottir, C. Fraley, ensembleMOS: ensemble model output statistics. R package version 0.7 (2013), <http://CRAN.R-project.org/package=ensembleMOS>. Accessed 23 Apr 2015
- M. Zappa, S. Jaun, U. Germann, A. Walser, F. Fundel, Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmos. Res.* **100**, 246–262 (2011)
- C. Ziehmann, Comparison of a single-model EPS with a multi-model ensemble consisting of a few operational models. *Tellus A* **52**, 280–299 (2000)

Part IV

Hydrological Models



Hydrological Cycles, Models, and Applications to Forecasting

Sharad K. Jain and Vijay P. Singh

Contents

1	Hydrology: An Overview	312
2	Hydrologic Cycle	313
3	Components of Hydrologic Cycle	314
3.1	Atmospheric	314
3.2	Surface Components of Hydrologic Cycle	315
3.3	Infiltration	316
3.4	Ground Water	317
3.5	Overland and Channel Flow	318
3.6	Base Flow	320
3.7	Scales for the Study of Hydrologic Cycle	321
3.8	Mathematical Representation of the Hydrologic Cycle	322
3.9	Influence of Human Activities and Land Use Change on Hydrologic Cycle	323
3.10	Impact of Climate Change on Hydrologic Cycle	324
4	Hydrological Modeling	324
4.1	Types of Hydrological Models	325
4.2	Deterministic Models	326
4.3	Black Box or Empirical Models	327
4.4	Lumped Conceptual Models	327
4.5	Fully Distributed, Physically Based Models	328
4.6	Advantages and Limitations of Physically Based Distributed Models	328
4.7	Statistical Models	330
5	History of Hydrologic Modeling	330
6	Integrated Modeling of Hydrologic Cycle	332
6.1	Model Calibration	333

S. K. Jain (✉)

Jal Vigyan Bhawan, National Institute of Hydrology, Roorkee, Uttarakhand, India
e-mail: skj.nihr@gov.in; s_k_jain@yahoo.com

V. P. Singh

Department of Biological and Agricultural Engineering and Zachry Department of Civil
Engineering, Texas A and M University, College Station, TX, USA
e-mail: vsingh@tamu.edu

6.2 Selection of Appropriate Model Type	334
6.3 Uncertainty in Hydrologic Modeling	335
7 Emerging Technology for Hydrologic Modeling	335
8 Future Outlook	336
References	337

Abstract

This chapter presents an overview of hydrology, water cycle, land surface processes (e.g., precipitation, snow, glaciers and frozen soils, evapotranspiration, surface and subsurface runoff, overland and river flow routing), and hydrologic modeling and its history. The chapter is concluded with an outlook for future.

Keywords

Hydrologic cycle · Watershed · Catchment · Models · Precipitation · Evapotranspiration · Surface water · Ground Water · Climate change · Black-box · Conceptual · Distributed · Calibration · Validation · Uncertainty · Data · Remote sensing · GIS

1 Hydrology: An Overview

All life on Earth is dependent, one way or another, on water. Hydrology can be defined as the science that deals with space-time characteristics of the quantity and quality of the waters of the Earth, encompassing their occurrence, movement, distribution, circulation, storage, development, and management. These characteristics are determined by the relation of water to the Earth. This definition of hydrology is not unique, but may suffice to indicate its scope.

Customarily, hydrology is partitioned into surface-water hydrology and ground-water hydrology. Surface-water hydrology is confined to the relation between water and the surface of the Earth. Groundwater hydrology deals with the relation between water and the lithosphere or the subsurface portion of the Earth. Between these two partitions is subsurface hydrology, often called vadose or unsaturated zone hydrology.

The definition of hydrology encompasses some aspects of a multitude of disciplines involving agriculture, biology, chemistry, geography, geology, glaciology, meteorology, oceanography, and physics. The involvement of hydrology with these sciences comes about due to the close association of water with the atmosphere and the Earth. Many branches of hydrology, therefore, have been distinguished. This association also points out that hydrology is an interdisciplinary science that touches almost all aspects of life. Frequently, hydrology is thought of as an element of agriculture, engineering hydraulics, forestry, geography, or geology. Present sociopolitical culture requires an environmental assessment of all changes in the natural relation of water to the surface of the Earth. Therefore, hydrology should be perceived in terms of the entire reaction of water with the environment.

2 Hydrologic Cycle

The *Hydrologic Cycle*, also known as the water cycle, is a fundamental concept in hydrology and is among a number of cycles operating in nature, such as the carbon cycle, the nitrogen cycle, and other biogeochemical cycles. The National Research Council (NRC 1982) defines the hydrologic cycle as “the pathway of water as it moves in its various phases to the atmosphere, to the Earth, over and through the land, to the ocean and back to the atmosphere.” This cycle has no beginning or end and water is present in the cycle in all the three states, viz., solid, liquid, and gas. It is necessary to study the hydrologic cycle, because water is essential for the survival of life and is an important input in many economic activities. But the needed quantity of water of the desired quality may not be available. A pictorial representation of the hydrologic cycle is given in Fig. 1.

The hydrologic cycle (Oki and Kanae 2006) considers the processes of motion, distribution, and storage of the Earth’s waters. It connects the atmosphere and two storages of the Earth system: the oceans and the landsphere (lithosphere and pedosphere). The water that is evaporated from the Earth and the oceans enters the atmosphere. From the atmosphere, water falls on the Earth and the oceans by precipitation. Oceans also receive streamflow and ground water flow from the landsphere. Water leaves oceans only through evaporation.

In hydrologic cycle, at some point in each phase, usually there is: (a) transport of water, (b) temporary storage, and (c) change of state. For example, in the atmospheric phase, there occurs vapor flow, vapor storage in the atmosphere, and condensation or formation of precipitation by change from vapor to the liquid or solid state. In the atmosphere, water is present in the vapor form, while it is mostly liquid in the oceans.

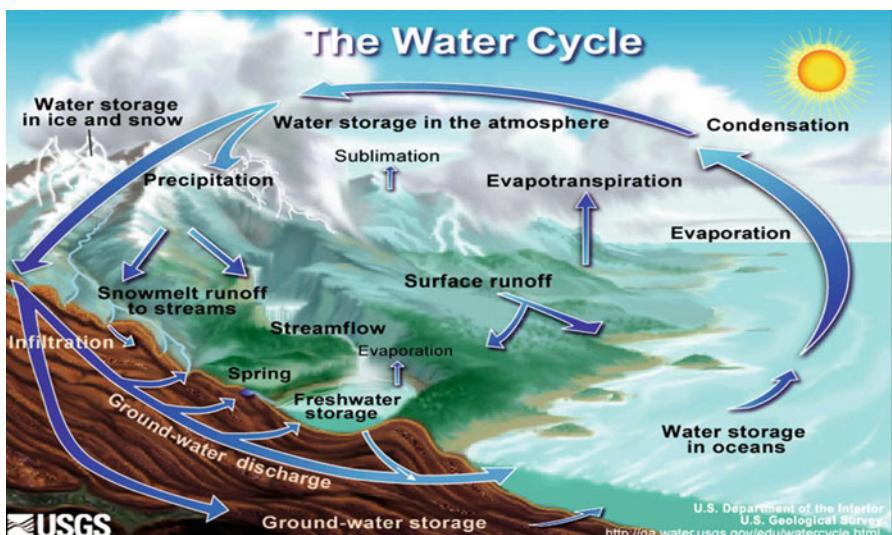


Fig. 1 Pictorial representation of hydrologic cycle (Source: <http://water.usgs.gov>, accessed on 12/6/2014)

Three major subsystems of the hydrologic cycle are readily identified. The atmosphere functions as the storehouse, carrier, and deliverer of water in the moisture form; the land is the user of water where it is also stored and the oceans are the biggest reservoir and source of water. Water availability at a particular place changes with time because of changes in the supply and consumption.

The landsphere receives water through precipitation. Water leaves land area through evapotranspiration (ET), streamflow, interflow, and ground water flow. ET and precipitation are the processes that take place in the vertical plane, while streamflow, interflow, and ground water flow occur mostly in the horizontal plane.

Shiklomanov (1999) called the exchange of water among the oceans, land, and the atmosphere as “the turnover.” Besides, water is a good solvent and hence geochemistry is an integral part of the hydrologic cycle, since water mixes with many chemicals and consequently its quality changes. The hydrologic cycle is, thus, the integrating process for the fluxes of water, energy, and chemical elements (NRC 1991).

The hydrologic cycle can also be visualized as a perpetual distillation and pumping system. In this endless circulation of water, the glaciers and snow packs are replenished, the quantity of river water is replenished, and its quality restored. From the point of view of utilization of water, the land phase of the hydrologic cycle is the most important.

3 Components of Hydrologic Cycle

The hydrologic cycle can be divided into the following major components: precipitation (rainfall, snowfall, hail, sleet, fog, dew, drizzle, etc.), interception, depression storage, evaporation, transpiration, infiltration, percolation, moisture storage in the unsaturated zone, and runoff (surface, interflow, and baseflow).

Water evaporates from the oceans and the land surface mainly due to solar energy. Therefore, sun is the prime mover of the hydrologic cycle. The moisture moves in the atmosphere in the form of water vapor which precipitates on land or oceans in the form of rain, snow, hail, sleet, etc. Part of the precipitation falling on land is intercepted by vegetation or buildings. Of the amount reaching the land, a part infiltrates into the soil and the remaining water runs off the land surface to join streams. Most streams finally discharge into the ocean. Some of the infiltrated water percolates deep to join groundwater. Depending upon the topography and geology, some of the percolated water returns to the streams or emerges out as springs.

A substantial quantity of moisture is added to the atmosphere by transpiration of water from vegetation. Living beings also supply water vapor to the atmosphere through perspiration. Gravity moves water on the Earth surface from high to low elevation; anthropogenic activities also play a role in the movement of water.

We now briefly describe the various components of hydrologic cycle.

3.1 Atmospheric

Precipitation is the most important atmospheric component of the hydrologic cycle.

3.1.1 Precipitation

Precipitation is received on the land surface in the form of rain, snow, hail, frost, and dew. Out of these, rainfall is the predominant component and primarily responsible for streamflow generation or floods in most natural rivers. In many places, rainfall is usually synonymous to precipitation. Rainfall is perhaps the most important and primary input to most hydrological models that are employed for planning, design, and operation of water resource projects. The pattern and magnitude of precipitation depend on the climatic factors, such as temperature, radiation, pressure, humidity, and wind speed. Temporal and spatial variation of these factors makes rainfall a function of both time and space.

Several conditions must be satisfied for the precipitation to occur: the atmosphere should contain moisture and a mechanism should be present to cool it. The cooled moisture should be able to condense. Hence, it must pass through a process of condensation and cloud formation. Since the moist air is lighter than the dry air at a given temperature, it moves upward and gets cooled. For a given amount of moisture, droplets of adequate size will form only in the presence of an optimum number of nuclei. The size of most water droplets in a rainfall event is 0.5–6.0 mm. Larger drops tend to break during fall. Snow is the solid form of precipitation which consists of ice crystals which generally combine to form flakes.

The total amount of precipitation reaching the ground in a stated period is expressed as the depth covering a horizontal projection of the given area in liquid form. In volumetric terms, the total amount of precipitation is the product of the depth and the catchment area. The snowfall is also expressed in terms of equivalent depth of water. The daily amount of precipitation is read to the nearest 0.1 mm.

3.2 Surface Components of Hydrologic Cycle

3.2.1 Interception

Where precipitation does not fall directly on bare soil, it is caught by vegetation or other surface covers and part of it may then be evaporated back to the atmosphere (never reaching the ground). This intercepted amount is known as the interception loss. The remainder of the precipitation eventually reaches the soil but with some delay after temporary storage on the surface cover. The amount of water stored on the wetted surface of the land cover is the interception storage. Interception has the greatest influence during low intensity rainstorms.

The amount of interception depends on the characteristics of precipitation and the form, density, and surface texture of the leaves, stems, or other surfaces, including layering of canopies in the vertical. Dunne and Leopold (1978) note that the total volume of rainfall is the factor used most successfully in the prediction of interception losses. The subtraction of interception loss from gross precipitation makes an insignificant impact during large rainstorms; interception does not affect the development of major floods.

3.2.2 Evaporation

Evaporation is the transfer of water from liquid to vapor state and back to the atmosphere. Evaporation occurs when some water molecules attain sufficient kinetic energy to escape the liquid surface. The rate of evaporation depends on the temperature of the evaporating surface and the ambient air and the difference in vapor pressure between the water surface and the atmosphere; this difference is called the vapor pressure deficit. As evaporation proceeds, the air above the water is gradually saturated and when it is unable to take up any more moisture, evaporation ceases. Since the replacement of saturated air by drier air helps evaporation, wind speed is an important factor in controlling the rate of evaporation. In addition, evaporation from a vegetated surface also depends on soil moisture. Evaporation is one of the most difficult components to quantify in the hydrologic cycle.

3.2.3 Transpiration

Transpiration is the loss of water from the cuticle or the stomatal openings in the leaves of plants. Water is vaporized within the leaf in the intercellular spaces and passes out of stomata by molecular diffusion. The stomata are pores on the undersurface of a leaf which open in sunshine, and when they are open, water vapor can diffuse from wet cells into the atmosphere. This transpired water is replaced by water taken by the roots of the plant from the soil. When computing water loss from a vegetated surface, it is usually impossible to separate transpiration and evaporation from the soil surface, ponds, lakes, and rivers. The term evapotranspiration (ET) represents the two processes together. Thus, ET is the total loss of water by both evaporation and transpiration from a land surface and its vegetation. The amount of ET varies according to the type of vegetation, its ability to transpire, and the availability of water in the soil.

Potential evapotranspiration (PET) is the amount of ET that would take place given an unlimited supply of moisture under the given meteorological conditions. If water is in limited supply at some time during the year, the actual ET may be less than the potential rate. ET is even more difficult to measure than precipitation, partly because this process is not visible.

ET from a reference surface, not short of water, is called the reference ET and is denoted by ET_0 (Allen et al. 1998). The reference surface is a hypothetical grass reference crop with specific characteristics. Reference ET is expressed in the units of depth/time, e.g., mm/day. Crop ET under standard conditions (ET_c) refers to the ET from excellently managed, disease-free, large, well-watered fields that achieve full production under given climatic conditions. To estimate ET_c , ET_0 is multiplied by an empirical crop coefficient which accounts for the difference between the standard surface and the crop. ET can be either measured with a lysimeter, water balance approach, or estimated from climatological data. FAO recommends the use of the Penman-Monteith (PM) method to compute reference ET from a grass surface (Allen et al. 1998).

3.3 Infiltration

From the Earth surface, water seeps into the ground through soil pores. The infiltrated water is useful for plant growth and irrigation demand arises when plants cannot

extract water from the soil pores in the root zone. The water that percolates further down meets the groundwater table and becomes part of the groundwater reservoir.

The rate at which the water enters ground is known as infiltration rate. Field capacity denotes the maximum amount of water that can be stored in the soil against gravitational forces. The permanent wilting point is the lower limit of water available in the soil for the use by plant roots. Thus, the field capacity and permanent wilting point represent the moisture availability under two extreme moisture situations.

The forces of soil water retention are known as matric forces because they result from the soil matrix. The matric suction is a function of soil water content. If the suction (expressed in cm of water column) is plotted on a logarithmic scale against the water content, the resulting curve is called a moisture retention or ψ - θ curve, where ψ refers to the suction head and θ is soil moisture content. The soil water retention property signifies the water storing capacity of soils. Whether water transmission actually takes place through soil pores depends on the property known as hydraulic conductivity or permeability.

3.4 Ground Water

The term “ground water” denotes subsurface water that exists at pressures greater than or equal to atmospheric pressure. Pressures of subsurface water in the capillary fringe and above are below atmospheric pressure and typically capillary water is not considered as ground water.

A geologic stratum that has porosity and hydraulic conductivity to store and transmit significant quantities of water is called an aquifer. Materials with sufficient porosity to store water but a very small capacity to transmit it are called aquiclude, e.g., clays and shales. Aquitard refers to a geologic material, whose hydraulic conductivity is too small to permit the development of wells or springs. Aquifers serve two main functions: They store water for varying periods in the underground reservoirs and also act as pathways to pass water. Some aquifers are more efficient as pathways (e.g., cavernous limestones) and some are more effective as storage reservoirs (e.g., sandstones); most aquifers perform both functions.

Water table is that surface in the groundwater body at which the water pressure is atmospheric. Aquifers may be classified as unconfined or confined, depending on the presence or absence of water table. For an unconfined aquifer, the water table serves as the upper surface of the zone of saturation. The water table in an unconfined aquifer is in contact with the atmosphere through pores in the unsaturated soil. Such aquifers are sometimes called water table aquifers. When a well is drilled in an unconfined aquifer, water will nearly remain at the level where it is first encountered. In a confined aquifer water is under pressure greater than the atmospheric. The upper boundary of a confined aquifer is an impermeable formation that “confines” water in the aquifer, separating it from the atmosphere. An imaginary surface passing through all points to which water will rise in wells penetrating a confined aquifer is called the piezometric surface. When water is first encountered during drilling in a confined aquifer, water will rise in the well and stand at a level above the top of the aquifer. Depending on local conditions, water in a

well tapping a confined aquifer may rise until it flows at the surface without pumping. Such a well is called an artesian well; confined aquifers are often called artesian aquifers.

The two most important aquifer parameters are transmissivity (T) and storage coefficient (S). Transmissivity can be defined as the rate of flow through a cross-section of unit width over the whole thickness of the aquifer under unit hydraulic gradient. It is the product of the average hydraulic conductivity and the thickness of the aquifer. Its common units are m^2/day or m^2/hr . The storage coefficient and the specific yield are defined as the volume of water released and stored per unit surface area of the aquifer per unit change in the component of head normal to that surface. The storage coefficient refers only to the confined parts of an aquifer and depends on the elasticity of the aquifer material and the fluid. It has an order of magnitude of 10^{-4} to 10^{-6} . The specific yield (S_y) refers to the unconfined parts of an aquifer.

3.5 Overland and Channel Flow

Based on the path taken by water, streamflow may be divided into surface flow, interflow, and base flow. Figure 2 shows components of a typical hydrograph. A number of conceptual models are available to describe runoff generation in catchments.

Overland flow frequently occurs as a saturation excess mechanism. All other things remaining the same, soil tends to saturate first where the antecedent soil moisture deficit is the smallest. This will be in valley bottom areas, where flow converges and slopes gradually decline towards the stream. Saturation rapidly occurs where soils are thin or have low permeability. The areas of saturated soil expand with increased wetting as rains continue and reduce after rainfall stops. This concept is called the dynamic contributing

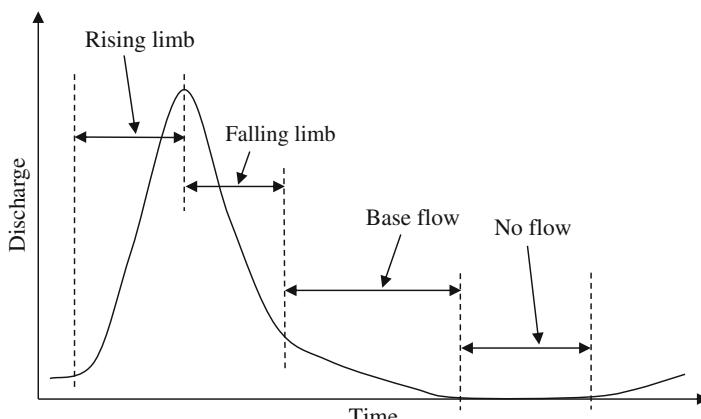


Fig. 2 Components of a streamflow hydrograph

area concept. In addition to contribution from rainfall, surface runoff from such a saturated area may also be due to the return flow of subsurface water (Fig. 3).

A similar concept may be applicable in areas whose responses are controlled by subsurface flows. When saturation starts to build up at the base of soil over a relatively impermeable bedrock, water will start to flow downslope. The connectivity of saturation in the subsurface is, however, important initially. It may be necessary to satisfy some initial bedrock depression storage before there is a consistent flow downslope. The dominant flow pathways may be localized, at least initially, related to variations in

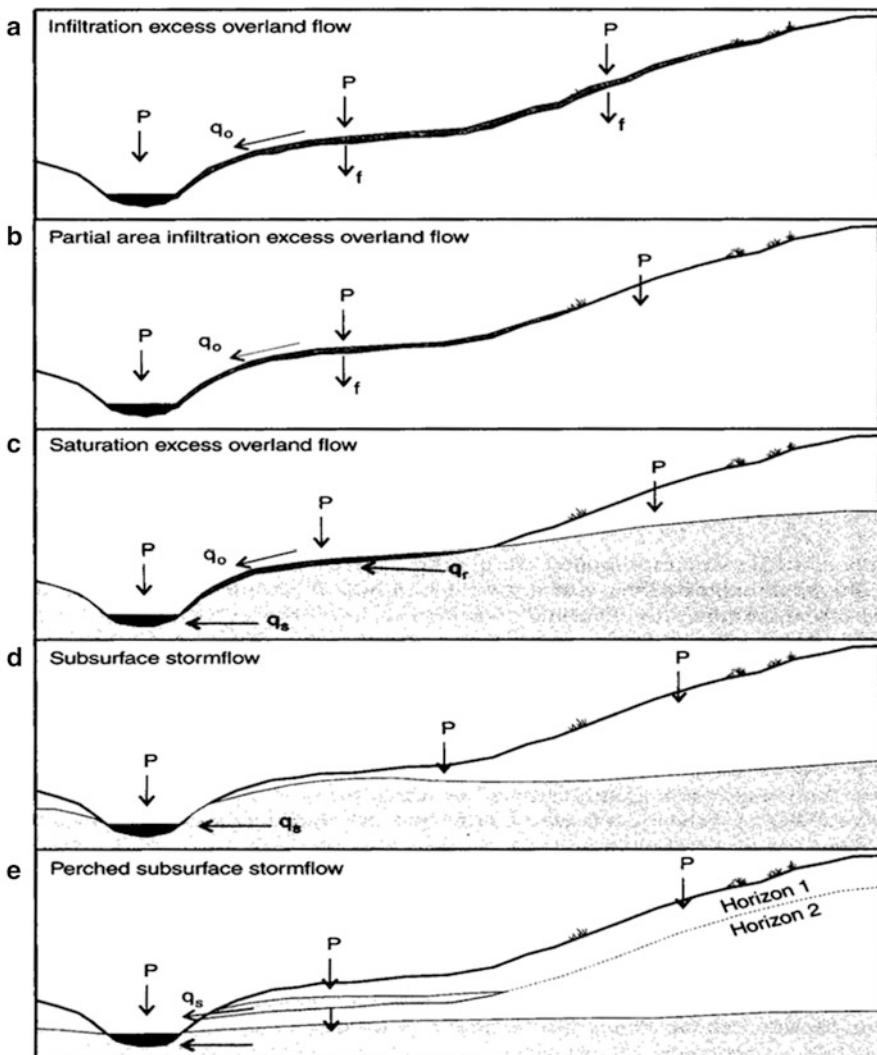


Fig. 3 Various hillslope runoff mechanisms (Source: Beven 2001)

the form of bedrock surface. In the catchments whose soils are deep and have high infiltration capacities, responses may be dominated by subsurface stormflow.

Traditionally, it has been usual to differentiate between different conceptualizations of catchment response based on the dominance of one set of processes over another. An example is the Hortonian model in which runoff is generated by an infiltration excess mechanism all over the hillslope (Fig. 3a). Many forested catchments have deep soils with high infiltration capacities. Response of these catchments during storms is often controlled by subsurface processes and surface runoff is restricted mainly to the channels (Fig. 3d).

Betson (1964) hypothesized that only a part of a catchment is likely to produce runoff in any storm. Since infiltration capacities decrease with increasing soil moisture and the downslope flow of water on hillslopes tends to result in wetter soils at the bottom of hillslopes, the area of surface runoff would tend to start near the channel and expand upslope. This partial area model (Fig. 3b) allowed for a generalization of the Horton conceptualization. It is now realized that the variation in overland flow velocities and the heterogeneities of soil characteristics and infiltration rates are important in controlling partial area responses. If runoff generated on one part of a slope flows onto an area of higher infiltration capacity further down-slope, it will infiltrate (the run-on process). When the high intensity rainfall producing overland flow is of short duration, it is also likely that water will infiltrate before it reaches the nearest channel.

3.6 Base Flow

ASCE (1996) defined base flow as the runoff that has reached the stream or river by passing first through the underlying aquifer, rather than by flowing directly on the ground surface. Thus, base flow is that portion of streamflow that is naturally and gradually withdrawn from groundwater storage or other delayed sources. The other names of base flow are groundwater flow, seepage flow, low flow, and fair weather flow.

Base flow contribution to streamflow varies widely, according to the geologic nature of the water-table aquifer. The lateral movement of groundwater is slower than vertical movement because the hydraulic gradient is smaller for lateral movement. The supply from groundwater to the channel will continue as long as the necessary gradient is present. If there is no additional infiltration to aquifer, the hydraulic gradient decreases as water moves to the stream from higher elevations, then lesser water will travel to the stream with time. This process is called base flow recession.

Perennial streams depend on base flow for discharge between runoff producing events. The presence of base flow around the year indicates humid climate and a shallow water table that is hydraulically connected with the stream. Base flow is absent in (semi)arid climates and areas of deep groundwater. Base flow depends on precipitation, the geologic conditions, and the hydrogeologic controls governing groundwater movement. Climate influences recession through recharge and ET.

3.7 Scales for the Study of Hydrologic Cycle

Depending on the purpose of study, the hydrologic cycle is studied over a range of spatial and temporal scales. Regarding space, two scales are readily distinct: the global scale and the catchment scale. From a global perspective, the hydrologic cycle can be considered to be comprised of three major systems: the oceans, the atmosphere, and the landsphere. Precipitation, runoff, and evaporation are the principal processes that transmit water from one system to the other. The study at the global scale helps understand the global fluxes and global circulation patterns. Results of these studies form important inputs for water resources management at national, regional, and local scales, weather/flow forecasting, and study of impacts of climate change.

While studying the hydrologic cycle on a catchment scale, the spatial coverage can range from a few hectares to thousands of square km. The timescale can be a short duration storm to a study spanning many years. For the water movement in the Earth system, three systems can be recognized: the land (surface) system, the subsurface system, and the aquifer (or geologic) system. In the hydrologic cycle of the land system, the dominant processes are precipitation, evapotranspiration, infiltration, and surface runoff. These subsystems subtract water from precipitation through interception, depression, and detention storage. The exchange of water among these subsystems takes place through the processes of infiltration, exfiltration, percolation, and capillary rise. Fig. 4 shows the schematic of the hydrologic cycle at global scale, in the Earth system, and microscale view of the cycle in the land system.

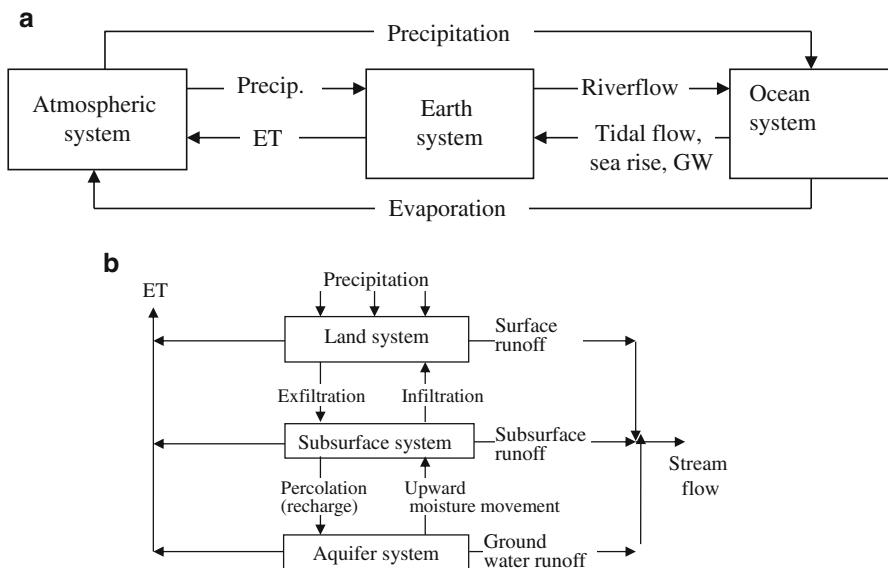


Fig. 4 (a) A global schematic of the hydrologic cycle (Source: Singh 1992). (b) A schematic of the hydrologic cycle of the Earth system (Source: Singh 1992)

The time required for the movement of water through various components of the hydrologic cycle varies considerably. Streamflow moves with much higher velocity compared to ground water. The time-step size for an analysis depends upon the purpose of study and the availability of data. The time step should be sufficiently small so that variations can be captured in required detail, but it should not be a burden on data collection and computational effort.

3.8 Mathematical Representation of the Hydrologic Cycle

The quantities of water going through the various components of the hydrologic cycle can be evaluated by the water balance equation which is a spatially lumped continuity or water budget equation:

$$I - Q = \Delta S \quad (1)$$

where I and Q are the inflow and outflow of water to the study area during any given time period, and ΔS is the change in storage of water in the given area during the time period. If I and Q vary continuously with time, then Eq. (1) can be written as

$$d[S(t)]/dt = I(t) - Q(t) \quad (\text{Uttarakhand}) \quad (2)$$

Integration of this equation yields

$$\begin{aligned} \int dS(t) &= \int [I(t) - Q(t)] dt \\ S(t) - S(0) &= \int_0^t I(t) dt - \int_0^t Q(t) dt = V_I(t) - V_0(t) \end{aligned} \quad (3)$$

where $S(0)$ is the initial storage at time $t = 0$, $S(t)$ is the storage at time t , $V_0(t)$ and $V_I(t)$ are the volumes of outflow and inflow at time t . Each of the terms of this lumped equation is the result of a number of other terms which can be subdivided and even eliminated from the equation, depending upon the temporal and spatial scales of the study. For a watershed, Eq. (1) may be written as

$$P + Q_{SI} + Q_{GI} - E - Q_{SO} - Q_{GO} - \Delta S - \varepsilon = 0 \quad (4)$$

where P is the precipitation, Q_{SI} is the surface inflow, Q_{GI} is the ground water inflow, E is the evaporation from the watershed, Q_{SO} is the surface water outflow, Q_{GO} is the ground water outflow, and ΔS is the change in the water storage in the watershed. For large watersheds, Q_{GI} and Q_{GO} are usually negligible. The discrepancy term ε is included, because the sum of all other terms may not be zero due to measurement errors and/or simplifying assumptions. However, a small value of ε does not necessarily mean that all other terms have been correctly measured/estimated. Finally, the

components of the hydrologic equation may be expressed in terms of the mean depth of water (mm), or as a volume of water (m^3), or in the form of flow rates (m^3/s or mm/s).

The hydrologic equation may be applied to any area, but the complexity of computation greatly depends on the size of the area under study. The smaller is the area, the more complicated is its water balance.

3.9 Influence of Human Activities and Land Use Change on Hydrologic Cycle

A host of factors influence the hydrologic regimes, and it is important to detect changes in the hydrologic cycle by separating natural variability from the variability and trends caused by other reasons. Natural hydrologic regimes at most places have been highly modified by increasing withdrawals and land use changes. These changes can both accelerate (e.g., by urbanization) and dampen (e.g., through afforestation) hydrologic responses. The hydrologic cycle is also modified by human intervention (dams, diversions, interbasin transfers), and application of river or ground water for irrigation and its return flows (Chen et al. 2016). While climate change is influencing the hydrologic cycle, other bio-geochemical cycles, energy generation, water supply and demand for irrigation, drinking, and the quality of water, its signals are difficult to detect and isolate. Already there are noticeable changes in many regions of the world in the key climate parameters, such as temperature and rainfall. However, the climate change signal in derived hydrologic variables, such as river runoff and ground water, is weak or not yet detectable in many parts of the world.

Most watershed changes can be distinguished as point changes or nonpoint changes. Structural changes, such as dam construction, channel improvement, and detention storage, are examples of point changes and affect watershed response in terms of evaporation, seepage, residence/travel time, etc. Afforestation, agriculture, mining, and urbanization are nonpoint changes that affect catchment response. A qualitative discussion of the hydrologic consequences due to watershed changes is given next.

Agricultural changes typically imply that a forested or a barren land is put to cultivation. As a result, the vegetal cover changes, the slope may be altered a little bit, and artificial bunds may be placed causing changes in water retention and infiltration. The effect on the hydrologic regime is pronounced and may be multiplicative. Large amounts of water may be withdrawn from the aquifer or canal irrigation may be introduced leading to noticeable changes in the water table behavior. The changes are also observed in evapotranspiration, overland flow, channel flow, and infiltration. Fertilizer, pesticide, and insecticide applications affect the quality of runoff from agriculture areas.

A land area under forest or agriculture might be transformed into an urban area, where houses, roads, parks, parking lots, sewers, etc., are constructed. A large increase in the paved (impervious) surface considerably reduces infiltration and the removal of storm water is accelerated. Urban development usually increases the

volume and peak of direct runoff, but the time of travel of water is reduced. Thus, the hydrologic effects of urbanization are: (a) increased water withdrawals from surface and subsurface sources; (b) increased peak flow and diminishing baseflow of streams; (c) reduced infiltration; (d) increased pollution of rivers and aquifers, endangering the ecology; and (e) changes in local microclimate.

3.10 Impact of Climate Change on Hydrologic Cycle

Increased emission of green-house gases is believed to be the cause of gradual increase in Earth's temperature. Global warming is likely to lead to higher evapotranspiration; changes in precipitation pattern, timing, and distribution; melting of polar ice caps; and recession of glaciers. Higher melting of polar ice and glaciers will cause sea water level rise and inundation of islands of low elevations as well as coastal cities. Most climate scientists agree that climate warming will intensify, accelerate, or enhance the global hydrologic cycle. Enhancement could be caused by increasing rates of evaporation, ET, precipitation, and streamflow. There are likely to be associated changes in atmospheric water content, soil moisture, ocean salinity, and glacier ice contents.

Given here are the broad impacts of global warming on the various components of the hydrologic cycle. The mode of precipitation is as important as the magnitude in determining hydrologic impacts, and precipitation variability at multidecadal scales can mask long-term trends. Increases in heavy precipitation events have been observed in some places where total precipitation has decreased. In addition, more precipitation now falls as rain rather than snow in northern regions. These changes are expected in a warmer atmosphere with a greater water-holding capacity. Results of reported studies suggest that over large areas of Asia and North America, on average, actual ET is increasing, even though pan evaporation is decreasing.

Worldwide glaciers have retreated since the mid-nineteenth century at varying rates, and this retreat is expected to accelerate on account of global warming and changes in precipitation amount and form. Although there is evidence of glacier retreat globally, all glaciers are not equally sensitive to climate change and there are pockets of anomalous behavior. Studies suggest that the number of days of snow cover is decreasing and snow melt is occurring earlier. Some studies suggest that these changes may have accelerated in the last several decades.

4 Hydrological Modeling

A hydrological model represents the physical/chemical/biological characteristics of the catchment and simulates the natural hydrological processes. A model aids in making decisions, particularly where data are scarce, understanding is incomplete, or there are large numbers of options to choose from (e.g., optimization of reservoir release rules) or it is not possible to experiment with the prototype system. The value of a model is in its ability, when correctly chosen and adjusted, to extract the

maximum amount of information from the available data and answer the question: What if?

4.1 Types of Hydrological Models

Broadly, hydrological models can be divided in two categories: physical (or laboratory) and mathematical (or intellectual). A physical model is a replica of the prototype and is constructed by some physical material, say concrete. These models are not much popular in hydrology. A mathematical model is a quantitative description of the processes or phenomena by using a collection of mathematical equations (often partial differential equations), logical statements, initial and boundary conditions, expressing relationships between input and output.

Commonly, the aim is to model the interactions of inputs (e.g., climate) with the system (e.g., a catchment) to produce an output (e.g., the outflow hydrograph) (Fig. 5). The mathematical functions employed in a model simulate the natural hydrological processes by using the available knowledge, mathematical constraints, data availability, and user requirements. Depending on the accuracy requirement, skills, funds, efforts needed for data collection and modeling, the natural system is represented in greater or smaller details.

The structure and architecture of a hydrologic model are determined by the objective for which the model is built. Hydrologic models can be classified in different ways but not all models fit in a given classification. A general classification of models is shown in Fig. 6. In a different classification, the models can be divided into the deterministic and the stochastic groups. These two groups can each be further divided into conceptual and empirical. Further subdivisions could be spatially lumped/or spatially distributed and linear or nonlinear models.

Singh (1995) classified hydrologic models based on (1) process description, (2) timescale, (3) space scale, (4) techniques of solution, (5) land use, and

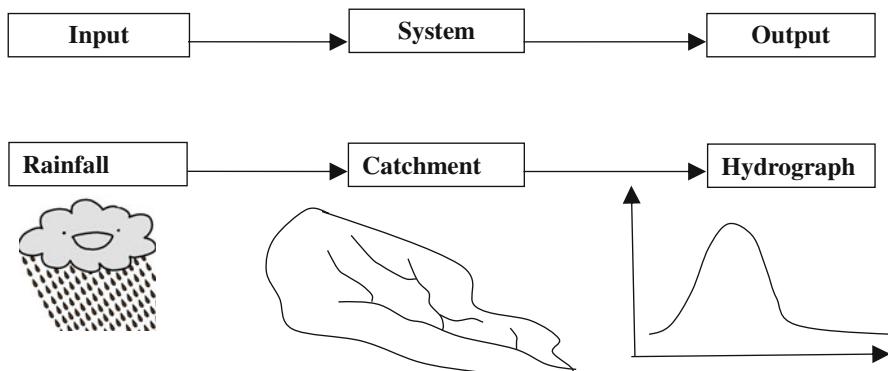


Fig. 5 Representation of input, system, and output for a mathematical model

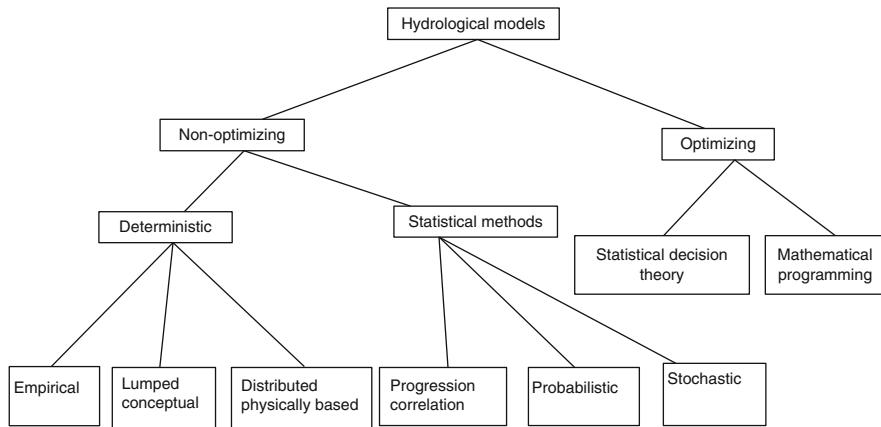


Fig. 6 Classification of hydrologic models

(6) model use. ASCE (1996) reviewed and categorized flood analysis models into (1) event-based precipitation-runoff models, (2) continuous precipitation-runoff models, (3) steady flow routing models, (4) unsteady-flow flood routing models, (5) reservoir regulation models, and (6) flood frequency analysis models.

Two main groups of hydrologic models are: deterministic and stochastic and further discussion follows along these lines.

4.2 Deterministic Models

Deterministic models can be classified according to whether the model describes the catchment as a spatially lumped or distributed system, and whether the description of the hydrological processes is empirical, conceptual, or based on physical laws. In practice, most conceptual models are (semi)lumped and most fully physically based models are distributed, so three main groups of deterministic models are identified in Fig. 6. Over the last many decades, model developments have followed a progression from black box models to grey box models (lumped, conceptual) and to increasingly sophisticated physically based, distributed models (white box). This progression has been supported by four factors: (a) improved understanding of the physics of hydrological processes; (b) increasing quantity, coverage, and quality of hydrological data collected by better sensors and satellite systems; (c) exponential advancements in computational technology; and (d) need for better forecasts and values of more variables in decision making.

Demands to address the increasingly complex problems arising from unwise and unsustainable water resources development, the impacts of landuse land-cover changes, increasing pollution of sources of water, and problems arising due to climate change are some of the reasons for increasingly sophisticated modeling.

4.3 Black Box or Empirical Models

Such models usually utilize relationship between input and output and parameters are calibrated from observed hydrometeorological records. A well-known black box model is the unit hydrograph model. Within the range of calibration data, empirical models may be highly successful because the mathematics of the model is backed with an implicit understanding of the physical system. However, extrapolation beyond the range of calibration is not advisable, since the implicit understanding may no longer be valid. Moreover, many black-box models are linear, while the real-world hydrological systems are nonlinear, which may make such extrapolation of dubious worth. The black box models cannot be employed for some practical problems, e.g., to predict the effects of land-use change on hydrologic response.

Black box models were in widespread use before advances in computer technology enabled the use of more physically correct models. These days black box models often form components of a larger model, e.g., the unit hydrograph is often used for streamflow routing in conceptual rainfall-runoff models.

4.4 Lumped Conceptual Models

These models occupy an intermediate position between fully physically based and black box models. Lumped conceptual models consist of a small number of components, each of which is a simplified representation of an element in the hydrologic system. Typically, each component of the model consists of a nonlinear reservoir in which the relationship between outflow (Q) and storage (S) is given by

$$Q_i = f(K, S^n) \quad (5)$$

where K and n are constants, to be calibrated from existing records. The model operation is normally a bookkeeping system which accounts for the movement of moisture in various storages at each time step. Nonlinearities in the behavior of the real system arising mainly in determining excess rainfall, soil moisture movement, and surface/subsurface runoff are taken care of by thresholds of different storages. Calibration of the lumped conceptual models is more a curve-fitting exercise which means that these models may not work well beyond the range of calibration data.

An example of the lumped conceptual models is the tank model developed by Sugawara (1967). It is a simple model which has proved to be effective in a range of studies. The HBV model, described in detail by Bergstrom (1976), was developed at the Swedish Meteorological and Hydrological Institute and comparable to the tank model. It has also been applied to many catchments and it is used operationally for forecasting floods and reservoir inflows at several hydro-power systems in Sweden and Norway.

4.5 Fully Distributed, Physically Based Models

These models are based on understanding of the physics of the processes which control catchment response; physics-based equations are used to describe the catchment processes. In these models, the transfers of mass, momentum, and energy are calculated by solving the governing partial differential equations, for example, the Saint Venant equations for surface flow, the Richards equation for unsaturated zone, and the Boussinesq equation for ground water flow. Usually these equations are solved by using numerical methods. By definition, physically based models are spatially distributed, since the underlying governing equations generally involve one or more space coordinates. These models simulate the spatial variation in hydrological conditions in a catchment and can give value of the variables, e.g., river flow, soil moisture, and actual ET, at any location in a catchment. However, all these features come at a cost. Such models are costly to develop, apply, and have huge computational time and data requirements.

A strong argument to use distributed models in hydrology has been that these models may be more realistic than the simpler models. Physically based distributed models treating a single component of the hydrological cycle have been developed and extensively applied since the late 1970s. For example, most of the groundwater models are of this type. However, physically based distributed catchment models which integrate submodels of the major components of the hydrological cycle came much later, largely because computer and data requirements of such models were quite high compared to the situation about 25 years ago. Further, there were numerical difficulties, such as the stability of numerical schemes, and mass balance errors which were to be overcome in modeling. Gradually, these difficulties were overcome and several physically based distributed models were developed and tested on small basins during the 1980s. Prominent among these is the SHE modeling system (Fig. 7) (Abbott et al. 1986), developed jointly by the Danish Hydraulic Institute, the Institute of Hydrology (UK), and SOGREAH (France). Initially, these models were applied on small well-instrumented basins, and these applications helped debug the models and make them ready for real-life problems. The SHE model was tested on catchments in a variety of environments and at scales ranging from tens of hectares to nearly 1000 km² (Jain et al. 1992).

4.6 Advantages and Limitations of Physically Based Distributed Models

The concept behind the physically based distributed models has advantages as well as limitations. While the black box models should not be used for the range of input data which is beyond the range of calibration data, physically based models can, in principle, be applied to any set of data subject to the range over which the underlying physical laws are valid. Black box models must be calibrated for each catchment because their parameters do not have any physical meaning, and therefore, these

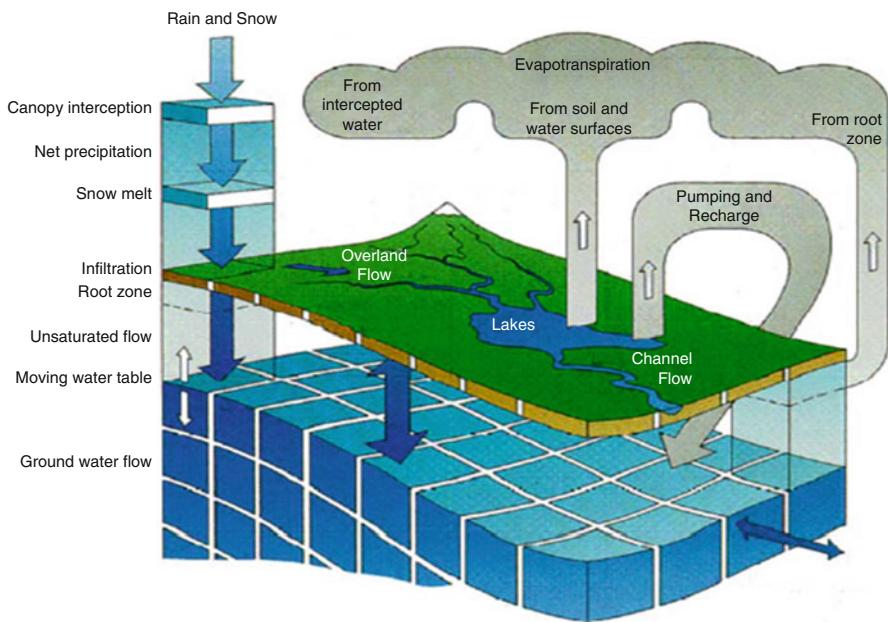


Fig. 7 Schematic diagram of a catchment and a quasi three-dimensional physically based distributed model: the SHE model (Source: <http://mikebydhi.com>)

cannot be derived from measurements of catchment characteristics. Hydro-meteorological records of sufficient length are required for calibration. In contrast, physically based models have parameters with a physical meaning and can in principle be evaluated from direct measurements. This allows such models to be applied to catchments without long data records and to the future state (changing land use) of catchments. In principle, physically based models should not require calibration. In practice, because of model approximations and simplifications, some calibration is required, but this can be carried out on the basis of a short data record.

A given black-box model may not apply to all catchments because the various hydrological processes are not accounted for separately in these models and hence the changes in their relative impacts cannot be easily allowed. Physically based models are applicable to a much wider range of catchments because the physical laws describing the hydrological processes are the same everywhere. Physically based models can use all available information (e.g., topography, soil and vegetation maps, understanding of soil physics and plant physiology, historical information on extreme event characteristics) and may evolve continuously as new insights into hydrological processes are developed.

Limitations

Computer and data requirements to run a physically based distributed catchment model are rather heavy. These include large storage capacity and high processing

speed, because calculations are repeated at each time step for a number of (grid) points. Short time steps may be necessary for stability of numerical schemes during periods of rapid changes. Some of these issues are not very critical these days because of steep decline in hardware costs. But the users are now adopting finer spatial scales, which means some of the advantages of faster processing speed are lost.

With distributed modes, there are problems parameterizing subgrid scale processes (Beven 2001). Theoretical understanding of hydrological processes is not always sufficient or mathematically tractable and amenable. Beven (2001) states that the problem of nonlinearity is at the heart of many problems faced in applying distributed models in hydrology.

4.7 Statistical Models

While most of the deterministic models rely on physics-based approach, statistical methods usually involve functional relationships between hydrological properties of various measured data. Statistical methods in hydrology have been developed extensively with support from basic statistical theory developed and applied in other fields.

5 History of Hydrologic Modeling

Hydrological modeling has a long history which can be traced back to the nineteenth century. The rational method was developed by Mulvany in 1850 and is a clear exposition of the concept of time of concentration and its relation to the maximum runoff (Todini 2007). Sherman (1932) introduced the unit hydrograph (UH) concept to relate the direct runoff from a catchment to rainfall excess. UH is the direct surface runoff hydrograph resulting from an excess rainfall of unit duration (e.g., 1 h) falling uniformly over a catchment. A storm rainfall can be divided into several unit durations and their responses combined to yield the storm hydrograph. In fact, much of the nonlinearity of the rainfall runoff process is taken care of while determining the excess rainfall. Another form of UH is used in hydrology. The impulse response of a linear system is represented by the Instantaneous Unit Hydrograph (IUH). An IUH is obtained when the unit duration of the rainfall excess is infinitesimally small. IUH has the advantage that the assumption of uniform rainfall during the unit duration is avoided. At about the same time when UH idea was developed, Horton (1931) developed a theory of infiltration to estimate rainfall excess. In 1945, Horton developed a concept of erosion and streamflow generation dominated by overland flow. This pioneering work presented a set of empirical laws, known as Horton's laws, which constituted the foundation of quantitative geomorphology.

Using simplified principles of physics, Green and Ampt in 1911 developed a theory of infiltration. Their formula is still popular for computing the infiltration capacity rate. In 1948, Thornthwaite and Penman made important contributions to models of evapotranspiration.

In 1956, the Soil Conservation Service (SCS) of the US Department of Agriculture (now called as the Natural Resources Conservation Service, NRCS) developed a method for computing the amount of storm runoff taking into account the abstractions. These abstractions depend on the land use, soil type, and antecedent soil moisture content which is specified by “Curve Number (CN).” Although originally intended to model daily runoff as affected by land use practices, the SCS-CN method has been very widely used to model infiltration as well as runoff hydrograph for continuous hydrologic simulation either as a standalone model or as a part of detailed hydrologic models, such as the SWAT model (Neitsch et al. 2002).

The subsurface phase of the hydrologic cycle was investigated by Theis (1935) who combined Darcy's law with the continuity equation to derive the relation between the lowering of the piezometric surface and the rate and duration of discharge of a well. Work by Theis laid the foundation of quantitative groundwater hydrology. The study of groundwater and infiltration led to the development of techniques for separation of baseflow and interflow in a hydrograph.

After an interregnum of nearly a quarter century, a major effort in the area of rainfall-runoff modeling employed the theory of linear systems which led to the theory of the instantaneous unit hydrograph by Nash (1957) and then the generalized unit hydrograph theory by Dooge (1959). In 1955, Lighthill and Whitham developed kinematic wave theory for flow routing in long rivers which now has become a main stay in watershed runoff modeling (Singh 1996, 1997). Nash (1957) visualized a catchment as a cascade of N linear reservoirs, each with a residence time of K units.

With the advent of computers, models of different components of the hydrologic cycle were integrated to simulate an entire watershed. A pioneering watershed model was the Stanford Watershed Model-SWM (now HSPF) developed by Crawford and Linsley (1966). SWM was probably the first comprehensive attempt to model the entire hydrologic cycle. At around the same time, a number of other watershed models were developed and applied to diverse problems of hydrologic design. Examples are the models by Dawdy and O'Donnell (1965), HEC-1 by US Army Corps of Engineers in 1968, NWS River Forecast System (Burnash et al. 1973), and the SSARR (Rockwood 1982). During the sixties, a number of conceptual models which represented the various watershed processes through storages or tanks were developed, e.g., the Tank Models developed by Sugawara (1967) and Sugawara et al. (1974). Backed by a large number of applications, many versions of HEC-1 model were brought out. With time, it migrated from mainframe to desktop computers (or PCs). It has been rechristened as HEC-HMS (Hydrologic Modeling System) (<http://www.hec.usace.army.mil/software/hec-hms/>) and the later versions have the GIS capabilities as well.

6 Integrated Modeling of Hydrologic Cycle

In the 1970s, it was hypothesized that runoff is produced by the basin when the soil moisture content reaches the field capacity. The Xinanjiang model developed by Zhao et al. (1980) was based on this concept. Explicit Soil Moisture Accounting (ESMA) is the name given to the models where a collection of storage elements that represent different processes that are important in controlling the catchment response are employed. Exchange of soil moisture fluxes between these elements is described by mathematical equations. ESMA models differ in the number of storage elements used as well as the functions and parameters describing moisture movement. Todini (1996) developed a new conceptual model which was applied to the Arno River and hence it was called as the ARNO model. Wood et al. (1992) further expanded the concept by including subgrid soil heterogeneity and soil layers in the Variable Infiltration Capacity (VIC) model.

Since the 1980s, there has been a proliferation of watershed hydrology models; the popular ones in the list include the Systeme Hydrologique Europeen (SHE) (Abbott et al. 1986), TOPMODEL (Beven and Kirkby 1979), Soil and Water Assessment Tool (SWAT), and Variable Infiltration Capacity (VIC) model (Liang et al. 1994; Gao et al. 2010). Some of these models have been significantly improved after their first appearance. SHE has been extended to include sediment transport and is applicable at the scale of a river basin (Bathurst et al. 1995) and was later packaged as a commercial suite. TOPMODEL has been extended to contain increased catchment information, more physically based processes, and improved parameter estimation.

Singh (1995) edited a book that summarized 26 popular models. Wurbs (1998) listed a number of generalized water resources simulation models in seven categories and discussed their dissemination. Singh and Frevert (2002a, b, 2006) summarized a large number of additional hydrologic models.

Although the mathematical equations embedded in watershed models are continuous in time and often space, analytical solutions cannot be obtained except in very simple circumstances. Numerical methods (finite difference, finite element, boundary element, boundary fitted coordinate) must be used for practical cases. The most general formulation would involve partial differential equations in three space dimensions and time. If the spatial derivatives are ignored, the model is said to be “lumped”; otherwise it is said to be “distributed” and the solution (output) is a function of space and time. Strictly speaking, if a model is truly distributed, then all aspects of the model must be distributed, including parameters, initial and boundary conditions, and sources and sinks. Practical limitations of data and discrete descriptions of watershed geometry and parameters to conform to the numerical solution grid or mesh do not permit a fully distributed characterization.

Several well-known general watershed models are in current use in many countries; some models are global and some are popular in a region. These models vary significantly in the model construct of each individual component process partly because these models serve somewhat different purposes. A number of catchment models are freely available in public domain. Some popular free hydrologic models

include SWAT, VIC, HEC-HMS, MODFLOW, and CROPWAT. HEC-HMS is frequently used for design of drainage systems, quantifying the effect of land use change on flooding, etc. The NWS model is the standard model for flood forecasting in USA. Mike and SHE are the standard models for hydrologic analysis in many European countries. The HBV model is the standard model for flow forecasting in Scandinavian countries. The ARNO, LCS, and TOPIKAPI models are popular in Italy. The tank models are well accepted in Japan. The Xinanjiang model is a commonly used model in China. SWAT and VIC models are popular in diverse studies, including the impact of climate change. CROPWAT is extensively used to compute crop water requirements and MODFLOW is the most popular groundwater flow model.

6.1 Model Calibration

Once one or more models have been chosen for a project, it is necessary to determine their parameters. In general, it is not possible to measure the parameters of models or estimate them *a priori*. Studies that have attempted these have generally found that even after intensive measurements, satisfactory estimates of parameter values could not be obtained. Prior estimation of feasible ranges of parameters also often results in wide ranges of predictions, which may still not always contain the measured responses.

A good automatic parameter estimation methodology requires four elements: (1) objective function, (2) optimization algorithm, (3) termination criteria, and (4) calibration data. The choice of an objective function influences parameter estimates as well as the quality of model results. Sorooshian and Gupta (1995) discussed several optimization methods, including local search methods (direct search methods and gradient search methods) and global search methods (random search methods, multistart algorithms, and shuffled complex algorithms). The shuffled complex evolution (SCE-UA) global optimization algorithm has been found to be consistent, effective, and efficient in locating the globally optimum hydrologic model parameters (Duan et al. 1992).

The model performance is typically evaluated from the comparison of simulated and observed discharge data by using statistical indices. Commonly used indices are: coefficient of determination, Nash and Sutcliffe efficiency, index of agreement, and root mean square error. In addition, visual comparison of the observed and computed values and scatter plot between them are always helpful.

There are two major reasons for difficulties in calibration. First, the scale of measurement techniques available is generally much less than the scales at which parameter values are required. For example, consider hydraulic conductivity which is a common parameter in watershed models. Techniques for measuring soil hydraulic conductivities generally integrate over areas of less than 1 m². However, the size of typical elemental area in a distributed model would be about 100 m² or more. Studies suggest that effective values might change with scale. Thus, the small-scale values that are typically measured and the effective values required at the model

element scale may be different. Hence, the parameter values for a particular model will need to be calibrated.

Most calibrations involve some form of optimization of the parameter values by comparing the simulated values with observed values of the variables of interest. The parameter values are adjusted after each model run, either manually or by some optimization algorithm, until some “best fit” parameter set is found.

It needs to be highlighted here that the model structure and the observations are not error-free. Thus, the optimum parameter set is model specific and may not remain optimum if the model structure or the calibration data changes. While one optimum parameter set can often be found, there will usually be many other parameter sets that are very nearly as good, perhaps from a different part in the parameter space. The idea of equifinality of parameters (Beven and Freer 2001) suggests that given the limitations of both the model structures and observed data, there may be many representations of a catchment that may be equally valid in terms of their ability to produce acceptable simulations of the available data.

6.2 Selection of Appropriate Model Type

In the presence of a large number of hydrological models, a frequent question is “which model is most appropriate for a particular problem?” This question cannot be answered by giving the name of a particular model. Instead, one may only recommend as to which of the above mentioned model types is most appropriate for the given hydrological problem, available data, and resources.

For some hydrological problems, the best model type is nearly obvious, e.g., probabilistic models for frequency analysis and stochastic models to generate long synthetic streamflow series. Empirical (black box) models are mainly employed for event-based modeling or as components of more complicated models. Lumped, conceptual models are suited to simulate the rainfall-runoff process when adequate data exist to calibrate the model. Typical applications of such models are extension of streamflow records based on long rainfall records, water balance, and real-time flood forecasting.

Theoretically, physically based distributed models can be applied to almost any hydrological problem. However, for many problems, the solutions can be obtained by less sophisticated empirical, lumped conceptual, or statistical models. Of course, there are complex problems, for which it is necessary to use a physically based distributed model. Some examples of their application are:

- Natural and anthropogenic changes in land-use and land cover, such as urbanization, forest clearance for agricultural purposes. The parameters of a physically based, distributed model have a direct physical interpretation. Hence, they can be estimated for the new state of the catchment and the impacts of changes can be examined before they occur.
- Ungauged catchments. Modeling of an ungauged catchment requires a program of fieldwork to provide data and parameters for calibration. Due to the physical

significance of the parameters, a physically based model can be applied to an ungauged basin or to a basin having shorter data record.

- To model the movement of pollutants and sediments, it is necessary to model the water flows which provide the basic frame work. Since water quality and sediment problems have a spatial aspect, distributed models are best suited for such problems.

6.3 Uncertainty in Hydrologic Modeling

Uncertainty is defined as a measure of imperfect knowledge or probable error which can occur during the data collection process, modeling and analysis of engineering systems, and prediction of a random process. In simple terms, uncertainty is the occurrence of events that are beyond human control. Uncertainty may also classified into two categories: (1) inherent or intrinsic, caused by randomness in nature; and (2) epistemic, caused by the lack of knowledge of the system or paucity of data. There are six sources of uncertainty in evaluating the reliability of environmental and water resources systems: (1) Natural uncertainties associated with random temporal and spatial fluctuations inherent in natural processes, e.g., climatic variability, occurrence of hydrologic extremes; (2) model structure uncertainty which reflects the inability of the simulation model to represent precisely the system's true behavior or process; (3) model parameter uncertainties which reflect the variability in determining the parameters to be used in a model or design; (4) data uncertainties arising due to measurement inaccuracy and errors, inadequacy of the data gaging network, and data handling and transcription errors; (5) computational uncertainties arise due to truncation and rounding off errors in doing calculations; and (6) operational uncertainties associated with construction, manufacturing, maintenance, and other human factors that are not accounted for in the modeling or design procedure. Montanari (2007) identified four types of techniques for assessing the uncertainty of the output of a hydrological model: (a) approximate analytical methods, (b) techniques based on the statistical analysis of model errors, (c) approximate numerical methods/sensitivity analyses, and (d) nonprobabilistic methods. To identify the uncertainty assessment method, one should take into account the following main issues: the type of model whose output uncertainty is to be inferred (simulation, forecasting) and the type of information available (observed data, information about model uncertainty).

7 Emerging Technology for Hydrologic Modeling

New data collection techniques, especially remote sensing, satellites, and radar, have received a great deal of attention and developments since the 1980s. Notable advances have been made in recent years which are gradually alleviating the scarcity of data which is one of the major difficulties in watershed modeling. Space-based technology provides data regarding topography, land use, land cover, soil

parameters, initial conditions; inventories of water bodies, such as dams, lakes, swamps, flooded areas, rivers; mapping of snow and ice conditions; water quality parameters; etc. (Engman and Gurney 1991). Satellite data are being increasingly used for the estimation of precipitation, temperature, evapotranspiration, and other meteorological inputs. Attempts are underway to estimate river flows from satellite data and these have the potential to overcome the handicaps due to missing river flow data in future.

A multitude of satellites, such as the Landsat Thematic Mapper (TM) Multispectral Scanner (MSS), the European Satellites, and the Indian Remote Sensing satellites, produce imageries which in conjunction with terrain data are successfully providing data for mapping and classification of land use, and vegetative cover. Similarly, the airborne Light Detection and Ranging (LIDAR) technology is being employed to provide real-time flood inundation maps. Special purpose satellites have been launched to measure precipitation, soil moisture, snow cover, and map topography at finer resolutions. Global Positioning System (GPS) has revolutionized field investigations. With the vastly improved capability, remote sensing and space technology is being increasingly coupled with watershed models for a variety of applications.

Physical characteristics of a watershed, such as soils, land use, and topography, vary spatially. Advances in digital mapping have provided essential tools to closely represent the three-dimensional nature of natural landscapes. One such tool is the digital terrain (DTM) or digital elevation (DEM) model. GIS systems now have the capability to automatically extract topographic features, such as basin geometry, stream networks, slope, aspect, flow direction, from raster DEMs.

8 Future Outlook

Mathematical models of watershed hydrology are now the most common and the best tools for all aspects of water resources management. The future is expected to witness a greater and growing integration of these models with environmental and ecological management. With growing technologies triggered by the information revolution, remote sensing technology, GIS, and data base systems, the hydrologic models are getting more sophisticated. These are increasingly being integrated with environmental, economic, and social models.

The future of hydrologic models will be shaped by several simultaneous factors. Two aspects that have begun to drive the application of hydrologic models are: (a) possible adverse impacts of climate change on society and water sector and what is a good adaptation strategy and (b) check the degradation of aquatic ecosystems due to faulty planning and indiscriminate exploitation. These issues cannot be handled without hydrologic models and this is gradually resulting in growing application of such models. In addition, increasing societal demand for integrated environmental management by incorporation of biological, chemical,

and physical aspects of the hydrological cycle, rapid advances in remote sensing, and geographical information systems (GIS) are setting the directions for changes and improvements in hydrological models. It is anticipated that the hydrology models will be required to be interfaced with economic and social models in future. These models will also become more global, not only in the sense of spatial scale but also in the sense of hydrologic details (Singh and Woolhiser 2002). New initiatives will lead to enhanced role of models in planning and decision making and growing demand for bundling models in a decision support system (DSS) framework. Users would expect clearer statements of reliability and risk associated with model results and decisions.

References

- M.B. Abbott, J.C. Bathurst, J.A. Cunge, P.E. O'Connell, J. Rasmussen, An introduction to the European hydrological system – system Hydrologique European “SHE”. History and philosophy of physically based distributed modelling system. *J. Hydrol.* **87**, 45–59 (1986)
- R.G. Allen, L.S. Pereira, D. Raes, M. Smith, Crop Evapotranspiration, Irrigation and Drainage Paper No. 56, Food and Agriculture Organization, Rome (1998)
- ASCE, *Handbook of hydrology*, ASCE Manual and Reports on Engrg. Pract. No. 28 (New York, 1996)
- J.C. Bathurst, J.M. Wicks, P.E. O'Connell, The SHE/SHESED basin scale water flow and sediment transport modeling system, Chapter 16, in *Computer Models of Watershed Hydrology*, ed. by V.P. Singh (Water Resources Publications, Littleton, 1995), pp. 563–594
- S. Bergstrom, Development and application of a conceptual runoff model for Scandinavian countries. SMHI Reports, No. 7, Norrkoping (1976)
- R.P. Betson, Water is watershed runoff? *J. Geophys. Res.* **69**(8), 1541–1552 (1964)
- K. Beven, *Rainfall-Runoff Modelling – The Primer* (Wiley, Chichester, 2001)
- K. Beven, J. Freer, Equifinality, data assimilation, and uncertainty estimation in mechanistic estimation of complex environmental systems using the GLUE methodology. *J. Hydrol.* **249**, 11–29 (2001)
- K.J. Beven, M.J. Kirkby, A physically-based variable contributing area model of basin hydrology. *Hydrol. Sci. Bull.* **24**(1), 43–69 (1979)
- R.J.C. Burnash, R.L. Ferral, R.A. McGuire, A generalized streamflow simulation system-conceptual modeling for digital computers. Rep., U. S. Dept. Commerce, National Weather Service and State of California, Department of Water Resource (March, 1973)
- J. Chen, H. Shi, B. Sivakumar, M.R. Peart, Population, water, food, energy and dams. *Renew. Sust. Energ. Rev.* **56**, 18–28 (2016)
- N.H. Crawford, R.K. Linsley, Digital simulation in hydrology: Stanford Watershed Model IV. Tech. Rep. No. 39 (Stanford University, Palo Alto, 1966)
- D.R. Dawdy, T. O'Donnell, Mathematical models of catchment behavior. *J. Hydraul. Div. ASCE* **91** (HY4), 123–127 (1965)
- J.C.I. Dooge, A general theory of the unit hydrograph. *J. Geophys. Res.* **64**(2), 241–256 (1959)
- O. Duan, V.K. Gupta, S. Sorooshian, Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resourc. Res.* **28**(4), 1015–1031 (1992)
- T. Dunn, L.B. Leopold, *Water in Environmental Planning* (W.H. Freeman, San Francisco, 1978), p. 818
- E.T. Engman, R.J. Gurney, *Remote Sensing in Hydrology* (Chapman and Hall, London, 1991)

- H. Gao, et al., Water budget record from Variable Infiltration Capacity (VIC) model Algorithm Theoretical Basis Document for Terrestrial Water Cycle Data Records (2010)
- Horton, R.E. (1931). The Field, Scope, and Status of the Science of Hydrology. Pp 189–202. in Trans. AGU, Reports and Papers, Hydrology. National Research Council, Washington, DC
- S.K. Jain, B. Storm, J.C. Bathurst, J.C. Refsgaard, R.D. Singh, Application of the SHE to catchments in India – Part 2: Field experiments and simulation studies with the SHE on the Kolar subcatchment of the Narmada river. *J. Hydrol.* **140**, 25–47 (1992)
- X. Liang, D.P. Lettenmaier, E.F. Wood, S.J. Burges, A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.* **99**(D7), 14415–14428 (1994)
- A. Montanari, What do we mean by ‘uncertainty’? The need for a consistent wording about uncertainty assessment in hydrology. *Hydrol. Process.* **21**, 841–845 (2007)
- J.E. Nash, The form of the instantaneous unit hydrograph. *Hydrol. Sci. Bull.* **3**, 114–121 (1957)
- S.L. Neitsch, J.G. Arnold, J.R. Kiniry, R. Srinivasan, J.R. Williams, *Soil and Water Assessment Tool, User Manual, Version 2000* (Grassland, Soil and Water Research Laboratory, Temple, 2002)
- NRC, *Scientific Basis of Water Resource Management* (National Research Council, National Academy Press, Washington D.C, 1982)
- NRC, Opportunities in the Hydrologic Sciences, in *Committee on ‘Opportunities in the Hydrologic Sciences’ of National Research Council* (National Academy Press, Washington, DC, 1991)
- T. Oki, S. Kanae, Global hydrological cycles and world water resources. *Science* **313**(5790), 1068–1072 (2006)
- D.M. Rockwood, Theory and practice of the SSARR model as related to analyzing and forecasting the response of hydrologic systems, in *Applied Modeling in Catchment Hydrology*, ed. by V.P. Singh (Water Resources Publications, Littleton, 1982), pp. 87–106
- L.K. Sherman, Stream flow from rainfall by the unit graph method. *Engrg. News Record*. **108**, 501–505 (1932)
- I.A. Shiklomanov, *World Water Resources: Modern Assessment and Outlook for the 21st Century. (Prepared in the Framework of IHP, UNESCO)* (State Hydrology Institute, St. Petersburg, 1999)
- V.P. Singh, *Elementary Hydrology* (Prentice Hall, Engelwood Cliffs, 1992)
- V.P. Singh, *Computer Models of Watershed Hydrology* (Water Resources Publications, Littleton, 1995)
- V.P. Singh, *Kinematic Wave Modeling in Water Resources: Surface-Water Hydrology* (Wiley, New York, 1996), p. 1399
- V.P. Singh, *Kinematic Wave Modeling in Water Resources: Environmental Hydrology* (Wiley, New York, 1997), p. 830
- V.P. Singh, D.K. Frevert, *Mathematical Models of Small Watershed Hydrology and Applications* (Water Resources Publications, Littleton, 2002a), p. 950
- V.P. Singh, D.K. Frevert, *Watershed Models* (CRC Press, Boca Raton, 2002b), p. 653
- V.P. Singh, D.K. Frevert, *Mathematical Models of Large Watershed Hydrology* (Water Resources Publications, Littleton, 2002c), p. 891
- V.P. Singh, D.A. Woolhiser, Mathematical modeling of watershed hydrology. *J. Hydrol. Eng. ASCE* **7**(4), 270–292 (2002)
- V.P. Singh, D.K. Frevert, *Watershed Models* (CRC Press, Boca Raton, 2006)
- S. Sorooshian, V.K. Gupta, Model calibration. Chapter 2, in *Computer Models of Watershed Hydrology*, ed. by V.P. Singh (Water Resources Publications, Littleton, 1995), pp. 23–68
- M. Sugawara, The flood forecasting by a series storage type model. *Internatl Symposium Floods and their Computation*, pp. 1–6, Leningrad, U.S.S.R., 1967
- M. Sugawara, et al., Tank model and its application to Bird Creek, Wollombi Brook, Bikin River, Kitsu River, Sanga River and Nam Mune. Research Note of the National Res. Center for Disaster Prevention, No. 11, (1974), pp. 1–64

- C.V. Theis, The relation between the lowering of the piezometric surface and the rate and duration of discharge of a well using ground-water storage: American Geophysical Union Transactions, 16th Annual Meeting, vol. 16, pt. 2 (1935), p. 519–524
- E. Todini, The ARNO rainfall-runoff model. *J. Hydrol.* **175**, 339–382 (1996)
- E. Todini, Hydrological catchment modelling: past, present and future. *Hydrol. Earth Syst. Sci.* **11**, 468–482 (2007). <https://doi.org/10.5194/hess-11-468-2007>
- E.F. Wood et al., A land-surface hydrology parameterization with subgrid variability for general circulation models. *J. Geophys. Res.-Atmos.* **97**(D3), 2717–2728 (1992)
- R.A. Wurbs, Dissemination of generalized water resources models in the United States. *Water Int.* **23**, 190–198 (1998)
- R.J. Zhao, Y.-L. Zhuang, L.R. Fang, X.R. Liu, Q.S. Zhang, The Xinanjiang model. Proceedings, Oxford Symposium on Hydrological Forecasting, IASH Pub. No. 129 (1980), pp. 351–356



Black-Box Hydrological Models

Chong-Yu Xu, Lihua Xiong, and Vijay P. Singh

Contents

1	Introduction	342
2	Antecedent Precipitation Index (API) Models	343
2.1	Calculations of API	344
2.2	Graphical API Models	345
2.3	Summary	345
3	Regression Models	347
3.1	Simple Linear Regression	347
3.2	Multiple Linear Regression Analysis	351
3.3	Summary	353
4	Time Series Models	353
4.1	Types of Hydrologic Time Series	354
4.2	Time Series Models for Stationary Data	354
4.3	Nonstationary Time Series Models	357
4.4	Summary	360
5	Artificial Neural Network (ANN) Models	361
5.1	Structure of ANN	361
5.2	Network Training	363
5.3	Summary	364
6	Fuzzy Logic Models	365
6.1	Basic Concepts of Fuzzy Systems	366
6.2	Operations with Fuzzy Sets	367

C.-Y. Xu

Department of Geosciences, University of Oslo, Oslo, Norway

e-mail: chongyu.xu@geo.uio.no; c.y.xu@geo.uio.no

L. Xiong

Department of Hydrology and Water Resources, Wuhan University, Wuhan, China

e-mail: 00011618@whu.edu.cn

V. P. Singh (✉)

Department of Biological and Agricultural Engineering and Zachry Department of Civil Engineering, Texas A and M University, College Station, TX, USA

e-mail: vsingh@tamu.edu

6.3	Types of Fuzzy Systems	370
6.4	Adaptive Neuro-Fuzzy Inference System (ANFIS)	372
6.5	Summary	372
7	Frequency Analysis Models	374
7.1	Graphical Method	374
7.2	Analytical Method – Frequency Factor Method	375
7.3	Data Sampling Methods	378
7.4	Outliers and Zeros	378
7.5	Regionalization	379
7.6	Summary	380
	References	381

Abstract

This chapter discusses different types of black-box hydrological models that are based on input-output relationships rather than physical principles. They include antecedent precipitation index (API) models, regression models, time series models, artificial neural network (ANN) models, fuzzy logic models, and frequency analysis models. The purpose of this chapter is neither to provide a complete discussion of the theory of hydrological systems nor to offer a complete coverage of the studies published in the literature. Rather, the chapter is focused on presenting general theories and methods of different types of black-box models, basic model forms, and related applications in hydrology and water resources engineering.

Keywords

Black-box · Gray-box · White-box models · Flood forecasting · Hydrology

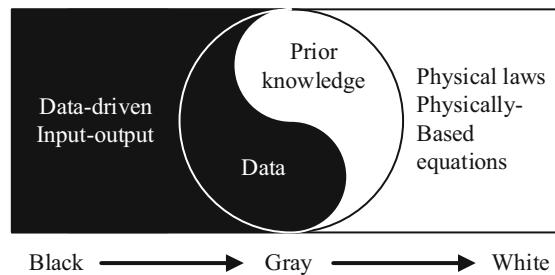
1 Introduction

Hydrological systems and their underlying processes are complicated. In real world, these systems are approximated by hydrological models, and a model is considered adequate if the difference between model prediction and measurement is small, satisfying a predetermined criterion.

Since the pioneering development of the rational method in the middle of the nineteenth century (Mulvaney 1850), development of hydrological models has gone through several stages, input-output models (black box), lumped conceptual models (grey box), and physically based distributed models (white box) (Fig. 1). As the simplest models, black-box hydrological models are based on input-output relations and do not describe the underlying hydrologic processes. In many cases, however, black-box models are adequate and served as a first step for modelers to conceptualize and simplify hydrologic systems. In fact, other types of hydrological models, i.e., gray-box and white-box models, more or less also owe their origin to black-box models.

Compared to other types of models, black-box models require the least input data. Precipitation and temperature are frequently used as input. Temperature is significant for hydrological modeling, especially, in climate regimes with snow.

Fig. 1 Black-box, gray-box, and white-box models



Other hydrometeorological variables can also be possibly employed in black-box models. Some models can even be run only with output data, such as runoff data. Spatial characteristics of catchments are seldom considered but they are useful in the estimation of model parameters and can aid model interpretation and analysis.

Further, running black-box models is more easily achieved in terms of computer resources. Most black-box models do not demand high computational capability. Hence, a personal laptop or computer is enough for most cases. Many software or computer languages, such as Matlab or R, are friendly in programming black-box models.

This chapter describes different types of black-box hydrological models that are based on input-output relationships, including graphical antecedent precipitation index (API) models, regression models, time series models, artificial neural network (ANN) models, fuzzy logic models, and frequency analysis models. In the following sections, the main types of black-box models are discussed with the focus placed on the definition of a model, mathematical description and schematic representation of the elements and structures of basic model forms, and examples of applications in hydrology and water resources engineering. It should be noted that a complete discussion of the theory of hydrological systems or a complete coverage of the studies published in the literature is not a major concern of this chapter. For each type of black-box hydrological models, the basic forms are presented and a general discussion of their strengths and weaknesses is also provided.

2 Antecedent Precipitation Index (API) Models

The antecedent precipitation, i.e., the amount of precipitation that has occurred prior to a single storm event, plays an important role in calculating the runoff response to rainfall, especially in catchments where runoff generation is dominated by soil water or groundwater storage, i.e., runoff generation follows the principle of the “variable source area” theory (Hewlett and Hibbert 1967). The antecedent precipitation index (API) is generally defined as the weighted summation of daily precipitation amounts that is used as an index of soil moisture, and is expressed by the following equation (Kohler and Linsley 1951):

$$\text{API} = \sum_{t=-1}^{-i} P_t k^{-t} \quad (1)$$

where P_t is the amount of precipitation on the t th day prior to the occurrence of storm, and k is normally a constant.

The above exponential model is based on the assumption that the greater the time lapse between a rainfall event and a given day, the less influence the rain has on the soil moisture content of that day (Saxton et al. 1967). The value of i is usually taken as 5, 7, or 14 days (Viessman and Lewis 1996; Ali et al. 2010).

2.1 Calculations of API

The decrease of API is usually assumed to follow a logarithmic decay. Thus, during periods of no precipitation:

$$\text{API}_i = k \times \text{API}_{i-1} \quad (2)$$

For periods with precipitation:

$$\text{API}_i = k \times (\text{API}_{i-1} + P_i) \quad (3)$$

This means that if any rain occurs, it should be added to the index (Fig. 2). In areas of snowfall, precipitation is applied to the model on the days when it melts rather than on the days when it falls. The value of k varies with basin physical characteristics and the meteorological condition, and a range of 0.85-0.90 over most of the eastern central parts of the United States was suggested (Viessman and Lewis 1996), which can be used as a reference for other regions.

To overcome the paradoxes that API often remains a subjectively determined and arbitrarily implemented parameter in rainfall-runoff modeling, Heggen (2001) proposed the use of a normalized antecedent precipitation index (NAPI) in place of API (Equation (4)). NAPI is defined as the ratio of the API on the day to the product of the average daily precipitation and the weighted sum of decay coefficients of the respective days before the storm.

$$\text{NAPI} = \frac{\sum_{t=-1}^i P_t k^{-t}}{\bar{P} \sum_{t=-1}^{-i} k^{-t}} \quad (4)$$

where \bar{P} is the average rainfall for antecedent days, and the other terms have been defined before. The soil moisture condition is assumed to be “dry” if $\text{NAPI} < 0.33$, the wet condition is defined as $\text{NAPI} > 3$, and the intermediate range $0.33 \sim 3$ is the “fair” condition (Hong et al. 2007).

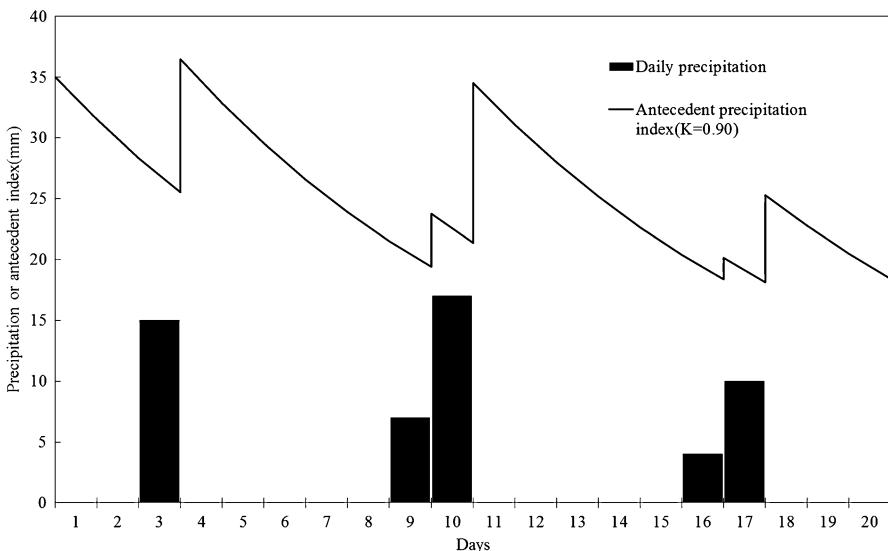


Fig. 2 API relation (modified from NWSRFS User Manual Documentation at http://www.nws.noaa.gov/ohd/hrl/nwsrfs/users_manual/htm/xrfsdocepdf.php)

2.2 Graphical API Models

During the 1950–1960s, scientists were seeking techniques which would (1) simplify the relationships of rainfall and runoff, (2) require less time for the calculation and forecasting especially when the computers were not yet available, and (3) not require information of soil and surface characteristics, vegetation differences, and land use, which were usually not available.

Because of the importance of antecedent soil moisture condition to runoff generation, many indices have been used to estimate the moisture condition, such as (1) days since last rain, (2) discharge at the beginning of the storm, and (3) antecedent precipitation. API, a rough representation of the initial soil-moisture condition, generally provides better results among the three indices and can also be easily determined. Based on API, Kohler and Linsley (1951) developed a relationship between storm runoff and precipitation by a graphical method of coaxial relations (Fig. 3). The graphical method consists of 3 three-variable relations, relating storm runoff as the dependent variable to the antecedent precipitation (API), date (week number), rainfall amount, and rainfall duration as independent variables.

2.3 Summary

Antecedent soil moisture condition is important for watershed modeling that ultimately provides information on flood forecasting, water resources management, hydroelectric power generation, and irrigation management. Because the observed

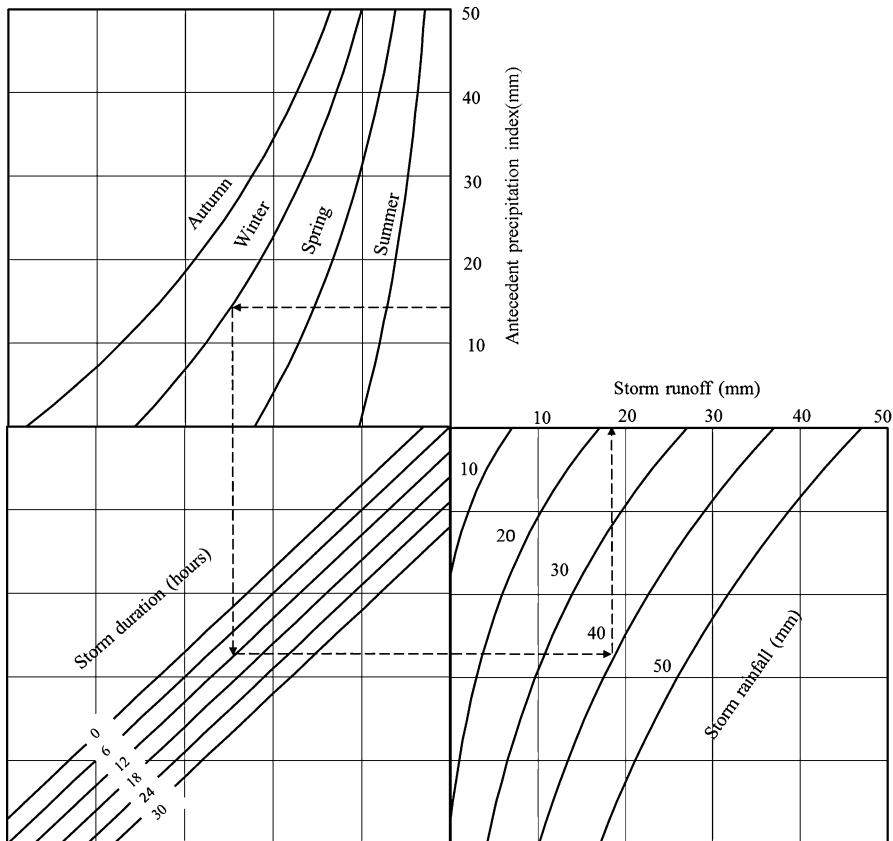


Fig. 3 Coaxial relation – antecedent precipitation index (modified from Sittner et al., 1969)

soil moisture data at a larger scale are usually not available, antecedent precipitation index, API, a rough representation of the initial soil-moisture condition, has been widely used in different hydrological modeling studies since the 1950s and the studies have generally shown that the use of API has potential to provide satisfactory results.

One typical example of using API as an important part of hydrological modeling is the model developed by the U.S. Soil Conservation Service (SCS) (1972) that incorporates initial losses or abstractions into a coefficient as a function of what is referred to as curve number, CN, which is a function of land use, soil type, hydrologic condition of basin, and the antecedent moisture condition (AMC) that generally is equivalent to the concept of antecedent precipitation index (API).

Other applications of API can be found in more complex rainfall-runoff models for storm runoff simulation (Saxton et al. 1967; Sittner et al. 1969; Fedora and Beschta 1989; Ali et al. 2010; Rajurkar et al. 2004; Dawson and Abrahart 2007), and in other models for landslide studies (Glade et al. 2000; Ma et al. 2014), global

runoff simulation (Hong et al. 2007), and the calculation of Forest Fire Danger Index (FFDI) (Liu et al. 2003), etc.

3 Regression Models

Regression (the term was first used by Pearson 1908) analysis is commonly used to describe quantitative relationships between a response variable and one or more explanatory variables. In hydrology, regression model is a useful tool for detecting relations between runoff and precipitation for the same watershed, between runoff (or precipitation) in different watersheds, between crop growth and precipitation, and so on.

An analytical problem to be solved by regression analysis involves (Riggs 1985): (1) selection of factors which are expected to influence the dependent variable; (2) describing these factors quantitatively; (3) selection of the regression model; (4) computing the regression equation, the standard error of estimate, and the significance of the regression coefficients; and (5) evaluation of results.

3.1 Simple Linear Regression

Regression analysis is a statistical tool that utilizes the relation between two or more quantitative variables so that one variable (dependent variable) can be predicted from others (independent variables).

The simplest statistical model for simple linear regression

$$Y_i = a + bX_i + e_i \quad i = 1, \dots, n \quad (5)$$

$$\hat{Y}_i = a + bX_i \quad (6)$$

where:

e_i is the error term or residual of the regression line, e_1, \dots, e_n are unobservable random variables, and are usually assumed as independent and normally distributed with mean zero and an unknown constant standard deviation, σ ; X_i and Y_i are the observed independent and dependent variables, respectively; \hat{Y}_i are the values estimated from the regression line; a and b are regression coefficients, where b is called the *slope* of the line and a is the *y-intercept*. The slope measures the amount Y increases/decreases when X increases/decreases by one unit. The *y-intercept* is the value of Y when $X = 0$.

3.1.1 Parameter Estimation

The goal is to find the equation of the straight line

$$\hat{Y}_i = a + bX_i$$

which would provide a “best” fit for the data points. That is to say, simple linear regression fits a straight line through the set of n points in such a way that makes the sum of squared *residuals* of the model as small as possible. In statistics, simple linear regression is the least squares estimator of a linear regression model. In other words, a (the y -intercept) and b (the slope) are solved by the following minimization problem:

$$\min_{a,b} \sum_{i=1}^n \hat{e}_i^2 = \sum_{i=1}^n (Y_i - a - bX_i)^2 \quad (7)$$

It can be shown that the values of a and b that minimize the objective function (7) are

$$b = \frac{\sum_{i=1}^n X_i Y_i - n\bar{X}\bar{Y}}{\sum_{i=1}^n X_i^2 - n\bar{X}^2} = \frac{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{cov}(X, Y)}{\text{var}(X)} \quad (8)$$

$$a = \bar{Y} - b\bar{X} \quad (9)$$

3.1.2 Model Evaluation

Coefficient of determination: After fitting a line to the data-points, we want to know how much of the variability in the dependent variable (Y) is explained by the regression. For this, the coefficient of determination (R^2) is often used and is expressed as:

$$R^2 = \frac{\text{Explained variance}}{\text{Total variance}}$$

The variability in the dependent variable Y is quantified as a sum of squares:

- $\sum (Y_i - \bar{Y})^2$ = total sum of squares corrected for the mean = total variance
- $\sum (\hat{Y}_i - \bar{Y})^2$ = the squared deviations of the predicted values from the mean value, explained variance by the regression line
- $\sum (Y_i - \hat{Y}_i)^2$ = the sum of squares of deviation from the regression = unexplained variance

The most general definition of the coefficient of determination is (Haan, 2002),

$$R^2 = \frac{\sum (\hat{Y}_i - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2} \quad (10)$$

R^2 ranges from 0 to 1, and it is normally expressed as a percentage.

Standard error of estimate (S): A measure of the variability of the regression line, i.e., the dispersion around the regression line is S . It tells how much variation there is in the dependent variable between the raw value and the expected value in the regression:

$$S = \sqrt{\frac{\text{SSE}}{n-2}} = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n-2}} \quad (11)$$

where SSE is the residual sum of squares or the sum of squares due to error. This S allows us to generate the confidence interval on the regression line as well as on regression coefficients.

Standard error (deviation) for parameters a and b : In regression analysis, the standard errors of the least square estimators for a (S_a) and b (S_b) are estimated by

$$S_a = S \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (12)$$

$$S_b = S \sqrt{\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (13)$$

3.1.3 Confidence Intervals

Confidence intervals are important for testing the statistical significance of the regression coefficient as well as the regression line.

A $100(1-\alpha)\%$ confidence interval for a and b is:

$$(\hat{a} - t_{\alpha/2} S_a, \hat{a} + t_{\alpha/2} S_a)$$

$$(\hat{b} - t_{\alpha/2} S_b, \hat{b} + t_{\alpha/2} S_b)$$

where α is the significance level, and $t_{\alpha/2}$ is the critical value of the t -distribution with degrees of freedom (d.f.) = $n-2$.

A $100(1-\alpha)\%$ confidence interval on the regression line is:

$$\left(\hat{y}_k \pm t_{\alpha/2} \sqrt{\frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)$$

where \hat{y}_k represents the predicted mean value of \hat{y}_k ; the confidence intervals are the narrowest at $x_k = \bar{x}$ and widen as x_k deviates from \bar{x} as can be seen from Fig. 4.

A $100(1-\alpha)\%$ confidence interval on the individual points is:

$$\left(\hat{y}_k \pm t_{\alpha/2} \sqrt{1 + \frac{1}{n} + \frac{(x_k - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \right)$$

where the symbols are the same as above.

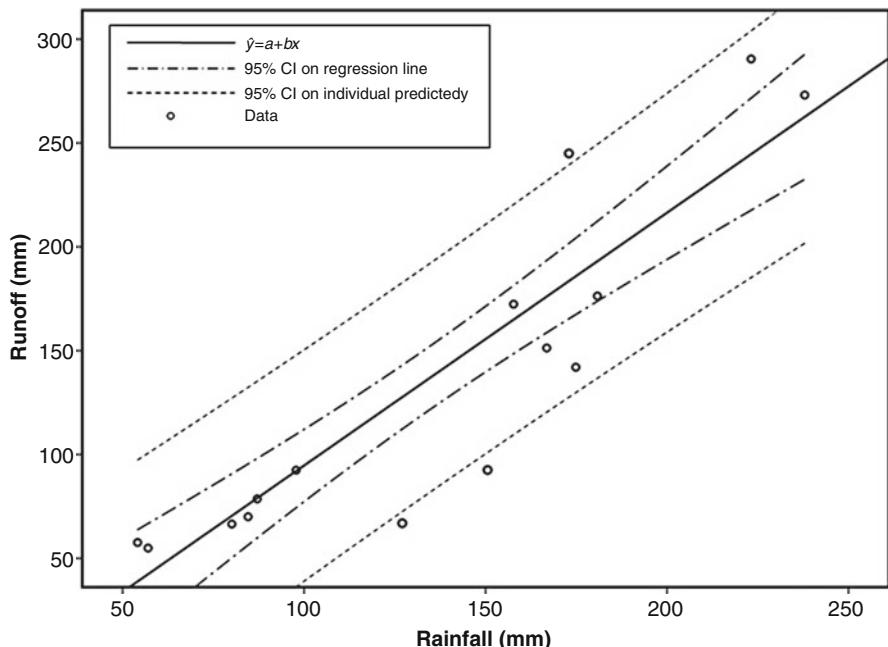


Fig. 4 A typical plot of simple regression with the 95% confidence intervals and 95% prediction intervals (modified from Haan, 2002)

3.1.4 Significance of Coefficients

The statistical significance of parameters a and b equal to or larger/smaller than a given value (including zero) can be tested based on the t -distribution against a one- or two-sided alternative, depending on the nature of the relation that is anticipated. For example, if we want to test if a is significantly different from zero, the null hypothesis would be $H_0: a = 0$, and the test statistic is $t = \frac{a-0}{s_a}$. H_0 will be rejected if $|t| \geq t_{1 - \alpha/2, n-2}$.

Similarly, if we want to test if b is significantly different from zero, the null hypothesis would be $H_0: b = 0$, and the test statistic is $t = \frac{b-0}{s_b}$. H_0 will be rejected if $|t| \geq t_{1 - \alpha/2, n-2}$.

3.2 Multiple Linear Regression Analysis

The general purpose of multiple linear regression is to learn more about the relationship of a dependent or criterion variable to several independent or predictor variables.

In general, a multiple regression procedure estimates a linear equation of the form:

$Y_i = a + b_1x_{i1} + b_2x_{i2} + \dots + b_px_{ip} + e_i \quad i = 1, \dots, n$ where x_{i1}, \dots, x_{ip} are the values of the input variables for the i th experimental run and Y_i is the corresponding response. The error terms e_i are usually assumed to be independent and normally distributed with mean zero and constant variance σ^2 . The unknown parameters are a , b_i , and σ^2 or σ .

3.2.1 Parameter Estimation

As in the simple linear regression, the first step in the multiple regression analysis is to obtain the least squares estimates of parameters a and b_i that minimize

$$\sum (y_i - a - x_{i1}b_1 - \dots - x_{ip}b_p)^2$$

In practice, n observations would be available on the dependent variable y and independent variables x_1 to x_p . The p unknown parameters are estimated from the n observations. Thus, n must be equal to or greater than p , and in practice, n should be at least 3 or 4 times as large as p .

Most of least squares analyses of multiple linear regression models are carried out with the aid of a computer.

3.2.2 Evaluation of Multiple Regression Model

Similar to simple linear regression, R^2 , the *coefficient of multiple determination* or *multiple coefficient of determination* is computed and used to evaluate how good the multiple regression is. The *multiple coefficient of determination* is defined as

$$\begin{aligned}
 R^2 &= \frac{\text{Sum of squares due to regression}}{\text{Sum of squares corrected for the mean}} = \frac{\text{Regression SS}}{\text{Total SS}} \\
 &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}
 \end{aligned} \tag{14}$$

where y_i are the observed values of the dependent variable, \bar{y} as its mean, and \hat{y}_i are the fitted values.

There are two reasons causing R^2 to tend to overestimate the variance accounted for, compared to an estimate that would be obtained from the population: a large number of predictors and a small sample size. Therefore, the calculated R^2 values usually need to be adjusted using Eq. (15). So, with a large sample and with few predictors, adjusted R^2 should be very similar to the R^2 value.

$$R^2_{\text{adjusted}} = 1 - (1 - R^2) \left(\frac{n - 1}{n - k - 1} \right) \tag{15}$$

where n is the number of data, and k is the number of independent variables used in the regression.

3.2.3 Stepwise Multiple Regression

In order to minimize the problem of over-parameterization, the independent variables need to be carefully chosen. The principle is that on one hand all relevant variables for theoretical or other reasons should be included, and on the other hand as few independent variables as possible (principle of parsimony) should be included, also because the more variables, the greater uncertainty, the larger type II error, and the fewer degrees of freedom.

In statistics, **stepwise regression** includes regression models in which the choice of predictive variables is carried out by an automatic procedure. The decision as to which variable should be included is determined by performing a sequence of F-tests or t-tests, but other techniques are also possible.

The main approaches are as follows:

Forward selection: In stage one, the independent variable best correlated with the dependent variable is included in the equation. In the second stage, the remaining independent variables with the highest partial correlation with the dependent variable, after the first independent variable is removed, is entered. The improvement of R^2 value in each step is checked by the F-test (Haan 2002).

$$F_c = \frac{(1 - R_{k-1}^2) \cdot (n - k - 1)}{(1 - R_k^2) \cdot (n - k - 2)} \tag{16}$$

where R_{k-1}^2 and R_k^2 are the determination coefficient with $k-1$ and k -independent variables, respectively; n is the number of data; and k is used as the number of independent variables. If $F_c > F_{1-\alpha, N-n-1, N-n-2}$, we know that the addition of X_n is significant.

We continue until no variables "significantly" explain the residual variation. At each step, the increment of R^2 is tested by the F-test.

Backward elimination which involves starting with all variables and eliminating independent variables one at a time until the elimination of one makes a significant difference in R-squared.

Multiple linear regression models are probably one of the most commonly used methods for hydrologic forecasting. However, various other types of statistically based regression models, e.g., nonlinear regression, principal component regression, partial least squares regression, are also used in hydrological studies (Eldaw et al. 2003; Sharda et al. 2008; Adamowski et al. 2012; Yasar et al. 2012).

3.3 Summary

Simple and multiple regression techniques are widely used in hydrology. Many applications of regression models can be found in the following categories: (1) hydrological forecasting, including streamflow forecasting (Yu and Liang 2007), rainfall prediction (Makarau and Jury 1997; Francis and Renwick 1998), water demand forecasting (Billings and Agthe 1998; Polebitski and Palmer 2009), etc.; (2) transferring information on hydrological behavior to ungauged catchments (Vogel and Kroll 1990; Pandey and Nguyen 1999); (3) infilling missing values of hydrologic variables, such as runoff, precipitation, temperature, soil moisture, etc. (Beauchamp 1989; Eischeid et al. 2000; Dumedah and Coulibaly 2011); and (4) regression models, which are also one of the four general categories of downscaling methods. The relationships between large-scale and local-scale climatic fields were established by regression-based schemes (Hewitson and Crane 1996; Wilby et al. 1999; Charles et al. 2007).

In regression models, it should be noted that the regression equation does not imply a cause-and-effect relationship of the dependent variable to independent variables. Both may be influenced by some other factors that are not readily measured. However, there should be some physical tie between the variables if the results can be considered meaningful. Thus, there should be a physically plausible argument for selecting the explanatory variables to estimate the dependent variable.

Like many other statistical procedures, the regression analysis method described above is built under the assumption that the data are normally distributed, but the types of data used in hydrology commonly are not normally distributed, and some have no probability distribution at all. Hydrologists must select procedures most nearly suitable to the characteristics of data and must interpret the results accordingly.

4 Time Series Models

This section deals with basic time series models, which have become a major tool in hydrology in the era of information technology. In hydrology, time series models are usually used for building mathematical models to generate synthetic hydrologic

records, to forecast hydrologic events, to detect trends and shifts in hydrologic records, and to fill in missing data and extend records (Salas et al. 1980, Salas 1993; Haan 2002).

A time series is a series of observations of a variable in the course of time, where time is discretized to a series of time points or moments. A complete observed time series, $y(t)$, can be decomposed into a number of components as expressed by:

$$y(t) = y_1(t) + y_2(t) + y_3(t) + y_4(t) \quad (17)$$

where $y_1(t)$ is the trend component, $y_2(t)$ is the periodic component, $y_3(t)$ is the catastrophic event, and $y_4(t)$ is the random stochastic component. The first two terms are deterministic and can be identified and quantified fairly easily; the last two are stochastic and cannot easily be identified and quantified.

4.1 Types of Hydrologic Time Series

Stationary and nonstationary series: If the statistics of the sample (mean, variance, covariance, autocorrelation, etc.) do not change with time or length of the sample, then the time series is said to be stationary to the second-order moment, weakly stationary, or stationary in a broad sense. Otherwise, it is a nonstationary series, i.e., if a definite trend is discernible in the series or there is periodicity in a series, then the time series is nonstationary.

Generally, annual hydrologic time series are considered to be stationary, although this assumption may not be strictly correct due to large-scale climate changes and human activities. On the other hand, hydrologic time series defined at time scales smaller than a year, such as monthly and daily series, are typically nonstationary.

White noise series: For a stationary time series, if the process is purely random and stochastically independent, the time series is called a white noise series. It is the simplest example of a stochastic process. Such processes contain no memory by construction, that is, for every t , element X_t is independent of every other element in the process.

Gaussian time series: A Gaussian random process is a process (not necessarily stationary) of which all random variables are normally distributed, and of which all simultaneous distributions of random variables of the process are normal. When a Gaussian random process is weakly stationary, it is also strictly stationary, since the normal distribution is completely characterized by its first- and second-order moments.

4.2 Time Series Models for Stationary Data

A time series model is an empirical model for stochastically simulating and forecasting the behavior of uncertain hydrologic systems. Time series models include

stochastic models for a purely random time series with known distribution, for stationary time series, e.g., autoregressive (AR) models, moving average (MA) models, autoregressive moving average (ARMA) models, as well as for nonstationary time series, e.g., autoregressive integrated moving average (ARIMA) models, Thomas-Fiering model, etc.

4.2.1 Time Series Models for Purely Random Series with Known Probability Distribution

Possibly, the simplest stochastic process to model is where the events can be assumed to occur at discrete times with the time between events constant, the events at any time are independent of the events at any other time, and the probability distribution of the event is known. Stochastic generation from a model of this type merely amounts to generating a sample of random observations from a univariate probability distribution.

Example: If X_t is a white noise series and normally distributed, i.e., $X_t \sim N(\mu_x, \sigma_x^2)$, then the model can be

$$X_t = \mu_x + \sigma_x Z \quad (18)$$

where μ_x and σ_x are the mean and standard deviation of the X_t series, and $Z \sim N(0, 1)$ is a random series having standard normal distribution, which can be generated by Monte Carlo simulation.

4.2.2 Autoregressive Models

The autoregressive models are used to model stationary time series when persistence (memory) is present. The general form of a p th-order autoregressive model, also called Markov type model, AR(p), is

$$\begin{aligned} y_t &= \mu + \beta_1(y_{t-1} - \mu) + \beta_2(y_{t-2} - \mu) + \cdots + \beta_p(y_{t-p} - \mu) + \varepsilon_t \\ &= \mu + \sum_{i=1}^p \beta_i(y_{t-i} - \mu) + \varepsilon_t \end{aligned} \quad (19)$$

where μ is the mean value of the series, p is the order of AR model, written as AR(p), β_i are the regression coefficients, and ε_t are the noise or prediction error, normally assumed as $N(0, \sigma_\varepsilon^2)$. There are $p+2$ parameters to be estimated: $\beta_1, \beta_2, \dots, \beta_p, \mu$, and σ_ε^2 , the variance of residuals.

The most frequently encountered AR processes are of first or second order, and AR(0) process is white noise.

The equation for the first-order autoregressive model is:

$$y_t = \mu + \beta_1(y_{t-1} - \mu) + \varepsilon_t \quad (20)$$

Parameters β_1, μ , and σ_ε of the model are estimated using the Yule-Walker equations. The relation between regression coefficient β and autocorrelation coefficient ρ is written as:

$$\rho_k = \sum_{j=1}^P \beta_j \rho_{k-j} \quad (21)$$

where ρ_k is the autocorrelation coefficient. The parameters are estimated as:

$$\hat{\beta}_1 = \rho_1; \hat{\sigma}_\varepsilon^2 = \sigma_y^2(1 - \beta_1^2); \hat{\mu} = \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad (22)$$

The procedure for generating a series of values for y_t using AR(1) model is:

- Estimate μ_y , σ_y , and β_1 by $\bar{y} = \mu_y$, $s_y = \sigma_y$, and $r_1 = \beta_1$, respectively, and $\sigma_\varepsilon^2 = \sigma_y^2(1 - \beta_1^2)$
- Select a z_t at random from an $N(0, 1)$ distribution
- Select an initial value for y_{t-1}
- Calculate y_t based on \bar{y} , s_y , and β_1 , and y_{t-1} by

$$y_t = \mu_y + \beta_1(y_{t-1} - \mu_y) + z_t \sigma_y \sqrt{(1 - \beta_1^2)} \quad (23)$$

- Delete the first 50 values to get rid of the influence of the initial values

4.2.3 Moving-Average Models

The moving-average model of order q process, denoted by $MA(q)$, is formulated as follows:

$$y_t = \mu_y + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q} \quad (24)$$

where ε_t is a white noise with $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$; θ_i are parameters of order q , i.e., parameter $\theta_k = 0$ for $k > q$.

The above equation gives the definition of the MA model: For a white noise or purely random series, ε_t is assumed to be normally distributed with zero mean and constant standard deviation. That is, a moving average model is conceptually a linear regression of the current value of the series against the previous (unobserved) white noise error terms or random shocks.

4.2.4 ARMA Models

Autoregressive moving-average (ARMA) models, sometimes called Box-Jenkins models (Box and Jenkins 1976), consist of two parts, an autoregressive (AR) part and a moving-average (MA) part. The model is usually then referred to as the $ARMA(p, q)$ model, where p is the order of the autoregressive part and q is the order of the moving-average part.

In this case, x_t is a mixed process where the output is a function of past outputs and current/past inputs

$$x_t = c + \sum_{i=1}^p x_{t-i}\beta_i + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j} \quad (25)$$

All notations have the same meaning as before. The error terms ε_t are generally assumed to be independent identically distributed random variables (i.i.d.) sampled from a normal distribution with zero mean: $\varepsilon_t \sim N(0, \sigma_\varepsilon^2)$, where σ_ε^2 is the variance of the error.

There are $p+q+2$ parameters (β_i , $i = 1, \dots, p$; θ_i , $i = 1, \dots, q$; c ; σ_ε). Some formulations transform the series by subtracting the mean of the series from each data point. This yields a series with a mean of zero. Whether one needs to do this or not is dependent on the software one uses to estimate the model.

In practice, the ARMA(1,1) model is often used:

$$x_t = \beta_1 x_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad (26)$$

Parameters β_1 , θ_1 , and σ_ε^2 can be estimated by solving the following equations:

$$\rho_1 = \frac{(\beta_1 - \theta_1)(1 - \theta_1\beta_1)}{1 + \theta_1^2 - 2\beta_1\theta_1} \quad (27)$$

$$\rho_2 = \beta_1 \rho_1 \quad (28)$$

$$\sigma_\varepsilon^2 = \frac{(1 - \beta_1^2)\sigma_x^2}{1 - 2\beta_1\theta_1 + \theta_1^2} \quad (29)$$

4.3 Nonstationary Time Series Models

4.3.1 ARIMA Models

The acronym ARIMA stands for "Auto-regressive integrated moving average." The ARMA models are suitable for data with the following two basic characteristics: (1) no apparent deviation from stationary assumption and (2) rapidly decreasing autocorrelation function. If these conditions are not met by a time series, a proper transformation should be performed to generate the time series with the above two conditions satisfied. This has usually been achieved by differencing, which is the essence of ARIMA models (Karamouz et al. 2012). It should be noted that the ARIMA models are nonstationary and cannot be used for synthetic generation of stationary time series, but they can be used for forecasting (Salas et al. 1980; Weeks and Boughton 1987).

4.3.2 First-Order Markov Process with Periodicity: Thomas-Fiering Model

The first-order Markov model of the previous section assumes that the process is stationary in its first three moments. It is possible to generalize the model so that the

periodicity in hydrologic data is accounted for to some extent. The main application of this generalization has been in generating monthly streamflow where pronounced seasonality in monthly flows exists. In its simplest form, the method consists of the use of 12 linear regression equations. If, say, 30 years of records are available, the thirty January flows and the thirty December flows are abstracted and the January flow is regressed upon the December flow; similarly, the February flow is regressed upon the January flow, and so on for each month of the year.

$$\begin{aligned} q_{\text{Jan}} &= \bar{q}_{\text{Jan}} + b_{\text{Jan}}(q_{\text{Dec}} - \bar{q}_{\text{Dec}}) + \varepsilon_{\text{Jan}} \\ q_{\text{Feb}} &= \bar{q}_{\text{Feb}} + b_{\text{Feb}}(q_{\text{Jan}} - \bar{q}_{\text{Jan}}) + \varepsilon_{\text{Feb}} \\ &\dots \end{aligned}$$

Fig. 5 shows a regression analysis of q_{j+1} on q_j , pairs of successive monthly flows for the months ($j+1$) and j over the years of record, where $j = 1, 2, 3, \dots, 12$ (January, February, ..., December) and when $j = 12$, $j+1 =$ January of next year (there would be 12 such regressions). If the regression coefficient of month $j+1$ on j is b_j , then the regression line values of a monthly flow, \hat{q}_{j+1} , can be determined from the previous month's flow q_j , by the equation:

$$\hat{q}_{j+1} = \bar{q}_{j+1} + b_j(q_j - \bar{q}_j)$$

To account for the variability in the plotted points about the regression line reflecting the variance of measured data about the regression line, a further random component is added:

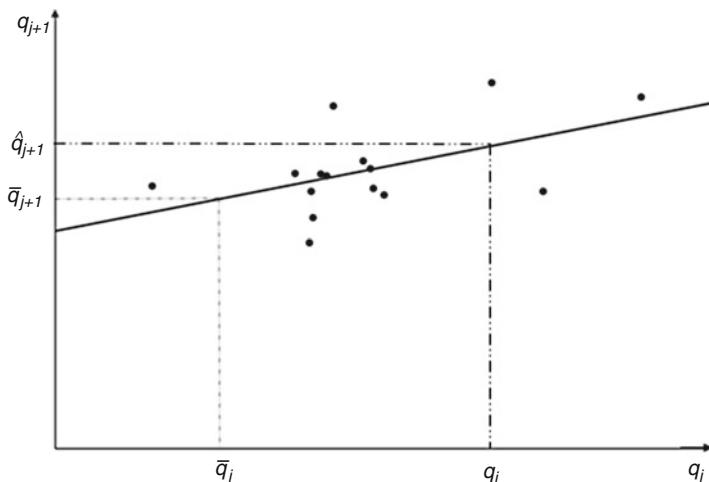


Fig. 5 Regression analysis of q_{j+1} on q_j

$$\varepsilon_j = Z \cdot S_{j+1} \sqrt{(1 - r_j^2)}$$

where s_{j+1} is the standard deviation of flows in month $j+1$, r_j is the correlation coefficient between flows in months $j+1$ and j throughout the record, and $Z = N(0, 1)$ a normally distributed random deviate with zero mean and unit standard deviation. The general form may be written as

$$\hat{q}_{j+1,i} = \bar{q}_{j+1} + b_j(q_{j,i-1} - \bar{q}_j) + Z_{j+1,i} \cdot s_{j+1} \sqrt{(1 - r_j^2)} \quad (30)$$

where $b_j = r_j \times s_{j+1}/s_j$. There are 36 parameters for the monthly model (\bar{q} , r , and s for each month). Subscript j refers to a month. For monthly synthesis, j varies from 1 to 12 throughout the year. Subscript i is a serial designation from year 1 to year n . Other symbols are the same as mentioned earlier.

The procedure for using the model is as follows:

1. For each month, $j = 1, 2, \dots, 12$, calculate

$$(a) \text{ Mean flow } \bar{q}_j = \frac{1}{n} \sum_i q_{j,i}; \quad (i = j, 12 + j, 24 + j, \dots)$$

$$(b) \text{ Standard deviation } S_j = \sqrt{\frac{\sum_i (q_{j,i} - \bar{q}_j)^2}{n-1}}$$

(c) The correlation coefficient with flow in the preceding month,

$$r_j = \frac{\sum_{i=1} (q_{j,i} - \bar{q}_j)(q_{j+1,i} - \bar{q}_{j+1})}{\sqrt{\sum_i (q_{j,i} - \bar{q}_j) \sum_i (q_{j+1,i} - \bar{q}_{j+1})}}$$

- (d) The slope of the regression equation relating the month's flow to flow in the preceding month:

$$b_j = r_j \frac{S_{j+1}}{S_j}$$

2. The model is then a set of 12 regression Eq. (30)

$$\hat{q}_{j+1,i} = \bar{q}_{j+1} + b_j(q_{j,i} - \bar{q}_j) + Z_{j+1,i} \cdot s_{j+1} \sqrt{(1 - r_j^2)}$$

where Z is a random normal deviate $N(0, 1)$

3. To generate a synthetic flow sequence, calculate (generate) a random number sequence $\{Z_1, Z_2, \dots\}$, and substitute in the model

4.3.3 ANMAX Model

Autoregressive moving-average models with exogenous inputs are denoted by ANMAX (p, q, b), which shows a model with p autoregressive terms (AR(p)), q moving-average terms (MA(q)), and b exogenous input terms as a linear combination of the last b terms of a known and external time series d_t (Bailite 1980). The model formulation is as follows:

$$y_t = \varepsilon_t + \sum_{i=1}^p \beta_i(p)y_{t-i} + \sum_{i=1}^q \theta_i\varepsilon_{t-i} + \sum_{i=1}^b \eta_id_{t-i} \quad (31)$$

where parameters η_1, \dots, η_b are related to the selected exogenous input. These models can be successfully utilized in cases where the historical data cannot completely cover the variations and behavior of the studied variables.

4.3.4 ARCH Model

Volatility (i.e., time-varying variance) clustering, in which large changes tend to follow large changes and small changes tend to follow small changes, has been well recognized in time series analysis. This phenomenon is called conditional heteroscedasticity, and can be modeled by ARCH-type (AutoRegressive Conditional Heteroscedasticity, ARCH) models, including the ARCH model introduced by Engle (1982) and their GARCH (the generalized ARCH) extension proposed by Bollerslev (1986). In these models, the key concept is the conditional variance, i.e., the variance conditional on the past.

4.3.5 Disaggregation Model

Disaggregation models are used to decompose time series into several subseries that are temporal or spatial fractions of the key time series. Valencia and Schaake (1973) and later extension by Mejia and Rousselle (1976) introduced the basic disaggregation model for temporal disaggregation of annual time series into seasonal time series. Disaggregation models of hydrologic time series are efficient techniques for cases where the preservation of statistical characteristics of both annual and seasonal scales is essential for the project under study. Most applications of disaggregation have been in the temporal domain, although some investigators have applied the same principle in the spatial domain.

4.4 Summary

Time series models are now a major tool in planning, operation, and decision making in hydrology and water resources. On one hand, time series models possess many appealing features. First, they can be used to model a time series without considering

its physical nature. Second, they can be used to extrapolate past patterns of behavior into the future. They allow a researcher, who has data only in past years, to forecast future events without having to search for other related time series data. Third, the time series approach also allows for the use of one time series to explain the behavior of another series, if the other time series data are correlated with a variable of interest and if there appears to be some cause for this correlation. On the other hand, some time series models, like ARIMA, are complex techniques, and require a great deal of experience and data. Although they often produce satisfactory results, those results depend on the researcher's level of expertise.

Machiwal and Jha (2006) reviewed both theoretical and applied research of time series models in the hydrological science. In hydrologic studies, time series models have been widely applied for detecting climatic changes (e.g., Kite 1989), investigating the long-term hydroclimatological trends (Lachtermacher and Fuller 1994), exploring the possible impact of climate change on hydrologic variables or water resources (Westmacott and Burn 1997), modeling precipitation (Janos et al. 1988), evapotranspiration (Mohan and Arumugam 1995), streamflow (Moatmari et al. 1999; Pekarova and Pekar 2006; Shao et al. 2009), groundwater (Houston 1983; Van Geer and Zuur 1997), drought (Mishra and Desai 2005), water quality (Ahmad et al. 2001), and water demand and consumption (Bougadis et al. 2005; Jorge 2007), etc.

5 Artificial Neural Network (ANN) Models

An artificial neural network (ANN) is a biologically inspired distributed computing processor system in parallel with certain performance characteristics resembling biological neural networks of the human brain, which differs from conventional computers in the way they process information (Haykin 1994). It has a distributed processing structure (Alp and Cigizoglu 2007) and consists of processing elements and connections between them with coefficients bound to the connections. Mathematically, ANNs may be treated as a universal approximator. They are able to extract the relation between inputs and outputs of a process without the physics being explicitly provided to them and to generalize the structure hidden within the whole dataset. ANN models are able to simulate nonlinear relationships through an automatic “training process” (Hsu et al. 1997). The ANN models have no limitations in the form of fixed assumptions or formal constraints and are faster compared with its conventional simulation methods, robust in noisy environments, flexible in many problems, and highly adaptive to the newer environments (Jain et al. 1999). There have been many standard ANN software that can be used to pursue intricate multipurpose nonlinear solutions.

5.1 Structure of ANN

ANNs are a computational model and have been developed as a generalization of mathematical models of human cognition or neural biology. An ANN is based on the following rules:

- Information processing occurs at many single elements called nodes, also referred to as units, cells, or neurons
- Signals are passed between nodes through connection links
- Each connection link has an associated weight that represents its connection strength
- Each node typically applies a nonlinear transformation called an activation function to its net input to determine its output signal

According to the absence or presence of feedback connections in a network, two types of architecture are distinguished: feedforward architecture and feedback architecture. A typical feedforward multilayer artificial neural network with a single hidden layer is illustrated in Fig. 6 (Friedman and Kandel 1999; Xiong et al. 2004).

This kind of ANNs can solve a wide variety of problems, such as classifying patterns, storing and recalling data, performing general mapping from input pattern (space) to output pattern (space), grouping similar patterns, or finding solutions to constrained optimization problems. It consists of input nodes $\{X_i(p)\}_{i=1}^n$ (and one input to the neuron, called a bias, has a constant value of 1 and is usually represented as a separate input), hidden nodes $\{Z_j(p)\}_{j=1}^l$ (and a bias), and output nodes $\{Y_k(p)\}_{k=1}^m$, where X , Z , and Y represent the input, hidden, and output layers, respectively, n , l , m represent the number of the nodes in each layer, and p denotes the training pattern. The weights associated with the connections between input and hidden nodes are denoted by v_{ij} , $0 \leq i \leq n$, $1 \leq j \leq l$. Those between the hidden and the output nodes are denoted by w_{jk} , $0 \leq j \leq l$, $1 \leq k \leq m$.

For node Z_j in the hidden layer (Fig. 6), its effective aggregated input signal, denoted by z_in_j , is calculated as:

$$z_in_j = v_{0j} + \sum_{i=1}^n v_{ij}x_i, \quad 1 \leq j \leq l \quad (32)$$

where x_i , $1 \leq i \leq n$ represents the input to each node in the input layer.

For node Z_j , its corresponding output signal, denoted by z_j , is obtained by using an activation function $f(x)$

$$z_j = f(z_in_j), \quad 1 \leq j \leq l \quad (33)$$

The most widely used activation function is the sigmoid function (Friedman and Kandel 1999). The sigmoid function is a bounded, monotonic, nondecreasing function that provides a graded and nonlinear response. Among several different sigmoid functions, the one most often used for ANNs is the logistic function

$$z_j = f(z_in_j) = \frac{1}{1 + e^{-\sigma z_in_j}} \quad (34)$$

where σ is an adjustable parameter used in the activation function $f(x)$. This function enables a network to map any nonlinear process.

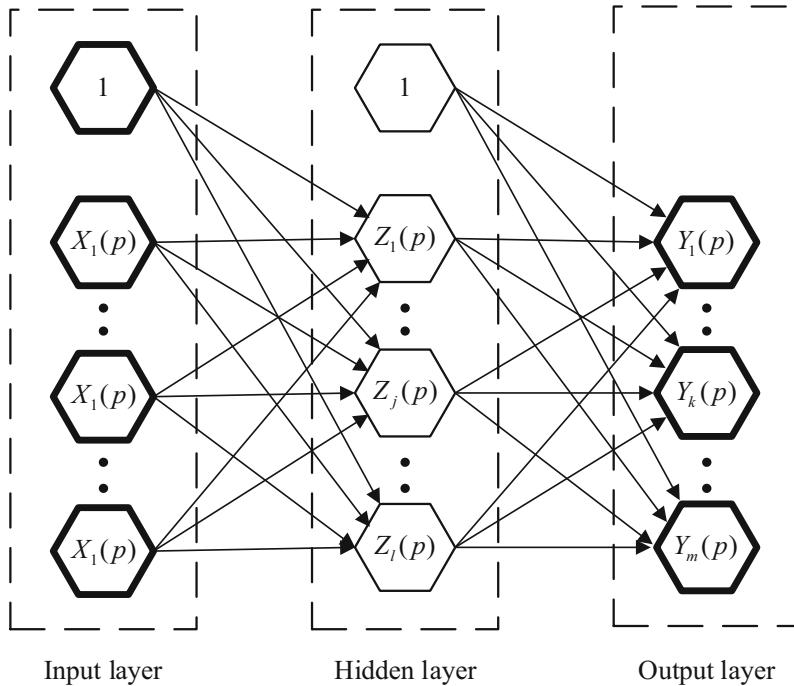


Fig. 6 A feedforward multilayer neural network with a single hidden layer (modified from Friedman and Kandel, 1999)

5.2 Network Training

In order for an ANN to generate an output vector $Y = (y_1, y_2, \dots, y_p)$ that is as close as possible to the target vector $T = (t_1, t_2, \dots, t_p)$, a training process, also called learning, is employed to find optimal weight matrices v_{ij} and w_{jk} , which minimize a predetermined error function that usually has the form:

$$E = \sum_p \sum_l (y_i - t_i)^2 \quad (35)$$

Here, t_i is a component of the desired output T ; y_i is the corresponding ANN output; l is the number of output nodes; and p is the number of training patterns. Training is a process by which the connection weights of an ANN are adapted through a continuous process of stimulation by the environment in which the network is embedded. The learning ability of a neural network is achieved by applying a learning (training) algorithm.

Training algorithms are mainly classified into three groups (Kasabov 1996):

- (1) Supervised. The training examples comprise input vectors x and the desired output vectors y . Training is performed until the neural network “learns” to

associate each input vector x to its corresponding and desired output vector y ; for example, a neural network can learn to approximate a function $y = f(x)$ represented by a set of training examples (x, y) . It encodes the examples in its internal structure.

- (2) Unsupervised. Only input vectors x are supplied; the neural network learns some internal features of the whole set of all the input vectors presented to it.
- (3) Reinforcement learning, sometimes called reward-penalty learning, is a combination of the above two paradigms; it is based on presenting input vector x to a neural network and looking at the output vector calculated by the network. If it is considered “good,” then a “reward” is given to the network in the sense that the existing connection weights are increased; otherwise, the network is “punished,” the connection weights, being considered as “not appropriately set,” decrease. Thus, reinforcement learning is learning with a critic, as opposed to learning with a teacher.

Learning is not an individual ability of a single neuron. It is a collective process of the whole neural network and a result of the training procedure. The connection weight matrix W has its meaning as a global pattern. It represents “knowledge” in its entirety. We do not know exactly how learning is achieved in the human brain. But learning (supervised or unsupervised) can be achieved in an artificial neural network. And there are some genetic laws of learning which have been discovered and implemented.

After training has been accomplished, it is hoped that the ANN will then be capable of generating reasonable results given new inputs. In contrast, an unsupervised training algorithm does not involve a teacher. During training, only an input data set is provided to the ANN that automatically adapts its connection weights to cluster those input patterns into classes that have similar properties. There are occasions when a combination of these two training strategies leads to reinforcement learning. A score or grade is used to rate the network performance over a series of training patterns. Most hydrologic applications have utilized supervised training. The manner in which the nodes of an ANN are structured is closely related to the algorithm that is used to train it.

5.3 Summary

ANNs have been utilized in many hydrologic problems and to evaluate if indeed all the strengths of ANNs have been effectively utilized in these applications. These applications include streamflow simulation (Shamseldin 1997; Hsu et al. 1995; Kişi, 2007; Wu and Chau 2011; He et al. 2014; Abrahart and See 2000; Aziz et al. 2014), water quality modeling (Rogers and Dowla 1994; Abyaneh 2014), ground water modeling (Aziz et al. 1992; Daliakopoulos et al. 2005), reservoir operation (Raman and Chandramouli 1996), water resources allocation and management (Raman and Sunilkumar 1995), evaporation estimation (Kumar et al. 2002; Shiri et al. 2014), hydrograph generation from hydrometeorological parameters (Ahmad

and Simonovic 2005), impact of climatic variations on flow discharge and dissolved organic carbon and nitrogen contents (Clair and Ehrman 1998), etc.

Zealand et al. (1999) claim that ANNs have the following beneficial model characteristics: (1) They infer solutions from data without prior knowledge of the regularities in the data. (2) ANNs are able to adapt to solutions over time to compensate for changing circumstances. (3) ANNs can generalize from previous examples to new ones, which is useful because real-world data are noisy, distorted, and often incomplete. (4) ANNs are also good at the abstraction of essential characteristics from inputs containing irrelevant data. (5) They are nonlinear, i.e., they can solve some complex problems more accurately than do linear techniques. (6) ANNs are highly parallel, containing many identical, independent operations that can be executed simultaneously, often making them faster than alternative methods.

However, ANNs also have several drawbacks for some applications. (1) Most of the ANN applications have been unable to explain the basic process in a comprehensibly meaningful way by which ANNs arrive at a decision. (2) When there is no learnable function or the data set is insufficient in size, they may fail to produce a satisfactory solution. (3) The optimum network geometry as well as the optimum internal network parameters are problem dependent and generally have to be found using a trial-and-error process. (4) The performance of an ANN deteriorates rapidly when the input vectors are far from the space of inputs used for training. (5) ANNs cannot cope with major changes in the system, because they are trained on historical data sets.

6 Fuzzy Logic Models

Fuzzy logic models are based on fuzzy logic system (Kruse et al. 1994; Klir and Yuan 1995; Kasabov 1996; Zimmermann 2001). Fuzzy logic systems, or fuzzy systems, are knowledge-based or rule-based systems. A fuzzy system is constructed from a set of fuzzy IF-THEN rules. A fuzzy IF-THEN rule is an IF-THEN statement in which some words are characterized by continuous membership functions (Wang 1997). For example, the following is an IF-THEN rule:

IF x is A , THEN y is B

The functioning of fuzzy systems is based on fuzzy sets theory (Zadeh 1965). The fuzzy sets theory, as an extension of the classical sets theory, are generally used to describe imprecision or vagueness. By translation into fuzzy IF-THEN rules, subjective knowledge can be incorporated in fuzzy logic systems in a natural and transparent way. Furthermore, the major strength of fuzzy logic systems resides in their ability to infer the behaviour of complex systems purely from data (data-driven), but still providing some insight about their internal operation. Finally, fuzzy systems are flexible modeling tools, as their architecture and the inference mechanisms can be adapted to the given modeling problem.

Fuzzy logic models consist of three steps, taking inputs, applying fuzzy rules, and producing outputs. Inputs to a fuzzy system can be either exact, crisp values, or fuzzy values. Output values from a fuzzy system can be fuzzy or exact (crisp). The process of transforming a single crisp value into a fuzzy value is called fuzzification, while the process of transforming a fuzzy value into a single crisp value is called defuzzification.

6.1 Basic Concepts of Fuzzy Systems

6.1.1 Fuzzy Sets and Membership Functions

Fuzzy sets, which may be generally used to describe imprecision or vagueness, have firstly been introduced by Zadeh (1965). Fuzzy sets are sets of objects without clear boundaries; in contrast with ordinary sets where for each object it can be decided whether it belongs to the set or not, a partial membership in a fuzzy set is possible. The traditional way of representing elements x of a set A is through the characteristic function:

$$\begin{aligned}\mu_A(x) &= 1, \text{ if } x \text{ is an element of set } A, \text{ and} \\ \mu_A(x) &= 0, \text{ if } x \text{ is not an element of set } A,\end{aligned}$$

that is, an object either belongs or does not belong to a given set.

An object can belong to a set partially in fuzzy sets. The degree of membership is defined through a generalized characteristic function called membership function:

$$\mu_A(x) : U \rightarrow [0, 1]$$

where U is called the universe and A is a fuzzy subset of U .

The values of the membership function are real numbers in the interval $[0, 1]$, where 0 means that the object is not a member of the set and 1 means that it belongs entirely to the set. Each value of the function is called a membership degree. One way of defining a membership function is through an analog function.

Fig. 7 (Kasabov 1996) shows three membership functions representing three fuzzy sets labeled as “short,” “medium,” and “tall,” all of them being fuzzy values of a variable “height.” As we can see, the value 174 cm belongs to the fuzzy set “medium” to a degree of 0.6 and at the same time to the set “tall” to a degree of 0.4.

To take another familiar example (Bárdossy et al. 1995), the set of young persons is fuzzy, as there is no generally accepted boundary between young and not young. The membership function of this set A may be defined as

$$\mu_A(x) = \begin{cases} 1 & \text{if } x \leq 25 \\ \frac{40 - x}{15} & \text{if } 25 < x \leq 40 \\ 0 & \text{if } x > 40 \end{cases}$$

Fuzzy set theory can be considered as an extension of ordinary set theory; compared to the classical set theory, fuzzy set theory is very flexible in describing

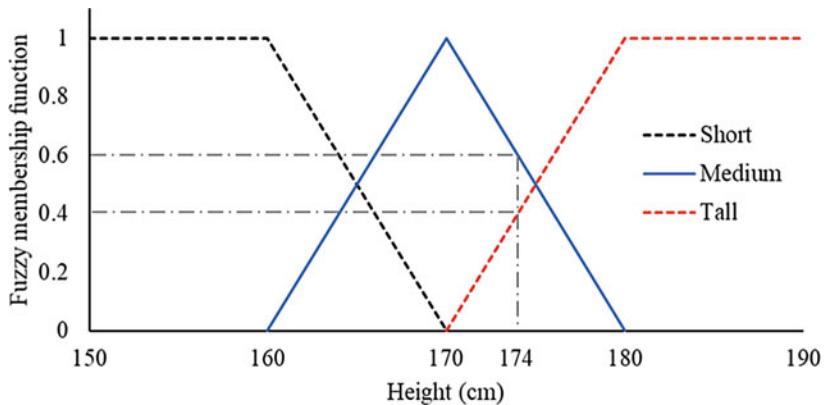


Fig. 7 Membership functions of representative three fuzzy sets for the variable “height” (modified from Kasabov, 1996)

the features of objects; it has advantages in expressing vague, uncertain, and imprecise information, which appears frequently in scientific and engineering fields (Zadeh 1965; Zimmermann 2001).

6.1.2 Fuzzy Rules

A fuzzy rule consists of a set of arguments in the form of fuzzy sets with membership functions and a response also in the form of a fuzzy set. For a general input vector, the rule is applied as:

$$\text{If } a_1 \text{ is } A_{i,1} \odot a_2 \text{ is } A_{i,2} \odot \dots \odot a_k \text{ is } A_{i,k}, \text{ then } B_i$$

where \odot is any logical operator, specified according to the application. Usually rules are formulated using AND/OR operators. For example, in modeling moisture movement in an unsaturated zone (Dou et al. 1999), it may consist of two premises (i.e. $k = 2$). $A_{i,1}$ may correspond to a class of upper layer moisture content (e.g., low, medium, high), and $A_{i,2}$ to a class of lower layer moisture content (e.g., very low, medium, high, saturated), and the response B_i may be the actual quantity of water flux between the two layers.

6.2 Operations with Fuzzy Sets

Fuzzy set theory can be considered as an extension of ordinary set theory, i.e., the classical sets are a special case of fuzzy sets, when two membership degrees only, 0 and 1, are used, and crisp borders between the sets are defined. The following operations over two fuzzy sets A and B defined over the same universe U are the most common in fuzzy theory (Zadeh 1965; Zimmermann 2001).

Containment, $A \subset B$

A is contained in B (or, equivalently, A is a subset of B , or A is smaller than or equal to B) if and only if $\mu_A \leq \mu_B$. In symbols

$$A \subset B \Leftrightarrow \mu_A \leq \mu_B$$

Intersection, $A \cap B$

The intersection of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , written as $C = A \cap B$, whose membership function is related to that of A and B by

$$\mu_C(x) = \min[\mu_A(x), \mu_B(x)], \quad x \in U$$

or in abbreviated form

$$\mu_C(x) = \mu_A(x) \wedge \mu_B(x), \quad x \in U$$

Union, $A \cup B$

The union of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , written as $C = A \cup B$, whose membership function is related to that of A and B by

$$\mu_C(x) = \max[\mu_A(x), \mu_B(x)], \quad x \in U$$

or in abbreviated form

$$\mu_C(x) = \mu_A(x) \vee \mu_B(x), \quad x \in U$$

Equality, $A = B$

Two fuzzy sets A and B are equal, written as $A = B$, if and only if

$$\mu_A(x) = \mu_B(x), \quad x \in U$$

Complement

The complement of a fuzzy set A is denoted by A' and is defined by

$$\mu_{A'}(x) = 1 - \mu_A(x), \quad x \in U$$

Concentration, $\text{CON}(A)$

$$\mu_{\text{CON}(A)}(x) = (\mu_A(x))^2, \quad x \in U$$

this operation is used as a linguistic modifier “very”

Dilation, $\text{DIL}(A)$

$$\mu_{\text{DIL}(A)}(x) = (\mu_A(x))^{0.5}, \quad x \in U$$

this operation is used as a linguistic modifier “more or less.”

Algebraic product, $A \cdot B$

The algebraic product of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , written as $C = A \cdot B$, whose membership function is related to those of A and B by:

$$\mu_C(x) = \mu_A(x) \cdot \mu_B(x), \quad x \in U$$

Algebraic sum, $A + B$

The algebraic sum of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , written as $C = A + B$, whose membership function is related to those of A and B by

$$\mu_C(x) = \mu_A(x) + \mu_B(x), \quad x \in U$$

The De Morgan laws are valid for the algebraic sum and difference.

Bounded product

The bounded product of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , whose membership function is related to those of A and B by

$$\mu_C(x) = \max[0, \mu_A(x) + \mu_B(x) - 1], \quad x \in U$$

Bounded sum

The bounded sum of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , whose membership function is related to those of A and B by

$$\mu_C(x) = \max[1, \mu_A(x) + \mu_B(x)], \quad x \in U$$

Bounded difference, $A/-B$

The bounded difference of two fuzzy sets A and B with respective membership functions $\mu_A(x)$ and $\mu_B(x)$ is a fuzzy set of C , written as $C = A/-B$, whose membership function is related to those of A and B by:

$$\mu_C(x) = \min[0, \mu_A(x) - \mu_B(x)], \quad x \in U$$

Normalization, $\text{NORM}(A)$

$$\mu_{\text{NORM}(A)}(x) = \mu_A(x) / \max\{\mu_A(x)\}, \quad x \in U$$

The operations over fuzzy sets have some properties, for example, they are associative, commutative, and distributive, that is,

Associative: $(a * b) * c = a * (b * c)$

Commutative: $a * b = b * a$ (not valid for the bounded difference)

Distributive: $a * (b * c) = (a * b) * (a * c)$

where $*$ and \star denote any operations from those listed above.

6.3 Types of Fuzzy Systems

To construct a fuzzy system, first we need to obtain a collection of fuzzy IF-THEN rules from human experts or based on domain knowledge. The next step is to combine these rules into a single system. Different fuzzy systems use different principles for this combination. There are three types of fuzzy systems that are commonly found in the literature: (i) pure fuzzy systems, (ii) Takagi-Sugeno-Kang (TSK) fuzzy systems, and (iii) fuzzy systems with fuzzifier and defuzzifier.

6.3.1 Pure Fuzzy Systems

The main feature with the pure fuzzy system is that its inputs and outputs are fuzzy sets, whereas the inputs and outputs are real-valued variables in engineering systems (Fig. 8) (Wang 1997).

Each single IF-THEN rule of pure fuzzy systems has the following general form:

Rule m : IF $(x_1 \text{ is } A_{1,m}) \text{ AND } (x_2 \text{ is } A_{2,m}) \text{ AND } \dots \text{ AND } (x_k \text{ is } A_{k,m})$ THEN y is... expressing the relation between k input variables x_1, x_2, \dots, x_k and output y . Terms $A_{k,m}$ in the antecedents of the rules (i.e., the IF part of the rules) represent fuzzy sets (Zadeh 1965) used to partition the input space into overlapping regions.

6.3.2 Takagi-Sugeno-Kang (TSK) Fuzzy Systems

In contrast to pure fuzzy systems, Takagi and Sugeno (1985) and Sugeno and Kang (1988) proposed another fuzzy system whose inputs and outputs are real-valued variables, named Takagi-Sugeno-Kang (TSK) fuzzy systems. The TSK fuzzy systems have the following general structure:

IF $(x_1 \text{ is } A_{1,m}) \text{ AND } (x_2 \text{ is } A_{2,m}) \text{ AND } \dots \text{ AND } (x_k \text{ is } A_{k,m})$ THEN $y = f_m(x_1, x_2, \dots, x_k)$

Each fuzzy rule in a TSK fuzzy inference system can be regarded as a local model of the system under consideration. The functions f_m are usually first-order polynomials, given by

$$f_m(x_1, x_2, \dots, x_k) = b_{0,m} + b_{1,m} \cdot x_1 + b_{2,m} \cdot x_2 + \dots + b_{k,m} \cdot x_k$$

Fig. 9 shows a schematic diagram of the functioning of a typical multiple-input single-output TSK fuzzy system (Jacquin and Shamseldin 2006). The first stage in the inference process of a TSK fuzzy model is the calculation of the degree of fulfilment (DOF) of each rule. The output of each rule is obtained by the evaluation of the corresponding function f_m . Finally, the overall fuzzy model response is obtained as the weighted average of the individual rule responses.

The degree of fulfilment of a rule evaluates the compatibility of a given input vector with the antecedent of the rule (i.e., the IF part). The degree of fulfilment is normally evaluated using a T-norm, such as the algebraic product:

$$\text{DOF}_m(x_1, x_2, \dots, x_k) = \mu_{A_{1,m}}(x_1) \cdot \mu_{A_{2,m}}(x_2) \cdot \dots \cdot \mu_{A_{k,m}}(x_k)$$

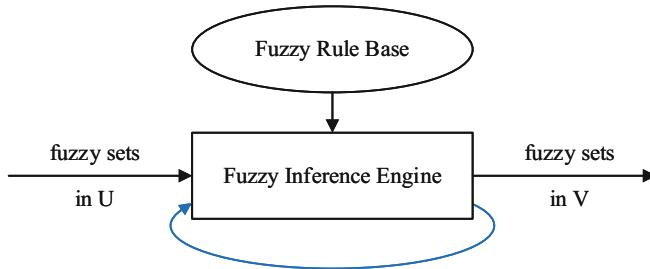


Fig. 8 Basic configuration of pure fuzzy systems (modified from Wang, 1997)

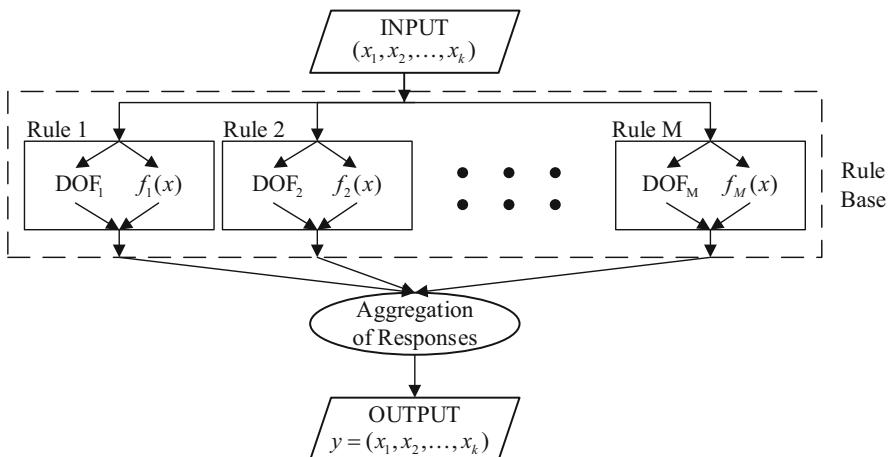


Fig. 9 Functioning of a multiple-input single-output Takagi-Sugeno-Kang fuzzy inference system (modified from Jacquin and Shamseldin 2006)

Several types of the membership functions can be used for the fuzzy sets in the antecedents of the rules (Zimmermann 2001; Piegat 2001). The Gaussian membership functions, which have the following analytical expression:

$$\mu_{k,m}(x) = \exp \left[-\frac{(x_k - c_{k,m})^2}{2\sigma_{k,m}^2} \right]$$

are a common choice (Chang et al. 2001; Gautam and Holz 2001; Xiong et al. 2001). In this case, each membership function has two parameters, namely the center $c_{k,m}$ and the spread $\sigma_{k,m}$.

Some problems with the Takagi-Sugeno-Kang fuzzy system are listed as follows: (i) its THEN part is a mathematical formula and therefore may not provide a natural framework to represent human knowledge, and (ii) there is not much freedom left to

apply different principles in fuzzy logic, so that the versatility of fuzzy systems is not well represented in this framework.

6.3.3 Fuzzy Systems with Fuzzifier and Defuzzifier

In order to use pure fuzzy systems for simulating engineering systems, a simple method is to add a fuzzifier, which transforms a real-valued variable into a fuzzy set, to the input, and a defuzzifier, which transforms a fuzzy set into a real-valued variable, to the output. Thus, we get a fuzzy system with fuzzifier and defuzzifier, as shown in Fig. 10 (Wang 1997).

6.4 Adaptive Neuro-Fuzzy Inference System (ANFIS)

During the past four decades, significant progress has been made in the two artificial intelligence techniques, i.e., fuzzy inference system (FIS) and artificial neural networks (ANNs). A judicious integration of FIS and ANN can produce a functional neural fuzzy system capable of learning, high-level thinking, and reasoning (Jang et al. 1997; Loukas 2001). It provides an effective approach for dealing with large imprecisely defined complex systems. An ANFIS works by applying neural learning rules to identify and tune the parameters and structure of an FIS.

A typical architecture of an ANFIS, in which a circle indicates a fixed node, whereas a square indicates an adaptive node, is shown in Fig. 11 (Jang et al. 1997). For simplicity, we assume that the examined FIS has two inputs and one output.

In Fig. 11, x and y are the two crisp inputs and A_i and B_i are the linguistic labels associated with the node function. For rainfall-runoff modeling in hydrology, the input and output nodes represent rainfall process and discharge observations, respectively.

The attractive features of an ANFIS include: easy to implement, fast and accurate learning, strong generalization abilities, excellent explanation facilities through fuzzy rules, and easy to incorporate both linguistic and numeric knowledge for problem solving (Jang et al. 1997). Due to these fascinating features of the ANFIS, it is widely used in hydrological science.

6.5 Summary

An important contribution of fuzzy system theory is that it provides a systematic procedure for transforming a knowledge base into a nonlinear mapping. On one hand, fuzzy systems are multi-input-single-output mappings from a real-valued vector to a real-valued scalar (a multi-output mapping can be decomposed into a collection of single-output mappings), and the precise mathematical formulas of these mappings can be obtained; on the other hand, fuzzy systems are knowledge-based systems constructed from human knowledge in the form of fuzzy IF-THEN rules.

The fields of hydrology and water resources commonly involve a system of concepts, principles, and methods for dealing with modes of reasoning that are

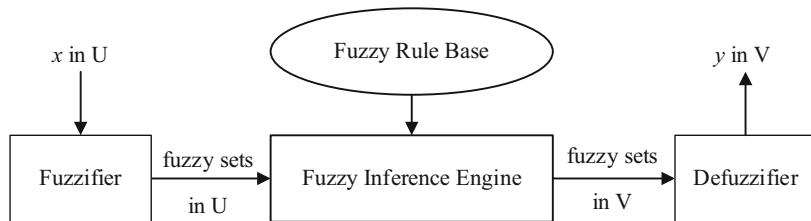
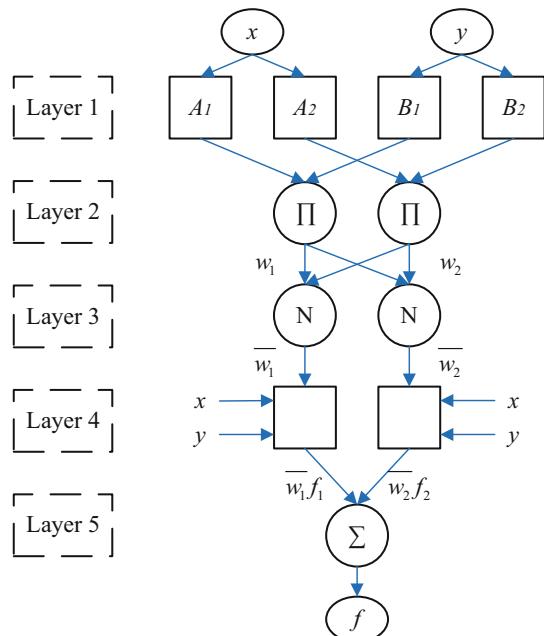


Fig. 10 Basic configuration of fuzzy systems with fuzzifier and defuzzifier (modified from Wang, 1997)

Fig. 11 Architecture of the ANFIS (modified from Jang et al., 1997)



approximate rather than exact. The capability of dealing with imprecision gives fuzzy logic great potential for hydrological analysis and water resources decision making.

In hydrology, the concept of fuzzy theory and its application have found many applications in a number of research areas, such as groundwater flow in the unsaturated zone (Bárdossy and Disse 1993; Bárdossy et al. 1995; Dou et al. 1995; Schulz and Huwe 1997; Dou et al. 1999; Hong et al. 2002; Afshar et al. 2007), the interdependence between global circulation and rainfall (Pongracz et al. 2001), reconstruction of missing precipitation events (Abebe et al. 2000; Coulibaly and Evora 2007), rainfall-runoff modeling (Yu and Yang 2000; Hundecha et al. 2001; Huang et al. 2010), flood forecasting (See and Openshaw 1999; Xiong et al. 2001), flood frequency analysis (Shu and Burn 2004), reservoir

operation (Russell and Campbell 1996), water resources allocation and management (Yurdusev and Firat 2009), drought prediction (Pesti et al. 1996), and evaporation estimation (Cobaner 2011; Shiri et al. 2013, 2014).

The advantages of a fuzzy logic model include the following: (1) The model is not sensitive to parameter changes, and can be easily programmed, codes remain simple, short, and require little computer time. (2) The model is transparent and easy to understand due to its rule-based structure, which imitates the human way of thinking. (3) The fuzzy rule-based model can encode the expert's knowledge. (4) The most distinguishing property of fuzzy logic is that it deals with fuzzy propositions, that is, propositions which contain fuzzy variables and fuzzy values; thus, the fuzzy systems are especially good at dealing with nonlinear relationships. However, it must be reminded that fuzzy logic models, just like other types of the black-box models, can only describe the input-output relationships without explicit consideration of the internal hydrologic processes that lead to this transformation.

7 Frequency Analysis Models

Flood frequency estimation has been fundamental in engineering hydrology since Fuller (1914) approached the temporal variability of flood flows of extremely high return periods. The primary objective of frequency analysis is to relate the magnitude of extreme events to their frequency of occurrence through the use of probability distributions (Chow et al. 1988). The purpose of frequency analysis is to analyze past records of hydrologic variables so as to estimate future occurrence probabilities of extreme events.

The hydrologic data analyzed in frequency analysis are assumed to be independent and identically distributed (i.e., the i.i.d. assumption), if the hydrologic system producing them (e.g., a storm rain system) is considered to be stochastic, space-independent, and time-independent. The hydrologic data employed in frequency analysis should be carefully selected so that the assumptions of independence and identical distribution are satisfied. The data used in the analysis must also be evaluated in terms of the objectives, length of records available, and completeness of records. A frequency analysis can be performed using single-site data, regional data, or both. It can also include historical information and reflect physical constraints.

7.1 Graphical Method

Plotting position refers to the probability value assigned to each value in a random sample. It is used to calculate and graphically display the empirical frequency curve by plotting each ranked value against a probability scale. Numerous methods have been proposed for the determination of plotting positions, most of which are empirical. If n is the total number of values to be plotted and m is the rank of a value in a list ordered by descending magnitude, the exceedance probability of the m^{th} largest value is x_m ; the general form of plotting positions can be written:

$$P(X \geq x_m) = \frac{m - b}{n + 1 - 2b} \quad (36)$$

where b is a parameter commonly varying between 0 and 0.5 for various formulae. For example, $b = 0.5$ for Hazen's formula, $b = 0.3$ for Chegodayev's formula. The most popular plotting position formula is the Weibull plotting position when $b = 0$, which is an expected (unbiased) form of the exceedance probability at the m th largest observation for all distributions.

The procedure of graphical method includes the following:

- Select a Q_{\max} value from each year
- Arrange the data in a decreasing order, i.e., $Q_1 \geq Q_2 \geq Q_3 \dots$
- Assign a frequency/probability of exceedance to each Q_i . The most common method is the Weibull formula:

$$P(Q > Q_m) = \frac{m}{n + 1} \quad (37)$$

where n is the total number of data and m is the order of Q . This means that $m(Q_{\max}) = 1$, and $m(Q_{\min}) = n$.

- Plot Q versus $P(Q > Q_m)$ or plot Q versus $T = 1/P(Q > Q_m)$, where T is return period
- On millimeter paper (without distribution assumption) – fit a curve
- On probability paper (with distribution assumption) – fit a straight line if the data fit the probability distribution as the probability paper presents
- Knowing the probability of $P(Q > Q_T)$ or T , the design flow Q_T can be read from the plot or knowing the magnitude of Q_T , we can read $P(Q > Q_T)$ or T from the plot

7.2 Analytical Method – Frequency Factor Method

A random variable X can be decomposed into two components and written as:

$$X = \bar{x} + \Delta X \quad (38)$$

where ΔX is the deviation from the mean, \bar{x} . A new quantity can be defined as: $K = \Delta X/s$, where s is the standard deviation of the data, and the former equation can be rewritten as:

$$X = \bar{x} + sK \quad (39)$$

For a design value X_T with return period of T , Eq. (39) can be written as

$$X_T = \bar{x} + K_T s \text{ or } X_T = \bar{x}(1 + C_v K_T) \quad (40)$$

where C_v is the coefficient of variation, and K_T is the *frequency factor* depending on the *probability distribution* being used and on the *return period*, T .

Equation (40) is the working equation for frequency analysis, which can be used to calculate the design value X_T for given design level T and probability distribution. It can also be used to estimate the return period of a given X value.

Examples of design flow calculation for different probability distributions are presented below.

7.2.1 Example for Normal Distribution

If X is normally distributed, i.e., $X \sim N(\bar{x}, s^2)$, from $X_T = \bar{x} + K_T s$ we get

$$K_T = \frac{X_T - \bar{x}}{s}$$

That means K_T is the standardized normal variate Z ($Z = \frac{X-\mu}{\sigma}$), i.e., $K_T = Z \sim N(0, 1)$. K_T can then be read from the standard normal distribution table for a given T or calculated from the related equation by numerical methods.

7.2.2 Example for Lognormal Distribution

X is said to be lognormally distributed if $Y = \ln(X)$ is normally distributed with mean μ_Y and variance σ_Y^2 .

The procedure for calculating X_T from log-normal distribution is:

- Let $y_i = \ln x_i$ for all x_i .
- Calculate $\bar{y} = \frac{1}{n} \sum y_i$ and $s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n-1}}$
- Read K_T from normal distribution table for a given T
- Calculate $Y_T = \bar{y} + K_T s_y$
- Calculate $X_T = e^{Y_T}$

It should be noted that if only the mean and standard deviation (\bar{x}, s_x) of a lognormally distributed variable X are available, then the mean and standard deviation (\bar{y}, s_y) of the associated normally distributed variable $Y = \ln(X)$ are calculated as:

$$\bar{y} = \ln \left(\frac{\bar{x}^2}{\sqrt{s_x^2 + \bar{x}^2}} \right), s_y = \sqrt{\ln \left(\frac{s_x^2}{\bar{x}^2} + 1 \right)}$$

7.2.3 Extreme Value Type I Distribution

The K_T value can either be calculated by using the equation below or read from the extreme value type I distribution table.

$$\begin{aligned}
K_T &= -\frac{\sqrt{6}}{\pi} \left\{ 0.5772 + \ln \left[\ln \left(\frac{T}{T-1} \right) \right] \right\}, \Rightarrow T \\
&= \frac{1}{1 - \exp \left\{ -\exp \left[-\left(0.5572 + \frac{\pi K_t}{\sqrt{6}} \right) \right] \right\}}
\end{aligned} \tag{41}$$

The design flow X_T can then be calculated using Eq. (40).

7.2.4 Pearson Type III Distribution

The Pearson type III distribution has three parameters, λ , β , and ε , which can be estimated through the calculation of mean, standard deviation, and coefficient of skewness.

The procedure for the Pearson type III distribution is described as follows:

- Compute the mean, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \rightarrow \lambda$
- Compute the standard deviation, $s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}} \rightarrow \beta$
- Compute the coefficient of skewness, $C_s = \frac{n \sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3} \rightarrow \varepsilon$
- Compute K_T by Eq. (42) or read from the Table of K_T values for P-III distribution

$$K_T = z + (z^2 - 1)k + \frac{1}{3}(z^3 - 6z)k^2 - (z^2 - 1)k^3 + zk^4 + \frac{1}{3}k^5 \tag{42}$$

where $K = C_s/6$

$$z = w - \frac{2.516 + 0.8029w + 0.01033w^2}{1 + 1.4328w + 0.1893w^2 + 0.00131w^3}$$

$$w = \left[\ln \left(\frac{1}{p^2} \right) \right]^{1/2}$$

$$p = \frac{1}{T}$$

- Compute $x_T = \bar{x} + K_T s$

7.2.5 Log-Pearson Type III Distribution

Similar to the lognormal distribution, the procedure for log-Pearson type-III distribution is as follows:

- Transform X to $Y = \log(X)$ or $\ln(X)$
- Compute the mean, y

- Compute the standard deviation, s_y
- Compute the coefficient of skewness, C_s
- Compute K_T by Eq. (42) or read from the Table of K_T values for P-III distribution
- Compute $y_T = \bar{y} + K_T s$
- Compute $x_T = 10^{y_T}$ or $x_T = e^{y_T}$

7.3 Data Sampling Methods

Data for frequency analysis studies can be compiled in several ways. In general, there are three data sampling methods: Annual Maximum Series, Partial Duration Series, and Annual Exceedance Series.

Annual Maximum Series: The Annual Maximum Series (AMS) data consist of the largest event in each year, regardless of whether the second largest event in a year exceeds the largest events of other years. An objection to using the AMS is that, in many cases, the second largest event in a year exceeds the largest event of other years.

Partial Duration Series: A Partial Duration Series (PDS) is a series of data which are selected so that their magnitude is greater than a predefined base value (Chow et al. 1988). Partial duration series or **peaks-over-threshold** (POT) pick all peaks above a threshold.

Annual Exceedance Series: If the base value of the PDS is selected so that the number of values in the series is equal to the number of years of record, the series is called an Annual Exceedance Series (AES) (Chow et al. 1988). An AES may be regarded as a special case of the PDS. Although AES is useful for some purposes, it may be difficult to verify that all the observations are independent.

The return period T_E of event magnitudes developed from an AES is related to the corresponding return period T for magnitudes derived from an AMS by (Chow 1964)

$$T_E = \left[\ln\left(\frac{T}{T-1}\right) \right]^{-1} \quad (43)$$

7.4 Outliers and Zeros

Outliers are data points that depart significantly from the trend of the remaining data. The retention or deletion of these outliers can significantly affect the magnitude of statistical parameters computed from the data. The U.S. Water Resources Council (1982) Bulletin 17B suggests that outliers can be identified from

$$X_H = \bar{X} + K_n S_X$$

$$X_L = \bar{X} - K_n S_X$$

where X_H and X_L are the threshold values for high and low outliers, and K_n can be approximated from

$$K_n \approx 1.055 + 0.981 \log_{10} n$$

where n is the number of observations.

The detailed description of the treatment of outliers is contained in the U.S. Water Resources Council (1982) Bulletin 17B.

Treatment of zeros: Most hydrologic variables are bounded on the left by zero. Zero values should not simply be ignored, nor do they necessarily reflect inaccurate measurements of the minimum flow in a channel. A zero in a set of data that is being logarithmically transformed requires special handling. One solution is to add a small constant to all of the observations. Another method is to analyze the nonzero values and then adjust the relation to the full period of record. This method biases the results as the zero values are essentially ignored. A third and theoretically more sound method would be to use the theorem of total probability (For details see Haan (2002, pp168–169)).

7.5 Regionalization

Two broad categories of regionalization procedures have been widely used in the field of frequency analysis: the index-flood approach (Dalrymple 1960) and the multiple-regression approach (Benson 1962).

7.5.1 Index-Flood Method

The key assumption of an index-flood procedure is that the stations form a homogeneous region, meaning that the frequency distributions of the N stations are identical apart from a site-specific scaling factor, the index flood. We may then write

$$Q_i(F) = \mu_i q(F), \quad i = 1, 2, \dots, N \quad (44)$$

where $Q_i(F)$, $0 < F < 1$, is the quantile function of the frequency distribution at site i ; μ_i is the index flood (Hosking and Wallis 1997); N is the number of sites; $q(F)$ is the regional growth curve, a dimensionless quantile function common to every site.

The index flood is estimated by $\hat{\mu}_i = \bar{Q}_i$, the sample mean of the data at site i , and the dimensionless rescaled data are computed by $q_{ij} = Q_{ij}/\hat{\mu}_i$, where Q_{ij} is the observed data at site i , $j = 1, 2, \dots, n_i$, and n_i is the sample size at site i .

Hosking and Wallis (1997) suggested an index-flood method where parameters are estimated separately at each site. They considered the use of a weighted average of the at-site estimates:

$$\hat{\theta}_k^R = \sum_{i=1}^N n_i \hat{\theta}_k^{(i)} / \sum_{i=1}^N n_i \quad (45)$$

where $\hat{\theta}_k^{(i)}$ stands for the L-moment of interest. The estimated regional quantile $\hat{q}(F) = q(F; \hat{\theta}_1^R, \dots, \hat{\theta}_P^R)$ is obtained by substituting the estimates $\hat{\theta}_k^{(i)}$ into $q(F)$ (Hosking

and Wallis 1993). The quantile estimates at site i can be obtained using the estimates of μ_i and $q(F)$.

$$\hat{Q}_i(F) = \hat{\mu}_i \hat{q}(F) \quad (46)$$

This index-flood procedure makes the following assumptions: (i) observations at any given site are identically distributed, and independent both serially and spatially; (ii) frequency distributions at different sites are identical apart from a scale factor; and (iii) the mathematical form of the regional growth curve is correctly specified.

7.5.2 Regional Regression

Regional regression models have long been used to predict flood quantiles at ungauged sites, and in a nationwide test, this method did as well as or better than more complex rainfall-runoff modeling in predicting flood quantiles (Newton and Herrin 1982).

Consider the traditional log-linear model for a statistic y_i which is to be estimated by using watershed characteristics such as drainage area and slope:

$$y_i = \alpha + \beta_1 \log(\text{area}) + \beta_2 \log(\text{slope}) + \dots + \varepsilon \quad (47)$$

A challenge in analyzing this model and estimating its parameters with available records is that one only obtains sample estimates, denoted \hat{y}_i , of the hydrologic statistics y_i . Thus, the observed error ε is a combination of (1) the time-sampling error in sample estimators of y_i (these errors at different sites can be cross-correlated if the records are concurrent) and (2) underlying model error (lack of fit) due to the failure of the model to exactly predict the true value of the y_i 's at every site. Often these problems have been ignored and standard ordinary least squares (OLS) regression has been employed (Thomas and Benson 1970).

7.6 Summary

Frequency analysis has been one of the earliest and most frequent uses of statistical methods in hydrology. In the earlier years (before 1960s), frequency analysis was mainly used for flood flow estimation, and nowadays, frequency analysis has been applied to almost every hydrological extreme variable, such as floods (Vogel et al. 1993; Vogel and Wilson 1996), low flows (Nathan and McMahon 1990; Lawal and Watt 1996; Durrans and Tomic 2001), rainfall events of various kinds (Pilgrim 1998; Öztekin 2007; Stedinger et al. 1993), droughts and dry spells (Lana and Burgueno 1998; Lana et al. 2008; Hallack-Alegria and Watkins 2007), etc. Hydrological frequency analysis (HFA) has been playing an essential role in the planning, design, and management of projects for flood control and water usages.

As to regional frequency analysis, recent advances mainly refer to the use of L-moments together with the index-flood method, as reported by Hosking and Wallis (1997). This methodology has been applied in modeling floods, rainfall extremes,

and low flows (Hosking et al. 1985; Vogel and Wilson 1996; Kjeldsen et al. 2001; Kumar et al. 2003; Yue and Wang 2004; Lim and Voeller 2009; Saf 2009).

The basic assumption of traditional HFA methods (both for one individual site and for a region) is that the hydrological data used are stationary, independent, and identically distributed over time. However, in the past decades, this stationarity assumption has been severely challenged because global climate change and/or large-scale human activities have altered the statistical characteristics of hydrological processes. Nonstationary frequency analysis is now a relatively new modeling approach and the number of studies is continuously increasing (Khaliq et al. 2006).

Studies of flood frequency analysis under nonstationary conditions have mostly assumed trends in time (Strupczewski et al. 2001; Renard et al. 2006; Leclerc and Ouarda 2007). In the last decade, some researchers have also explored the possibility of incorporating climate indices as external forcing into models for flood frequency analysis, assuming linear and nonlinear dependences (Sankarasubramanian and Lall 2003; El Aldouni et al. 2007; Ouarda and El-Aldouni 2011). Results have shown the feasibility of incorporating climate indices as covariates in the models, thus enabling the models to better describe changes in flood regimes over time.

Acknowledgment We are indebted to Yixing Yin, Yukun Hou, Qiang Zeng, and Xin-e Tao for their help in preparation of this chapter with proofreading and in supplying references, drawing figures, rewriting parts of the text, etc. We are also thankful to the two anonymous reviewers whose comments improved this chapter.

References

- A.J. Abebe, D.P. Solomatine, R.G.W. Venneker, Application of adaptive fuzzy rule-based models for reconstruction of missing precipitation events. *Hydrol. Sci. J.* **45**(3), 425–436 (2000)
- R.J. Abrahart, L. See, Comparing neural network and autoregressive moving average techniques for the provision of continuous river flow forecasts in two contrasting catchments. *Hydrol. Process.* **14**, 2157–2172 (2000)
- H.Z. Abyaneh, Evaluation of multivariate linear regression and artificial neural networks in prediction of water quality parameters. *J. Environ. Health Sci. Eng.* **12**, 40 (2014). <http://www.ijehse.com/content/12/1/40>
- J. Adamowski, H.F. Chan, S.O. Prasher, B. Ozga-Zielinski, A. Sliusarieva, Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban water demand forecasting in Montreal, Canada. *Water Resour. Res.* **48**(1) (2012). <https://doi.org/10.1029/2010WR009945>
- A. Afshar, M.A. Mariño, M. Ebtehaj, J. Moosavi, Rule-based fuzzy system for assessing ground-water vulnerability. *J. Environ. Eng.* **133**(5), 532–540 (2007)
- S. Ahmad, S.P. Simonovic, An artificial neural network model for generating hydrograph from hydro-meteorological parameters. *J. Hydrol.* **315**, 236–251 (2005)
- S. Ahmad, I.H. Khan, B.P. Parida, Performance of stochastic approaches for forecasting river water quality. *Water Res.* **35**(18), 4261–4266 (2001)
- S. Ali, N.C. Ghosh, R. Singh, Rainfall-runoff simulation using a normalized antecedent precipitation index. *Hydrol. Sci. J.* **55**(2), 266–274 (2010)
- M. Alp, H.K. Cigizoglu, Suspended sediment load simulation by two artificial neural network methods using hydrometeorological data. *Environ. Model Softw.* **22**, 2–13 (2007)

- A. Aziz, R. Abd, K.F.V. Wong, A neural-network approach to the determination of aquifer parameters. *Ground Water* **30**(2), 164–166 (1992)
- K. Aziz, A. Rahman, G. Fang, S. Shrestha, Application of artificial neural networks in regional flood frequency analysis: a case study for Australia. *Stoch. Env. Res. Risk A.* **28**(3), 541–554 (2014)
- R.T. Bailite, Predictions from ARMAX models. *J. Econ.* **12**, 365–374 (1980)
- A. Bárdossy, M. Disse, Fuzzy rule-based models for infiltration. *Water Resour. Res.* **29**(2), 373–382 (1993)
- A. Bárdossy, A. Bronstert, B. Merz, 1-, 2- and 3-dimensional modeling of water movement in the unsaturated soil matrix using a fuzzy approach. *Adv. Water Resour.* **18**(4), 237–251 (1995)
- J.J. Beauchamp, Comparison of regression and time-series methods for synthesizing missing streamflow records. *Water Resour. Bull.* **25**, 961–975 (1989)
- M.A. Benson, Factors influencing the occurrence of floods in a humid region of diverse terrain. U.S. Geol. Surv., Water-Supply Pap., 1580-B, 64 pp. (1962)
- R.B. Billings, D.E. Agthe, State-space versus multiple regression for forecasting urban water demand. *J. Water Resour. Plann. Manag.* **124**(2), 113 (1998)
- T. Bollerslev, Generalized autoregressive conditional heteroskedasticity. *J. Econometrics* **31**(3), 307–327 (1986)
- J. Bougadis, K. Adamowski, R. Diduch, Short-term municipal water demand forecasting. *Hydrol. Process.* **19**, 137–148 (2005). <https://doi.org/10.1002/hyp.5763>
- G.E.P. Box, G.M. Jenkins, *Time Series Analysis: Forecasting and Control* (Holden-Day, San Francisco, 1976)
- F.J. Chang, H.F. Hu, Y.C. Chen, Counterpropagation fuzzy–neural network for streamflow reconstruction. *Hydrol. Process.* **15**(2), 219–232 (2001)
- S.P. Charles, M.A. Bari, A. Kitsios, B.C. Bates, Effect of GCM bias on downscaled precipitation and runoff projections for the Serpentine catchment, western Australia. *Int. J. Climatol.* **27**, 1673–1690 (2007)
- V.T. Chow, *Handbook of Applied Hydrology* (McGraw-Hill, New York, 1964)
- V.T. Chow, D.R. Maidment, L.W. Mays, *Applied Hydrology* (McGraw-Hill, New York, 1988)
- T.A. Clair, J.M. Ehrman, Using neural networks to assess the influence of changing seasonal climates in modifying discharge, dissolved organic carbon, and nitrogen export in eastern Canadian rivers. *Water Resour. Res.* **34**(3), 447–455 (1998)
- M. Cobaner, Evapotranspiration estimation by two different neuro-fuzzy inference systems. *J. Hydrol.* **398**, 292–302 (2011)
- P. Coulibaly, N.D. Evora, Comparison of neural network methods for infilling missing daily weather records. *J. Hydrol.* **341**, 27–41 (2007)
- I. Daliakopoulos, P. Coulibaly, I. Tsanis, Ground water level forecasting using artificial neural networks. *J. Hydrol.* **309**(1-4), 229–240 (2005)
- T. Dalrymple, Flood frequency methods. U.S. Geol. Surv. Water Supply Pap., 1543-A, 11–51 (1960)
- C.W. Dawson, R.J. Abrahart, Evaluation of two different methods for the antecedent precipitation index in neural network river stage forecasting. *Geophys. Res. Abstr.* **07522**, 9 (2007)
- C. Dou, W. Woldt, I. Bogardi, M. Dahab, Steady State Groundwater Flow Simulation With Imprecise Parameters. *Water Resour. Res.* **31**(11), 2709–2719 (1995)
- C. Dou, W. Woldt, I. Bogardi, Fuzzy rule-based approach to describe solute transport in the unsaturated zone. *J. Hydrol.* **220**(1), 74–85 (1999)
- G. Dumedad, P. Coulibaly, Evaluation of statistical methods for infilling missing values in high-resolution soil moisture data. *J. Hydrol.* **400**(1-2), 95–102 (2011)
- S.R. Durrans, S. Tomic, Comparison of parametric tail estimators for low-flow frequency analysis. *J. Am. Water Resour. Assoc.* **37**(5), 1203–1214 (2001)
- J.K. Eischeid, P.A. Pasteris, H.F. Diaz, M.S. Plantico, N.J. Lott, Creating a serially complete, national daily time series of temperature and precipitation for the Western United States. *J. Appl. Meteorol.* **39**, 1580–1591 (2000)

- S. El Aldouni, T. Ouarda, X. Zhang, R. Roy, B. Bobee, Generalized maximum likelihood estimators for the nonstationary generalized extreme value model. *Water Resour. Res.* **43**, 1–13 (2007)
- A.K. Eldaw, J.D. Salas, L.A. Garcia, Long-range forecasting of the Nile River flows using climatic forcing. *J. Appl. Meteorol.* **42**(7), 890–904 (2003)
- R.F. Engle, Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica* **50**(4), 987–1007 (1982)
- M.A. Fedora, R.L. Beschta, Storm runoff simulation using an antecedent precipitation index (API) model. *J. Hydrol.* **112**, 121–133 (1989)
- R.I.C.C. Francis, J.A. Renwick, A regression-based assessment of the predictability of New Zealand climate anomalies. *Theor. Appl. Climatol.* **60**, 21–36 (1998)
- M. Friedman, A. Kandel, *Introduction to Pattern Recognition: Statistical, Structural, Neural, and Fuzzy Logic Approaches*, Series in Machine Perception Artificial Intelligence, vol 32 (World Scientific, Singapore, 1999)
- W.E. Fuller, Flood flows. *Trans. ASCE* **77**(1293), 564–617 (1914)
- D. Gautam, K. Holz, Rainfall-runoff modelling using adaptive neuro-fuzzy systems. *J. Hydroinf.* **3**, 3–10 (2001)
- T. Glade, M.J. Crozier, P. Smith, Applying probability determination to refine landslide-triggering rainfall thresholds using an empirical (Antecedent Daily Rainfall Model). *Pure Appl. Geophys.* **157**(6/8), 1059–1079 (2000)
- C.T. Haan, *Statistical Methods in Hydrology*, 2nd edn. (Iowa State University Press, Ames, 2002.) 496 pp
- M. Hallack-Alegria, D.W. Watkins, Annual and warm season drought intensity-duration-frequency analysis for Sonora, Mexico. *J. Clim.* **20**(9), 1897–1909 (2007)
- S. Haykin, *Neural Networks* (MacMillan, London, 1994)
- Z. He, X. Wen, H. Liu, J. Du, A comparative study of artificial neural network, adaptive neuro fuzzy inference system and support vector machine for forecasting river flow in the semiarid mountain region. *J. Hydrol.* **509**, 379–386 (2014)
- R.J. Heggen, Normalized antecedent precipitation index. *J. Hydrol. Eng. ASCE* **6**(5), 377–381 (2001)
- B.C. Hewitson, R.G. Crane, Climate downscaling: techniques and application. *Clim. Res.* **7**, 85–95 (1996)
- J.D. Hewlett, A.R. Hibbert, Factors affecting the response of small watersheds to precipitation in humid regions, in *Forest Hydrology*, ed. by W. E. Sopper, H. W. Lull, (Pergamon Press, Oxford, 1967), pp. 275–290.
- Y.S. Hong, M.R. Rosen, R.R. Reeves, Dynamic fuzzy modeling of storm water infiltration in urban fractured aquifers. *J. Hydrol. Eng.* **7**(5), 380–391 (2002)
- Y. Hong, R.F. Adler, F. Hossain, S. Curtis, G.J. Huffman, A first approach to global runoff simulation using satellite rainfall estimation. *Water Resour. Res.* **43**, W08502 (2007). <https://doi.org/10.1029/2006WR005739>
- J.R.M. Hosking, J.R. Wallis, *Regional frequency analysis: an approach based on L-moments* (Cambridge University Press, Cambridge, 1997)
- J.R.M. Hosking, J.R. Wallis, E.F. Wood, An appraisal of the regional flood frequency procedure in the UK Flood Studies Report. *Hydrol. Sci. J.* **30**, 85–109 (1985)
- J.R.M. Hosking, J.R. Wallis, Some statistics useful in regional frequency analysis: Water Resource Research, **29**(2), 271–281 (1993)
- J.F.T. Houston, Groundwater system simulation by time series techniques. *Ground Water* **21**, 301–310 (1983)
- K.-L. Hsu, H.V. Gupta, S. Sorooshian, Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **31**, 2517–2530 (1995)
- K.L. Hsu, X. Gao, S. Sorooshian, H.V. Gupta, Precipitation estimation from remotely sensed information using artificial neural networks. *J. Appl. Meteorol.* **36**(9), 1176–1190 (1997)
- Y. Huang, X. Chen, et al., A fuzzy-based simulation method for modelling hydrological processes under uncertainty. *Hydrol. Process.* **24**(25), 3718–3732 (2010)

- Y. Hundecha, A. Bardossy, H.W. Werner, Development of a fuzzy logic-based rainfall-runoff model. *Hydrol. Sci. J.* **46**(3), 363–376 (2001)
- A.P. Jacquin, A.Y. Shamseldin, Development of rainfall-runoff models using Takagi–Sugeno fuzzy inference systems. *J. Hydrol.* **329**, 154–173 (2006)
- S.K. Jain, A. Das, D.K. Srivastava, Application of ANN for Reservoir Inflow Prediction and Operation. *J. Water Resour. Plan. Man. ASCE*, **125**(5), 263–271 (1999)
- J.S.R. Jang, C.T. Sun, E. Mizutani, *Neuro-Fuzzy and Soft Computing* (Prentice-Hall, Englewood Cliffs, NJ, 1997)
- B. Janos, D. Lucien, H.R. Omar, Practical generation of synthetic rainfall event time series in a semi-arid climatic zone. *J. Hydrol.* **103**, 357–373 (1988)
- C. Jorge, Forecasting water consumption in Spain using univariate time series models. *Proc IEEE Span. Comput. Intell. Soc.* **2007**, 415–423 (2007)
- M. Karamouz, S. Sara Nazif, M. Falahi, *Hydrology and Hydroclimatology: Principles and Applications* (CRC Press, Boca Raton, 2012)
- N.K. Kasabov, *Foundations of Neural Networks, Fuzzy Systems, and Knowledge Engineering* (MIT Press, Cambridge, MA, 1996)
- M.N. Khalil, T.B.M.J. Ouarda, J.-C. Ondo, P. Gachon, B. Bobée, Frequency analysis of a sequence of dependent and/or non-stationary hydro-meteorological observations: A review. *J. Hydrol.* **329**, 534–552 (2006)
- Ö. Kişi, Streamflow forecasting using different artificial neural network algorithms. *J. Hydrol. Eng.* **12**(5), 532–539 (2007)
- G. Kite, Use of time series analyses to detect climatic change. *J. Hydrol.* **111**, 259–279 (1989)
- T.R. Kjeldsen, J.C. Smithers, R.E. Schulze, Flood frequency analysis at ungauged sites in the KwaZulu-Natal Province, South Africa. *Water SA* **27**(3), 315–324 (2001)
- G.J. Klir, B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications* (Prentice Hall, Upper Saddle River, 1995)
- M. A. Kohler, R. K. Linsley, Predicting the Runoff from Storm Rainfall, U.S. Weather Bureau Research Paper No. 34 (1951)
- R. Kruse, J.E. Gebhardt, F. Klöwn, *Foundations of Fuzzy Systems* (Wiley, New York, 1994)
- M. Kumar, N. Raghuwanshi, R. Singh, W. Wallender, W. Pruitt, Estimating evapotranspiration using artificial neural network. *J. Irrig. Drain. Eng.* **128**(4), 224–233 (2002)
- R. Kumar, C. Chatterjee, S. Kumar, A.K. Lohani, R.D. Singh, Development of regional flood frequency relationships using L-moments for Middle Ganga Plains Subzone 1(f) of India. *Water Resour. Manag.* **17**(4), 243–257 (2003)
- G. Lachtermacher, J.D. Fuller, Backpropagation in hydrological time series forecasting, in *Stochastic and Statistical Methods in Hydrology and Environmental Engineering*, (Springer, Cham, 1994), pp. 229–242
- X. Lana, A. Burgueno, Daily dry-wet behavior in Catalonia (NE Spain) from the viewpoint of the first and second order Markov chains. *Int. J. Clim.* **18**, 793–815 (1998)
- X. Lana, M.D. Martínez, A. Burgueño, C. Serra, Return period maps of dry spells for Catalonia (northeastern Spain) based on the Weibull. *Hydrol. Sci. J.* **53**, 48–64 (2008)
- S.A. Lawal, W.E. Watt, Non-zero lower limit in low flow frequency analysis? *Water Resour. Bull.* **32**(6), 1159–1166 (1996)
- M. Leclerc, T.B.M.J. Ouarda, Non-stationary regional flood frequency analysis at ungauged sites. *J. Hydrol.* **343**, 254–264 (2007)
- Y.H. Lim, D.L. Voeller, Regional flood estimations in Red River using L-moment-based index-flood and Bulletin 17B Procedures. *J. Hydrol. Eng.* **14**(9), 1002–1016 (2009)
- S. Liu, L. Leslie, M. Speer, R. Bunker, Predicting forest fire danger using improved model derived soil-moisture and antecedent precipitation. International Congress on Modelling and Simulation, Townsville, Australia, July 14–17, 2003
- Y.L. Loukas, Adaptive Neuro-Fuzzy Inference System, An Instant and Architecture-Free Predictor for Improved QSAR Studies. *J. Med. Chem.* **44**(17), 2772–2783 (2001)

- T. Ma, C. Li, Z. Lu, B. Wang, An effective antecedent precipitation model derived from the power-law relationship between landslide occurrence and rainfall level. *Geomorphology* **216**, 187–192 (2014)
- D. Machiwal, M.K. Jha, Time series analysis of hydrologic data for water resources planning and management: A review. *J. Hydrol. Hydromech.* **54**(3), 237–257 (2006)
- A. Makarau, M.R. Jury, Predictability of Zimbabwe summer rainfall. *Int. J. Climatol.* **17**, 1421–1432 (1997)
- J.M. Mejia, J. Rouselle, Disaggregation Models in hydrology revisited. *Water Resour. Res.* **12**, 185–186 (1976). <https://doi.org/10.1029/WR012i002p00185>
- A.K. Mishra, V.R. Desai, Drought forecasting using stochastic models. *Stoch. Environ. Res. Risk Assess.* **19**, 326–339 (2005)
- A. Moatmari, M. Longoni, R. Rosso, A seasonal long memory stochastic model for the simulation of daily river flows. *Phys. Chem. Earth (B)* **24**(4), 319–324 (1999)
- S. Mohan, N. Arumugam, Forecasting weekly reference crop evapotranspiration series. *Hydrol. Sci. J.* **40**(6), 689–720 (1995)
- T.J. Mulvaney, On the use of self-registering rain and flood gauges. *Trans. Inst. Civ. Eng. Ireland* **4**(2), 1–8 (1850)
- R.J. Nathan, T.A. McMahon, Practical Aspects of Low-Flow Frequency Analysis. *Water Resour. Res.* **26**(9), 2135–2141 (1990)
- D.W. Newton, J.C. Herrin, Assessment of commonly used methods of estimating flood frequency. *Transportation Research Record*, Series 896, Washington, DC, pp. 10–30 (1982)
- T. Ouarda, S. El-Aldouni, Bayesian nonstationarity frequency analysis of hydrological variables. *J. Amer. Water Resour. Assoc.* **47**, 496–505 (2011)
- T. Öztekin, Wakeby distribution for representing annual extreme and partial duration rainfall series. *Meteorol. Appl.* **14**, 381–387 (2007)
- G.R. Pandey, V.T.V. Nguyen, A comparative study of regression based methods in regional flood frequency analysis. *J. Hydrol.* **225**, 92–101 (1999)
- K. Pearson, On the generalized probable error in multiple normal correlation. *Biometrika* **6**, 59–68 (1908)
- P. Pekarova, J. Pekar, Long-term discharge prediction for the Turnu Severin station (the Danube) using a linear autoregressive model. *Hydrol. Process.* **20**, 1217–1228 (2006)
- G. Pesti, B.P. Shrestha, L. Duckstein, I. Bogárdi, A fuzzy rule-based approach to drought assessment. *Water Resour. Res.* **32**(6), 1741–1747 (1996)
- A. Pieglat, *Fuzzy Modeling and Control. Studies in Fuzziness and Soft Computing* (Book 69) (Springer, Berlin, 2001)
- D. H. Pilgrim (ed.), *Australian Rainfall and Runoff: A Guide to Flood Estimation, Volumes I and II* (Institution of Engineers Australia, Canberra, 1998)
- A.S. Polebitski, R.N. Palmer, Seasonal residential water demand forecasting for census tracts. *J. Water Resour. Plan. Manag.* **136**(1), 27–36 (2009)
- R. Pongracz, J. Bartholy, I. Bogardi, Fuzzy rule-based prediction of monthly precipitation. *Phys. Chem. Earth B Hydrol. Oceans Atmos.* **26**(9), 663–667 (2001)
- M.P. Rajurkar, U.C. Kothiyari, U.C. Chaube, Modeling of the daily rainfall-runoff relationship with artificial neural network. *J. Hydrol.* **285**, 96–113 (2004)
- H. Raman, V. Chandramouli, Deriving a general operating policy for reservoirs using neural networks. *J. Water Resour. Plan. Manag.* **122**(5), 342–347 (1996)
- H. Raman, N. Sunilkumar, Multi-variate modeling of water resources time series using artificial neural networks. *Hydrol. Sci. J.* **40**, 145–163 (1995)
- B. Renard, M. Lang, P. Bois, Statistical analysis of extreme events in a non-stationary context via a Bayesian framework: case study with peak-over-threshold data. *Stoch. Env. Res. Risk Assess.* **21**, 97–112 (2006)
- H.C. Riggs, *Streamflow characteristics: Developments in water science* 22 (Elsevier, Amsterdam, 1985)

- L.L. Rogers, F.U. Dowla, Optimization of groundwater remediation using artificial neural networks with parallel solute transport modeling. *Water Resour. Res.* **30**(2), 457–481 (1994)
- S.O. Russell, P.F. Campbell, Reservoir operating rules with fuzzy programming. *J. Water Resour. Plann. Manag.* **122**(3), 165–170 (1996)
- B. Saf, Regional flood frequency analysis using L-moments for the west Mediterranean region of Turkey. *Water Resour. Manag.* **23**(3), 531–551 (2009)
- J.D. Salas, Analysis and modeling of hydrologic time series, Chapter 19, in *Handbook of Hydrology*, ed. by D. R. Maidment (McGraw-Hill, New York, 1993)
- J.D. Salas, J.W. Deutcher, V. Yevjevich, W.L. Lane, *Applied Modeling of Hydrologic Time Series* (Water Resources Publications, Littleton, 1980)
- A. Sankarasubramanian, U. Lall, Flood quantiles in a changing climate: seasonal forecasts and causal relations. *Water Resour. Res.* **39**, 4.1–4.12 (2003)
- K.E. Saxton, A.M. ASCE, A.T. Lenz, F. ASCE, Antecedent retention indexes predict soil moisture. *J. Hydraul. Div. ASCE* **93**(HY4, Proc. Paper 5351), 223–241 (1967)
- K. Schulz, B. Huwe, Water flow modeling in the unsaturatedzone with imprecise parameters using a fuzzy approach. *J. Hydrol.* **201**, 211–229, (1997). [https://doi.org/10.1016/S0022-1694\(97\)00038-3](https://doi.org/10.1016/S0022-1694(97)00038-3).
- L. See, S. Openshaw, Applying soft computing approaches to river level forecasting. *Hydrol. Sci. J.* **44**(5), 763–778 (1999)
- A.Y. Shamseldin, Application of a neural network technique to rainfall-runoff modeling. *J. Hydrol.* **199**, 272–294 (1997)
- V.N. Sharda, S.O. Prasher, R.M. Patel, P.R. Ojasvi & Chandra Prakash. Performance of Multivariate Adaptive Regression Splines (MARS) in predicting runoff in mid-Himalayan micro-watersheds with limited data / Performances de régressions par splines multiples et adaptives (MARS) pour la prévision d'écoulement au sein de micro-bassins versants Himalayens d'altitudes intermédiaires avec peu de données. *Hydrol. Sci. J.* **53**(6), 1165–1175 (2008). <https://doi.org/10.1623/hysj.53.6.1165>
- Q.X. Shao, H. Wong, M. Li, W.C. Ip, Streamflow forecasting using functional-coefficient time series model with periodic variation. *J. Hydrol.* **368**, 88–95 (2009)
- J. Shiri, A.H. Nazemi, A.A. Sadraddini, G. Landeras, O. Kisi, A.F. Fard, P. Marti, Global cross-station assessment of neuro-fuzzy models for estimating daily reference evapotranspiration. *J. Hydrol.* **480**, 46–57 (2013)
- J. Shiri, P. Marti, A.H. Nazemi, A.A. Sadraddini, O. Kisi, Local vs. external training of neuro-fuzzy and neural networks models for estimating reference evapotranspiration assessed through k-fold testing. *Hydrol. Res.* **46**(1), 72–88 (2014). <https://doi.org/10.2166/nh.2013.112>
- C. Shu, D.H. Burn, Homogeneous pooling group delineation for flood frequency analysis using a fuzzy expert system with genetic enhancement. *J. Hydrol.* **291**(1-2), 132–149 (2004)
- K. Schulz, B. Huwe, Water flow modeling in the unsaturatedzone with imprecise parameters using a fuzzy approach. *J. Hydrol.* **201**, 211–229 (1997). [https://doi.org/10.1016/S0022-1694\(97\)00038-3](https://doi.org/10.1016/S0022-1694(97)00038-3)
- W.T. Sittner, C.E. Schauss, J.C. Monro, Continuous hydrograph synthesis with an API type hydrological model. *Water Resour. Res.* **5**(5), 1007–1022 (1969)
- J.R. Stedinger, R.M. Vogel, E. Foufoula-Georgiou, Frequency analysis of extreme events, Chapter 18, in *Handbook of Applied Hydrology*, ed. by D. R. Maidment (McGraw-Hill, New York, 1993), pp. 1–66
- W.G. Strupczewski, V.P. Singh, H.T. Mitosek, Non-stationary approach to at-site flood frequency modelling, III. Flood analysis of Polish rivers. *J. Hydrol.* **248**, 152–167 (2001)
- M. Sugeno, G.T. Kang, Structure identification of fuzzy model. *Fuzzy Sets Syst.* **28**, 15–33 (1988)
- T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modeling and control. *IEEE Trans. Syst. Man Cybern. SMC-15*(1), 116–132 (1985)
- D.M. Thomas, M.A. Benson, Generalization of streamflow characteristics from drainage-basin characteristics, US Geological Survey, Water Supply Paper, 1975 (1970)

- United States Water Resources Council, Guidelines for determining flood flow frequency Bull. 17B, U.S. Water Resour. Coun. Hydrol. Comm., Washington, DC (1982)
- R.D. Valencia, J.C. Schaake, Disaggregation processes in stochastic hydrology. *Water Resour. Res.* **9**, 580–585 (1973). <https://doi.org/10.1029/WR009i003p00580>
- F.C. Van Geer, A.F. Zuur, An extension of Box-Jenkins transfer/noise models for spatial interpolation of groundwater head series. *J. Hydrol.* **192**, 65–80 (1997)
- W. Viessman Jr., G.L. Lewis, *Introduction to Hydrology*, 4th edn. (Harper Collins, New York, 1996)
- R.M. Vogel, C.N. Kroll, Generalized low-flow frequency relationships for ungaged sites in Massachusetts. *Water Resour. Bull.* **26**(2), 241–253 (1990)
- R.M. Vogel, I. Wilson, Probability distribution of annual maximum, mean, and minimum streamflows in the United States. *ASCE J. Hydrol. Eng.* **1**(2), 69–76 (1996)
- R.M. Vogel, W.O. Thomas Jr., T.A. McMahon, Flood-flow frequency model selection in Southwestern United States. *J. Water Resour. Plann. Manag.* **119**(3), 353–366 (1993)
- L.X. Wang, *A Course in Fuzzy Systems and Control* (Prentice-Hall, Englewood Cliffs, 1997)
- W.D. Weeks, W.C. Boughton, Tests of ARMA model forms for rainfall-runoff modeling. *J. Hydrol.* **91**, 29–47 (1987)
- J.R. Westmacott, D.H. Burn, Climate change effects on the hydrologic regime within the Churchill-Nelson River Basin. *J. Hydrol.* **202**(1–4), 263–279 (1997)
- R.L. Wilby, L.E. Hay, G.H. Leavesley, A comparison of downscaled and raw GCM output: implications for climate change scenarios in the San Juan River basin, Colorado. *J. Hydrol.* **225**, 67–91 (1999)
- C.L. Wu, K.W. Chau, Rainfall-runoff modeling using artificial neural network coupled with singular spectrum analysis. *J. Hydrol.* **399**, 394–409 (2011)
- L. Xiong, A.Y. Shamseldin, K.M. O'Connor, A non-linear combination of the forecasts of rainfall-runoff models by the first-order Takagi-Sugeno fuzzy system. *J. Hydrol.* **245**, 196–217 (2001)
- L. Xiong, K.M. O'Connor, S. Guo, Comparison of three updating schemes using artificial neural network in flow forecasting. *Hydrol. Earth. Sys. Sci.* **8**(2), 247–255 (2004)
- A. Yasar, M. Bilgili, E. Simsek, Water demand forecasting based on stepwise multiple nonlinear regression analysis. *Arab. J. Sci. Eng.* **37**(8), 2333–2341 (2012)
- X.Y. Yu, S.Y. Liang, Forecasting of hydrologic time series with ridge regression in feature space. *J. Hydrol.* **332**, 290–302 (2007)
- P.S. Yu, T.C. Yang, Fuzzy multi-objective function for rainfall-runoff model calibration. *J. Hydrol.* **238**, 1, 1–1, 14 (2000)
- S. Yue, C.Y. Wang, Possible regional probability distribution type of Canadian annual streamflow by L-moments. *Water Resour. Manag.* **18**(5), 425–438 (2004)
- M.A. Yurdusev, M. Firat, Adaptive neuro fuzzy inference system approach for municipal water consumption modeling, An application to Izmir, Turkey. *J. Hydrol.* **365**(3–4), 225–234 (2009)
- L.A. Zadeh, Fuzzy sets. *Inf. Control.* **8**, 338–353 (1965)
- C.M. Zealand, D.H. Burn, S.P. Simonovic, Short term streamflow forecasting using artificial neural networks. *J. Hydrol.* **214**, 32–48 (1999)
- H.J. Zimmermann, *Fuzzy Set Theory and Its Applications*, 4th edn. (Kluwer Academic Publishers, Boston, 2001)



Conceptual Hydrological Models

Zhaofei Liu, Yamei Wang, Zongxue Xu, and Qingyun Duan

Contents

1	Introduction	390
2	Hydrological Processes Described by Conceptual Hydrological Models	391
2.1	Precipitation	391
2.2	Infiltration	393
2.3	Soil Moisture Storage	396
2.4	Evapotranspiration	398
2.5	Runoff Generation	399
2.6	River Routing	402
3	Typical Conceptual Hydrological Models	406
3.1	The Tank Model	406
3.2	The Xinanjiang Model	407
3.3	The Sacramento Model	409
	References	410

Abstract

Conceptual hydrological models, sometimes also called gray-box models, are precipitation-runoff models built based on observed or assumed empirical relationships among different hydrological variables. They are different from black-box models which consider precipitation-runoff relationship only statistically. They are also different from the physically based distributed hydrological models which are

Z. Liu (✉)

Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China
e-mail: zfliu@igsnrr.ac.cn

Y. Wang · Q. Duan

Faculty of Geographical Science, Beijing Normal University, Beijing, China

Z. Xu (✉)

College of Water Science, Beijing Normal University, Beijing, China
e-mail: zxxu@bnu.edu.cn

based on solving differential equations describing the physical laws of mass, energy, and momentum conservations. This chapter describes how conceptual hydrological models represent the different hydrological processes involved in converting precipitation to runoff over land, and then to streamflow discharge at the basin outlet, including precipitation, snow accumulation and ablation, infiltration, soil moisture storage, evapotranspiration, runoff generation, baseflow, and river routing. Some of the well-known models are also used for illustration.

Keywords

Conceptual hydrological model · Precipitation · Infiltration · Soil moisture storage · Evaporation · Evapotranspiration · Runoff generation · River routing · Tank model · Xinjiang model · Sacramento model

1 Introduction

Hydrological cycle refers to the continuous circulation of water between the Earth and atmosphere (see Fig. 1). Water moves between land, sea, and atmosphere via processes such as evaporation, condensation, precipitation, deposition, runoff, infiltration, sublimation, transpiration, melting, and groundwater flow. Although these processes are fairly easy to grasp, they are far from easy to understand and quantify in detail. In order to do this, abstraction is necessary and various types of hydrological models have been created. In general, all hydrological models are designed to meet one of the two primary objectives: (1) to study the system operation, and (2) to predict its system behavior.

Hydrological models can be classified into three categories (1) black-box models, (2) conceptual model, and (3) physically based model. Black-box models, also referred as empirical models, consider the system input-output relationships from a statistical viewpoint. They do not aid in physical understanding of the system behaviors. Physical-based model (sometimes called white-box models or theoretical models), on the other hand, describe hydrological processes in details by solving differential equations describing the physical laws of mass, energy, and momentum conservations. Those equations are generally solved over some grid structure representing a spatial domain. Therefore, physically based models are often called distributed hydrological models. Conceptual models (sometimes called gray-box models) consider physical laws but in highly simplified forms. A conceptual model is a descriptive representation of hydrologic system that incorporates the modeler's understanding of the relevant physical, chemical, and hydrologic conditions. Conceptual rainfall-runoff models are designed to predict magnitude of streamflow by conceptualizing rainfall-to-runoff generating processes and simulating internal variables, such as soil moisture, by various types of response functions. Such models consider six major contributing physical processes and their relationships, including precipitation, infiltration, soil moisture storage, evapotranspiration, runoff generation (including both surface and subsurface runoff), and river routing.

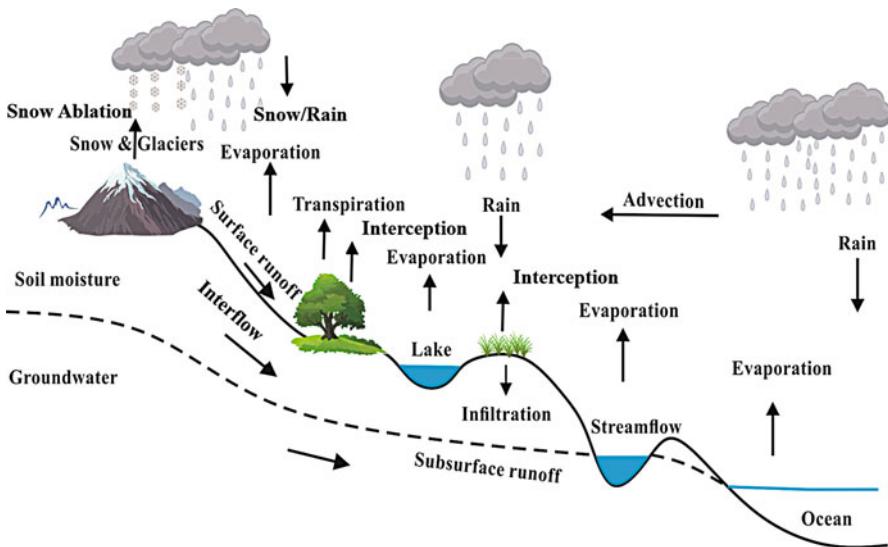


Fig. 1 A schematic description of hydrologic cycle

In the following sections, how those processes are represented in conceptual hydrological models are briefly described, followed by illustration with several well-known and widely used conceptual hydrological models.

2 Hydrological Processes Described by Conceptual Hydrological Models

2.1 Precipitation

Precipitation is the most important input to all hydrological models. Precipitation includes rainfall, snowfall, and other forms by which water falls from the air to the land surface (e.g., hail and sleet). The first two forms constitute the major part of precipitation and are of great importance to hydrological models. There are several methods to estimate areal rainfall from observed rain-gauge station data. In general, most models use areal average rainfall, i.e., arithmetic average or weighted average values, as inputs. The Thiessen polygon method (Thiessen 1911) is the most widely used method to estimate the station weights. Other popular methods include inverse distance and Kriging methods. In many models, rainfall input is implicitly assumed to be uniform in space. This can be problematic if the underlying basin is large. Some researchers use an exponential distribution function to account for spatial variability of rainfall for large river basins (Koren 1993). For example, the exponential distribution function is used for modeling areal rainfall in the simple water balance (SWB) model (Schaake et al. 1996). Since the development of remote sensing technology, remote-sensed precipitation products from radars and satellites have been widely

used as precipitation data sources. The advantage of remote sensing precipitation product is that it can account for the areal distribution of precipitation. On the other hand, remote sensing precipitation data contain significant uncertainty because it is derived from radiative signals, which can be easily interfered by objects or noises.

Snowfall plays a significant part in the hydrological regime in many parts of the world, especially at middle and high latitude basins and high elevation basins. Whether precipitation falls as rain or snow can have a very significant influence on the estimation of runoff, especially for spring snowmelt-induced runoff. Model outputs are therefore sensitive to whether the form of precipitation is determined correctly. The determination of precipitation form is usually based on concurrent air temperature records. If air temperature is lower than a certain threshold value (usually set to 0°C), the precipitation form is snow, otherwise rainfall. It may be noted that some models use a fixed value for temperature threshold, whereas others treat it as a calibration parameter. Methods used in conceptual hydrological models for distinguishing the rainfall and snowfall could be summarized as following:

$$P_t = \begin{cases} P_r & \text{if } T \geq T_0 \\ P_s & \text{if } T < T_0 \end{cases} \quad (1)$$

where P_r is the amount of precipitation in the form of rain (mm), P_s is the amount of precipitation in the form of snow (mm), P_t is the total precipitation (mm), T is the daily air temperature (K), and T_0 is the threshold temperature (K).

When snowfall is detected, the snow module in the hydrological model is activated. Snow accumulation and ablation is generally estimated by using an energy balance approach or a degree-day approach.

The energy balance of a snow cover can be expressed as (Anderson 1976; Price and Dunne 1976):

$$\Delta Q = Q_n + Q_e + Q_h + Q_m + Q_g \quad (2)$$

where ΔQ is the change in snow cover energy, Q_n is the net radiative heat flux, Q_e is the latent heat flux, Q_h is the sensible heat flux, Q_m is the heat gained from precipitation, and Q_g is the heat transfer across the snow-soil interface. The unit of each term is $\text{cal}\cdot\text{cm}^{-2}$. A positive energy balance warms the snow cover, and results in melt. In general, frozen ground at the base of the snowpack persisted throughout the melt period, so that Q_g is usually assumed to be zero. Therefore, the four remaining components are as follows,

$$Q_n = Q_i - Q_r + \varepsilon Q_a - \Delta t \cdot \varepsilon \cdot \sigma \cdot T_s \quad (3)$$

where Q_i and Q_r are the incident (incoming) and reflected (outgoing) solar radiation ($\text{cal}\cdot\text{cm}^{-2}$), respectively, Q_a is the incoming long-wave radiation ($\text{cal}\cdot\text{cm}^{-2}$), ε is the emissivity in the long-wave portion of the energy spectrum, σ is the Stefan-Boltzmann constant ($1.355 \times 10^{-12} \text{ cal}\cdot\text{cm}^{-2}\cdot\text{K}^{-4}\cdot\text{s}^{-1}$), Δt is the time interval (s), and T_s is the snow surface temperature (K).

$$Q_e = 0.1 L_S \cdot \rho_w \cdot f(U_a) \cdot (e_a - e_s) \quad (4)$$

$$f(U_a) = a U_a + b \quad (5)$$

where L_S is the latent heat of sublimation ($677 \text{ cal}\cdot\text{g}^{-1}$), ρ_w is the density of water ($\text{g}\cdot\text{cm}^{-3}$), U_a is the wind travel (mean wind speed multiplied by time interval) (km), e_a and e_s are the air and snow surface vapor pressure (mPa), a and b are the fitted coefficients.

$$Q_h = 0.16 \rho_w \cdot c_a \cdot p \cdot f(U_a) \cdot (T_a - T_s) \quad (6)$$

where c_a is the specific heat of dry air ($\text{cal}\cdot\text{g}^{-1}\cdot\text{K}^{-1}$), p is the atmospheric pressure (mPa), and T_a is the air temperature (K).

$$Q_m = 0.1 c_w \cdot \rho_w \cdot P \cdot (T_w - 273.15) \quad (7)$$

where c_w is the specific heat of water ($\text{cal}\cdot\text{g}^{-1}\cdot\text{K}^{-1}$), P is the precipitation (mm), T_w is the wet-bulb temperature (K).

The degree-day method is used in most conceptual hydrological models as it needs the least amount of data (Hock 2003). In this approach, a simple degree-day expression to estimate snowmelt based on air temperature can be described as follows:

$$SM = DD \cdot (T - T_b) \quad (8)$$

where SM = snowmelt ($\text{mm}\cdot\text{day}^{-1}$), DD = degree-day factor ($\text{mm}\cdot\text{K}^{-1}\cdot\text{day}$) indicating the snowmelt depth resulting from 1 degree-day, and T_b = a base temperature (K).

In most cases, T_b is assumed to be a constant (i.e., $T_b = 273.15$); it could also be estimated by model calibration. This method is used in the HBV model. Besides that, the SRM also considers snow-covered area in simulating snowmelt, which could be represented by a simplified equation,

$$SM = DD \cdot (T - T_b) \cdot A \quad (9)$$

where A is the snow-covered area in a region.

2.2 Infiltration

Infiltration is a key process in the hydrological cycle. Infiltration models usually employ simplified concepts of the infiltration rate or cumulative infiltration volume. It assumes that surface runoff begins when the precipitation rate exceeds the soil surface infiltration rate. Because of its fundamental role in land-surface and subsurface hydrology, infiltration has received a great deal of attention from soil and water scientists, and a large number of infiltration models have been developed to compute

it. There are three types for infiltration models: empirical-based, semiempirical-based, and physically-based models. Physically-based models specify appropriate boundary conditions and normally require detailed data input. It requires solution of the Richards' equation (Richards 1931), which describes water flow in soils in terms of the hydraulic conductivity and the soil water pressure as functions of soil water content, for specified boundary conditions. However, it is extremely difficult to obtain all of data input required in the physically-based models. Therefore, for many applications, equations that simplify the concepts involved in the infiltration process are desirable for practical use (Rawls et al. 1993). The empirical and semiempirical approaches, which use simplified concepts for infiltration processes, are used in conceptual hydrological models. The empirical approaches generally relate infiltration rate or volume to elapsed time modified by certain soil properties. The most common empirical approaches are equations of Kostiakov, Horton, and Holtan. The semiempirical approaches such as of the Green and Ampt equation and the Philip equation apply the physical principles governing infiltration for simplified boundary and initial conditions.

The Kostiakov Equation

Kostiakov (1932) and independently Lewis (1937) proposed a simple empirical infiltration equation based on curve fitting from field data. It relates infiltration to time as a power function:

$$f_p = K_k \cdot t^{-\alpha} \quad (10)$$

where f_p is the infiltration capacity ($\text{mm} \cdot \text{s}^{-1}$), t is the time after infiltration starts (s), and K_k and α are constants depending on the soil type and initial conditions.

The parameters K_k and α must be evaluated from measured infiltration data, since they have no physical interpretation. The equation describes the measured infiltration curve and, given the same soil and same initial water condition, allows prediction of an infiltration curve using the same constants developed for those conditions.

The Kostiakov equation is widely used because of its simplicity, ease of determining the two constants from measured infiltration data, and reasonable fit to infiltration data for many soils over short-time periods (Clemmens 1983). Mezencev (1948) proposed a modification to Kostiakov's equation by adding a constant to the equation that represents the final infiltration rate reached when the soil becomes saturated after prolonged infiltration. The Kostiakov and modified Kostiakov equations tend to be the preferred models used for irrigation infiltration, probably because it is less restrictive as to the mode of water application than some other models.

The Horton Equation

Possibly, the best-known infiltration expression is that known as Horton's equation (11), which was proposed in 1940. Horton recognized that infiltration capacity (f_p) decreased with time until it approached a minimum constant rate (f_c). He attributed this decrease in infiltration primarily to factors operating at the soil surface rather than to flow processes within the soil. The equation is

$$f_p = f_c + (f_o - f_c) e^{-kt} \quad (11)$$

where f_c is the final or equilibrium infiltration rate ($\text{mm} \cdot \text{s}^{-1}$), f_o is the initial infiltration capacity at $t = 0$ ($\text{mm} \cdot \text{s}^{-1}$), and k is a constant dependent on soil type and the initial moisture content.

The parameters f_c , k , and f_o can be evaluated from measured infiltration data.

Horton's equation has advantages over the Kostiakov equation. First, at time that t equals 0, the infiltration capacity is not infinite but takes on the finite value f_o . Also, as t approaches infinity, the infiltration capacity approaches a nonzero constant minimum value of f_c (Horton 1940; Hillel 1998). Horton's equation has been widely used because it generally provides a good fit to data. Although the Horton equation is empirical in that β , f_c , and f_o must be estimated from experimental data, rather than measured in the laboratory, it does reflect the laws and basic equations of soil physics (Chow et al. 1988).

However, the Horton equation is cumbersome in practice since it contains three constants that must be evaluated experimentally (Hillel 1998). A further limitation is that it is applicable only when rainfall intensity exceeds f_c (Rawls et al. 1993). Horton's approach has also been criticized because he neglects the role of capillary potential gradients in the declination of infiltration capacity over time and attributes control almost entirely to surface conditions (Bevin 2004). Another criticism of the Horton model is that it assumes that hydraulic conductivity is independent of the soil water content (Novotny and Olem 1994).

The Holtan Equation

Holtan (1961) described an empirical equation based on a storage exhaustion concept. The infiltration rate is expressed in terms of cumulative infiltration, initial soil water content, and other soil variables:

$$f_p = f_c + a \cdot F_p^n \quad (12)$$

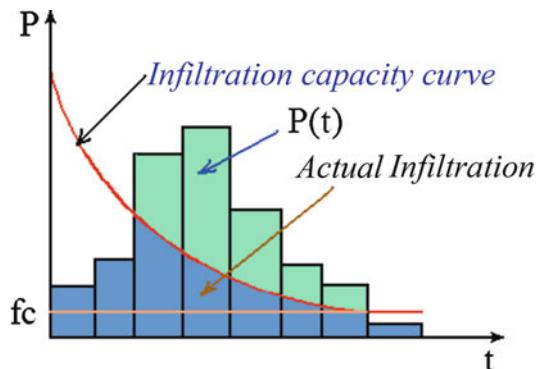
where F_p is the unfilled capacity of the soil to store water (equal to the initial available moisture storage minus the volume of water already infiltrated), a is a constant between 0.25 and 0.80, and n is a constant dependent on soil type.

The exponent n has been found to be about 1.4 for many soils. The value of F_p ranges from the maximum of the available water capacity to zero. This expression is well suited for inclusion in a watershed model, because it links infiltration capacity to the soil moisture level and is not time dependent. The Holtan model for infiltration has the advantage over the Horton model, in that it has a more physical basis and can describe infiltration and the recovery of infiltration capacity during periods low or no rainfall.

More details about infiltration approaches could be found in "Handbook of Hydrology" (Maidment 1993).

Infiltration is the process by which water enters the soil. In some conceptual hydrological models, surface runoff or overland runoff occurs when the infiltration capacity is smaller than the precipitation intensity. If the infiltration capacity is

Fig. 2 Infiltration process modeling in conceptual hydrological models



greater than the precipitation intensity, all the water enters the soil profile and no overland flow occurs. In general, runoff generation based on infiltration in conceptual hydrological models can be illustrated in Fig. 2.

The red curve is the infiltration capacity curve. It could be one of the infiltration equations, such as the Horton and Holtan equations. With time, the value of the infiltration capacity decreases to the final or equilibrium infiltration rate f_c . At first, because the precipitation intensity is smaller than the infiltration capacity, actual infiltration (marked blue) equals the precipitation intensity. In other words, all precipitation enters the soil without any surface runoff generated. When the precipitation intensity exceeds the infiltration capacity, surface runoff generates and equals to the difference value between the precipitation intensity and the infiltration capacity.

2.3 Soil Moisture Storage

When water infiltrates into the soil, it can storage and process considerable amounts of water. The soil would storage water until it is saturated. In this process, water infiltrates the soil surface and then moves laterally through the soil toward the stream channels, either as subsurface runoff or as underground (groundwater) runoff. Water storage in the soil can be quantified on the basis of its volumetric or gravimetric water content. The volumetric water content is the volume of water per unit volume of soil, expressed as a percentage of the volume. The gravimetric water content is the mass of water per unit mass of dry (or wet) soil. The volumetric water content is equal to the gravimetric water content times the soil's bulk density (on a dry soil basis).

In the conceptual hydrological models, it usually uses the soil moisture storage capacity to represent this process. The maximum soil moisture storage capacity is used in most of the conceptual hydrological models, including Tank, SCS, TOPMODEL, etc. In this concept, all the precipitation falling over the soil infiltrates unless the soil water content reaches the maximum soil moisture storage capacity (saturation).

Besides that, the soil moisture storage capacity curve is also applied in some conceptual hydrological models, including Xinanjiang, ARNO, and HBV models. It provides a nonuniform distribution of soil moisture storage capacity over the basin. In other words, the proportion of saturated areas, which generate surface runoff, is represented by the soil moisture storage capacity curve.

The Xinanjiang model was developed in 1973 and internationally published in 1980 (Zhao et al. 1980). The soil moisture storage capacity curve is the key of the model. It could be described as following,

$$\frac{f}{F} = 1 - \left(1 - \frac{W_m}{W_{mm}}\right)^B \quad (13)$$

where f is the saturated areas, F is the basin area, W_m is the soil moisture storage capacity, W_{mm} is the maximum soil moisture storage capacity, and B is the homogeneity of soil moisture storage capacity in the basin.

The soil moisture storage capacity curve used in the ARNO model is the same as that in the Xinanjiang model.

In the HBV model, the soil moisture accounting routine computes an index of the wetness of the entire basin and integrates interception and soil moisture storage. It is controlled by three free parameters, FC, BETA, and LP, as shown in Fig. 3. FC is the maximum soil moisture storage in the basin and BETA determines the relative contribution to runoff from rain or snowmelt at a given soil moisture deficit. LP controls the shape of the reduction curve for potential evaporation. At soil moisture values below LP, the actual evapotranspiration will be reduced.

The ideal soil is considered to be one which is homogeneous throughout the profile, and in which all of the pores are interconnected by capillaries. In addition, it is assumed that the applied rainfall falls uniformly over the soil surface. Because the movement of water into the soil is areally uniform, the infiltration process can be considered to be one dimensional. For this ideal case, perhaps the most important factors which affect the infiltration capacity are soil type and moisture content. The soil type determines the size and number of the capillaries through which the water must flow. However, actual soil conditions are seldom uniformly distributed in a basin. Therefore, a concept of variable infiltration capacity (VIC) was proposed by Liang et al. (1994). It is also similar to soil moisture storage capacity curve in the Xinanjiang model, and could be expressed by the following equation,

$$i = i_m (1 - A)^{\frac{1}{b_i}} \quad (14)$$

where i is the infiltration capacity, i_m is maximum infiltration capacity, A is the fraction of an area for which the infiltration capacity is less than i , and b_i is the infiltration shape parameter, which is a measure of the spatial variability of the infiltration capacity, defined as the maximum amount of water that can be stored in the soil column.

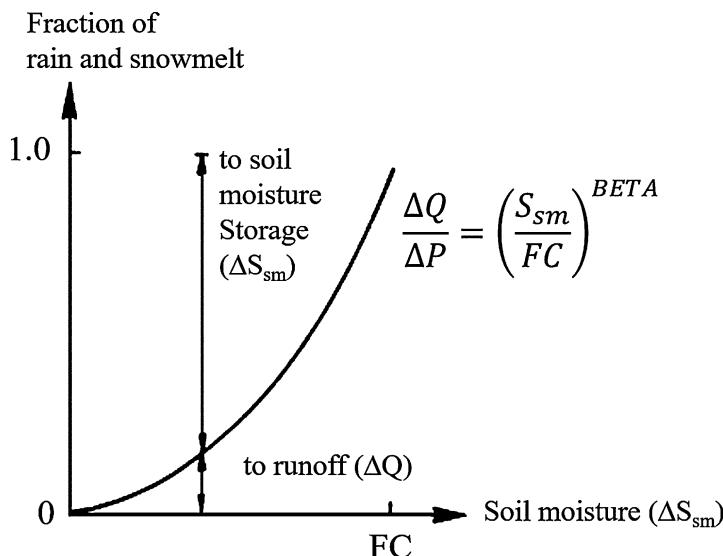


Fig. 3 The soil moisture storage capacity curve in the HBV model

Antecedent soil moisture is the degree of soil water content prior to a precipitation event. It is one of the most important factors controlling hydrological processes. It is usually estimated as a parameter in many conceptual hydrological models, including SCS, Xinanjiang, et al.

2.4 Evapotranspiration

Evapotranspiration, which include evaporation from soil and water surface and transpiration from vegetation, is usually a key variable in hydrological models. Generally, actual evapotranspiration is considered as a function of soil and vegetation properties, or a function upon potential evapotranspiration (PET). PET is generally considered to be the maximum rate of evaporation from vegetation-covered land surfaces when water is freely available. Following infiltration, evaporation from bare soil follows an atmosphere-controlled stage where evaporation is largely independent of soil moisture content and evaporation occurs near the free-water rate. Then there is a soil-controlled stage in which evaporation rate is determined by the rate at which water can be conducted to the surface rather than by atmospheric conditions.

In conceptual hydrological models, a linear relationship between actual and potential evaporation is usually used to estimate actual evaporation. Potential evaporation could be represented by pan evaporation rate, which is observed at most regions. Therefore, actual evaporation is estimated from pan evaporation rate by a linear formula in some conceptual hydrological models, when pan evaporation

data is available. For example, in the Xinanjiang and the PRMS models, measured pan evaporation is used for estimating actual evaporation.

However, pan evaporation data is lacking at many regions. Potential evaporation is usually estimated by evaporation models. A great number of evaporation models have been developed and validated through field measurements, from the single climatic variable-driven equations to the energy balance and aerodynamic principle combination methods. Among them, the Thornthwaite equation (Thornthwaite 1948), Penman equation (Penman 1948), and Penman–Monteith equation (Allen et al. 1998) are widely used.

The Streamflow Synthesis and Reservoir Regulation model directly applied the Thornthwaite equation to estimate potential evapotranspiration as a function of air temperature. Other climatic variables are also used to estimate potential evapotranspiration in different conceptual hydrological models. For example, in the PRMS model, it include two equations,

$$PET = C_m \cdot SD \cdot \rho \quad (15a)$$

$$PET = C_m \cdot (T_{mean} - T_b) \cdot R_d \quad (15b)$$

where PET is the potential evapotranspiration, C_m is a monthly coefficient, SD is the Sunshine duration (hours), ρ is the absolute humidity (g/m^3), T_{mean} is the daily mean air temperature, T_b is a coefficient, and R_d is the daily solar radiation.

In the HBV model, the following equation is used to estimate potential evapotranspiration,

$$PET = (1 + C \cdot (T_{mean} - T_m)) \cdot PE_m \quad (16)$$

where C is an empirical model parameter, T_m is the monthly long-term average temperature, and PE_m is the monthly average potential evapotranspiration.

In the UBC model, maximum air temperature is used to estimate potential evapotranspiration,

$$PET = 0.133 C_m \cdot T_{max} \quad (17)$$

where C_m is a monthly coefficient and T_{max} is the maximum air temperature.

If water is not readily available from soil surface, actual evapotranspiration is usually considered as a function upon soil moisture deficiency and potential evapotranspiration.

2.5 Runoff Generation

In a basin, runoff generation is equal to net precipitation, which is precipitation amount subtracted by precipitation losses. Precipitation losses include infiltration, soil moisture storage, and evapotranspiration as mentioned above. Net precipitation

need to be transformed into runoff (flow) at the basin outlet, because point net precipitation flow to the basin outlet is different in physical meanings. However, duration of runoff generation from net precipitation is considered as uniformly over the entire area of the basin in conceptual hydrological models.

1. Unit Hydrograph Method

The unit hydrograph method is widely used for runoff generation from net precipitation. It is defined as the direct runoff hydrograph resulting from a unit volume of net precipitation of constant intensity and uniformly distributed over the drainage area. The duration of the unit volume of net precipitation, sometimes referred to as the effective duration, defines and labels the particular unit hydrograph. The unit volume is usually considered to be associated with 1 cm (1 inch) of net precipitation distributed uniformly over the basin area. The fundamental assumptions implicit in the use of unit hydrographs for modeling hydrological systems are: (a) Watersheds respond as linear systems; (b) The net precipitation is uniformly distributed over the entire basin; (c) Net precipitation is of constant intensity throughout the precipitation duration; (d) The duration of runoff hydrograph depends on the net precipitation duration.

The discrete convolution equation allows the calculation of flow for a given net precipitation

$$Q_n = \sum_{m=1}^{n \leq M} P_m \times U_{n-m+1} \quad (18)$$

where Q is the flow, P is the net precipitation, U is the unit hydrograph, M is the net precipitation steps, n is the flow steps, and The unit hydrograph has $N-M + 1$ pulses.

The determination of unit hydrographs for particular basins can be carried out either using the theoretical developments of linear system theory or using empirical techniques. For either case, simultaneous observations of both precipitation and flow must be available. In other words, UH is applicable only for gauged basins and for the point on the stream where data are observed. Thus, the resultant UH is specific to a basin defined by the point on the stream where flow observations were made. When no direct observations are available, or when UH's for other locations on the stream in the same basin or for nearby basins of similar characteristics, the synthetic unit hydrograph method needs to be used.

$$Q_p = \frac{640 \times C_p \times A}{C_t \times (L \times L_c)^{0.3}} \quad (19)$$

where Q_p is the peak flow rate in English unit, cubic feet per second (cfs), 640 is the parameter (it is 2.75 for metric system), C_p is a storage coefficient ranging from 0.4 to 0.8 where larger values of C_p are associated with smaller values of C_t , A is the basin area in

square miles (mi^2), C_t is a coefficient ranging from 1.8 to 2.2, L is the length of the basin outlet to the basin divide in miles (mi), and L_c is the length along the main stream to a point nearest the basin centroid (in mi).

More details about the unit hydrograph method could be found in Ramírez (2000).

2. Soil Conservation Service (SCS) Method

The SCS method is also widely used for estimating runoff at both gauged and ungauged basins. It has been adopted as the required procedure by many municipal and regional authorities. It was developed originally as a procedure to estimate runoff volume and peak discharge for design of soil conservation works and flood-control projects. It could be described as follows,

$$Q = \frac{(P - I_a)^2}{P - I_a + S} \quad (20)$$

where Q is the actual runoff, P is the actual precipitation, I_a is the initial precipitation losses, and S is potential maximum retention after runoff begins. The unit of each item is inches.

The initial precipitation losses include infiltration, soil moisture storage, and evapotranspiration as mentioned above. It can be determined from observed rainfall-runoff events for small basins, where lag time is minimal, as the rainfall that occurs before runoff begins. Interception and surface depression storage may be estimated from cover and surface conditions, but infiltration during the early part of the storm is highly variable and dependent on such factors as rainfall intensity, soil crusting, and soil moisture. Establishing a relationship for estimating I_a is not easy. Thus, I_a is assumed to be a function of the maximum potential retention, S . In the SCS method, an empirical relationship between I_a and S was expressed as

$$I_a = 0.2S \quad (21)$$

Therefore, the equation could be simplified as

$$Q = \frac{(P - 0.2S)^2}{P + 0.8S} \quad (22)$$

For convenience and to standardize application of this equation, the maximum potential retention is expressed in the form of a dimensionless runoff curve number CN, where

$$CN = \frac{1000}{10 + S} \quad (23a)$$

If S is in millimeters:

$$CN = \frac{1000}{10 + \frac{S}{25.4}} \quad (23b)$$

The variability in the CN results from rainfall intensity and duration, total rainfall, soil moisture conditions, cover density, stage of growth, and temperature. Its practical range is from 40 to 98. CN is also used only as an integer value.

The CN is related to soil type, soil infiltration capability, land use, and the depth of the seasonal high water table. To account for different soils' ability to infiltrate, NRCS has divided soils into four hydrological soil groups. In practical use, the CN value could be found in National Engineering Handbook Hydrology Chapters 9:(<http://www.nrcs.usda.gov/wps/portal/nrcs/detailfull/national/water/?cid=stelprdb1043063>).

2.6 River Routing

Hydrologic routing is a way to predict how water moves from an upper-stream location to a downstream location. There exist two kinds of routing: routing surface runoff from hillslope to the nearest stream and routing water in a river from upstream to downstream and eventually to the basin outlet. The latter is also called river routing. Here we focus on river routing.

In simple terms, river routing is a way to describe the movement of water from one point to another along a river. River routing methods account for storage as water moves through stream channels and water control structures. Generally, river routing is described as the difference between the inflow at the upstream end and the outflow at the downstream end and is equal to storage changes.

$$I(t) - O(t) = \frac{dS}{dt} \quad (24)$$

where I is the inflow at the upstream, O is the outflow at the downstream, and S is the storage in the river.

Solution of this equation for $O(t)$ with various approximations for the storage constitutes lumped flow routing. Both graphical and mathematical techniques for solving this equation have been used. The advantage of lumped flow routing is its relative simplicity compared to physically flow routing. However, lumped flow routing methods for rivers neglect backwater effects and are not accurate for rapidly rising hydrographs routed through mild to flat sloping rivers. Those methods simulate stage and discharge in stream channels. There are several lumped river routing methods used in conceptual hydrological models. Some common techniques include the simple stage-storage method and the modified PULS method and the slightly more complicated Muskingum method.

1. The Stage-Storage Method

The simplest method of flood routing defines the storage in terms of the mean gage height in the reach. Thus a gage-height record for both ends of the reach must be available if the flood is to be routed. The necessary stage-storage relation is, generally defined on the basis of past flood-discharge records, although in certain reaches it may be defined from topographic data. It could be described as follows,

$$O = I - A \frac{\Delta h}{\Delta t} \quad (25)$$

where A is the average area of water surface in the reach during time Δt and Δh is the average change in water surface elevation in the reach in time Δt .

The water-surface area at a given stage is the slope of the stage-storage curve and may be easily computed from a stage-storage table, since it is the “first difference.” Hence, the slope or first difference is a function of the mean stage in the reach. The outflow may be computed from the mean stage, the rate of change of stage, and the inflow.

2. The Modified PULS

The Modified PULS routing method utilizes the simple concept that storage is a function of outflow. Correct computation of the outflow hydrograph rests on the assumption that storage depends primarily, if not solely, on outflow rate. For this reason, the Modified PULS routing method is typically used for reservoir routing where a unique storage-outflow relation is likely. Strelkoff (1980) stated that determination of this relationship is a key factor in the application of the Modified PULS method.

To perform the routing, a relationship between storage and outflow is calculated and plotted as a curve. The following form of the continuity equation is then solved for each time step.

$$\frac{S_2}{\Delta t} + \frac{O_2}{2} = \left(\frac{S_1}{\Delta t} + \frac{O_1}{2} \right) - O_1 + \left(\frac{I_1 + I_2}{2} \right) \quad (26a)$$

or

$$\frac{2S_2}{\Delta t} + O_2 = I_1 + I_2 + \left(\frac{S_1}{\Delta t} - O_1 \right) \quad (26b)$$

where S is the storage in the river, O is the outflow at t_1 and t_2 (m^3/s), and I is the inflow at t_1 and t_2 (m^3/s), $\Delta t = t_2 - t_1$.

The Modified PULS routing method proves valid for reservoirs when the effects of a flood wave (differences in storage due to rising and falling stages)

are damped, if not eliminated, by the reservoir. The Modified PULS method can be used for channel routing in a similar manner where each subsection of the reach is considered to behave like a cascading reservoir. It is assumed that a unique and single-valued stage-storage outflow relationship exists for each reach, and that changing downstream conditions will not alter this relationship.

In the application, this method requires either a known stage-storage-discharge relationship, or hydraulic geometry data adequate to calculate this relationship for each reach. An appropriate computation time step also must be selected, which requires an estimate of the travel time through the reach.

The Modified PULS method is available in the HEC-1 model. The HEC-1 program allows the user to enter the storage-outflow relationships directly, or they may be computed from eight-point cross-sectional data provided to the model. If cross section data are entered, the normal depth is calculated for each cross section using Manning's equation. The danger in using this method is that downstream effects cannot be taken into account for each cross section. Also, if this eight-point cross section is not truly representative of the reach, then the stage-storage relationships cannot be developed accurately.

3. The Muskingum Method

In a stream channel (river), a flood wave may be reduced in magnitude and lengthened in travel time, i.e., attenuated, by storage in the reach between two sections. The storage in the reach may be divided into two parts – prism storage and wedge storage, since the water surface is not uniform during the floods. The volume that would be stored in the reach if the flow were uniform throughout, i.e., below a line parallel to the stream bed, is called “prism storage” and the volume stored between this line and the actual water surface profile due to outflow being different from inflow into the reach is called “wedge storage.” During rising stages, the wedge storage volume is considerable before the outflow actually increases, while during falling stages, inflow drops more rapidly than outflow, the wedge storage becoming negative.

In the case of stream-flow routing, the solution of the storage equation is more complicated, than in the case of reservoir routing, since the wedge storage is involved. While the storage in a reach depends on both the inflow and outflow, prism storage depends on the outflow alone and the wedge storage depends on the difference ($I - O$). A common method of stream flow routing is the Muskingum method (McCarthy 1938) where the storage is expressed as a function of both inflow and outflow in the reach as

$$S = K [xI + (1 - x) O] \quad (27)$$

where I is the inflow at the upstream, O is the outflow at the downstream, K is slope of storage – weighted discharge relation and has the dimension of time, and x is a dimensionless constant which weights the inflow and outflow.

It assumes that the water-surface profile is uniform and unbroken between the upstream and downstream points on the reach, that the stage and discharge are uniquely defined at these two places, and that K and x are sensibly constant throughout the range in stage experienced by the flood wave.

The factor x is chosen so that the indicated storage volume is the same whether the stage is rising or falling. The value of x ranges from 0 to 0.50 with a value of 0.25 as average for natural river reaches. If $x = 0.5$, the storage depends equally on inflow and outflow. If $x = 0$, the storage depends only on the outflow, as in the case of a large body of water such as a reservoir. No way is known for determining the value of x from the hydraulic characteristics of a channel system in the absence of discharge records.

The factor K has the dimension of time and is the slope of the storage-weighted discharge relation, which in most flood problems approaches a straight line. Analysis of many flood waves indicates that the time required for the center of mass of the flood wave to pass from the upstream end of the reach to the downstream end is equal to the factor K . The time between peaks only approximates the factor K . Ordinarily, the value of K can be determined with much greater ease and certainty than that of x .

After determining the values of K and x , the outflow O from the reach may be obtained by combining and simplifying the two equations.

$$\left(\frac{I_1 + I_2}{2}\right) \times t - \left(\frac{Q_1 + Q_2}{2}\right) \times t = S_2 - S_1 \quad (28a)$$

$$S_2 - S_1 = K [x (I_2 - I_1) + (1 - x) (O_2 - O_1)] \quad (28b)$$

For a discrete time interval, the following equation may be obtained

$$O_2 = C_0 I_2 + C_1 I_1 + C_2 O_1 \quad (29a)$$

$$C_0 = -\frac{Kx - 0.5t}{K - Kx + 0.5t} \quad (29b)$$

$$C_1 = -\frac{Kx + 0.5t}{K - Kx + 0.5t} \quad (29c)$$

$$C_0 = -\frac{K - Kx - 0.5t}{K - Kx + 0.5t} \quad (29d)$$

where t is the routing period. The routing period should be less than the time of travel for the flood wave through the reach; otherwise, it is possible that the wave crest may pass completely through the reach during the routing period.

3 Typical Conceptual Hydrological Models

Below we provide a brief introduction of three popular conceptual models: the Tank model, the Xinanjiang model, and the Sacramento model. These models are formulated from different conceptual views of the rainfall-runoff processes. The Tank model assumes that the watershed consists of a series of linear reservoirs (i.e., tanks), with each tank representing certain physical processes (i.e., evaporation and surface runoff, interflow, subsurface flow, and baseflow). The key concept contained in the Xinanjiang model is the empirical water storage capacity curve which plays a critical role in computing infiltration, runoff, evaporation, and soil water storage change. The Sacramento model also uses linear reservoir concepts. But the relationships between the reservoirs are much more complicated than that of the Tank model. There are two reservoirs representing the shallow soil water storages and three reservoirs representing the deep soil water storages, with each depicting certain physical processes such as evapotranspiration, surface runoff, interflow, and fast and slow baseflow.

3.1 The Tank Model

The Tank model (Sugawara 1972) is known as a lumped conceptual model, and is recommended by the World Meteorological Organization (WMO) as a hydrological forecasting model (WMO 1981). The model is simple and practical, and has been successfully applied to river basins in Asia, Africa, Europe, and USA (WMO 1981).

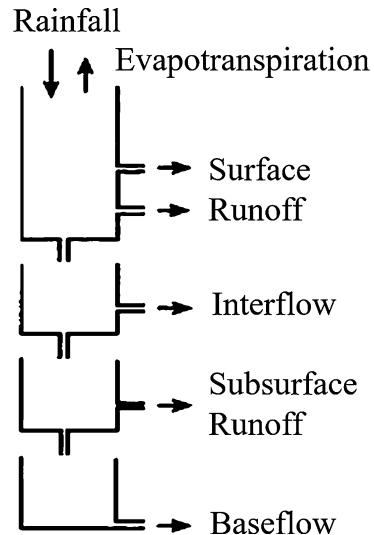
A Tank model is a simple concept that uses one or more tanks are illustrated as reservoirs in a watershed that considering rainfall as the input and generate the output as the surface runoff, subsurface flow, intermediate flow, sub-base flow, and base flow as output, as well as the phenomenon of infiltration, percolation, deep percolation, and water storages in the tank can be explained by the model. Generally, it is composed of a few (usually four) tanks laid vertically in series as shown in Fig. 4.

The top tank has two side outlets corresponding to the conceptual structure of the surface discharge, and one bottom outlet representing the infiltration. The second and third tanks have two outlets each, while the fourth tank has only one outlet.

Precipitation is put in the top tank, and evaporation or evapotranspiration is subtracted from the top tank. If there is no water in the top tank, evaporation or evapotranspiration is subtracted from the second tank; if there is no water in both the top and the second tank, evaporation or evapotranspiration is subtracted from the third tank; and so on.

The output from the side outlets are the calculated runoff. The output from the second tank is as intermediate runoff, the output from the third tank as sub-base runoff, and the output from the fourth tank as base flow.

Fig. 4 Schematic diagram of the Tank model



Water in the second tank partly moves to the stream channel through the side outlet and this corresponds to the interflow. This model structure may be considered to correspond to the zonal structure of the surface and subsurface water. The process of water inflow to soil is considered as infiltration and if the infiltration is constant the percolation is appeared.

3.2 The Xinanjiang Model

The Xinanjiang model was developed in 1973 and was initially used for streamflow forecasting (Zhao et al. 1995). The model is based on the concept of saturation excess overland flow, i.e., runoff occurs when soil content in the zone of aeration reaches field capacity. Runoff then is the rainfall excess infiltration without further loss. Saturation area where runoff generates is described by an empirical storage capacity curve (Fig. 5a):

$$\alpha = 1 - \left(1 - \frac{WM'}{WMM} \right)^b, \quad (30)$$

where α is the fraction of area where storage less or equal to WM' ; b is a parameter related to the watershed characteristics; WM' is the water storage in the zone of aeration at each point in the basin; WMM is the maximum of WM' ; and the mean watershed storage (WM) is estimated as

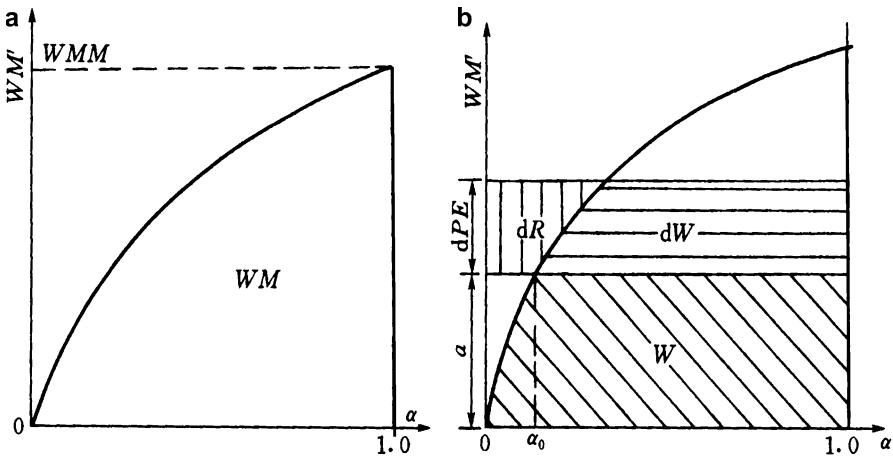


Fig. 5 Illustration of the Xinanjiang model storage capacity curve and runoff generation
(a) Storage capacity curve. **(b)** Runoff generation

$$WM = \frac{WMM}{1 + b}. \quad (31)$$

For an initial watershed storage of W and effective precipitation of PE , runoff generation (Fig. 5b) is calculated by

$$R = P + W - WM + WM \left(1 - \frac{P + a}{WMM} \right)^{b+1}, \quad (32)$$

where a is the maximum value of W .

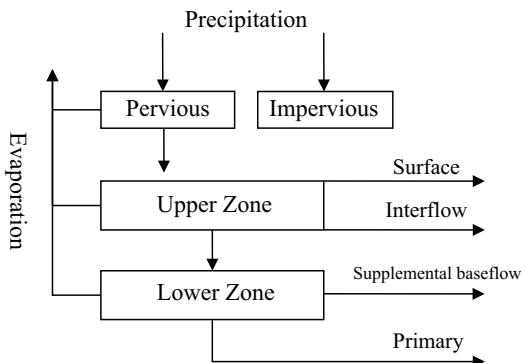
Evaporation in the Xinanjiang model is usually estimated based on pan-evaporation (or potential evaporation) by an adjustment coefficient.

In the three-source Xinanjiang model, total runoff consists of surface runoff (RS), subsurface runoff (RI), and deep ground runoff (RG). Surface runoff generates when soil water excess its storage capacity (SM), and subsurface runoff $RI = KI \times SM$, and ground runoff $RG = KG \times SM$; when soil water is not excess its storage capacity, no surface runoff generates, and $RI = KI \times PN$, $RG = KG \times PN$, where PN is net rainfall, KI and KG are outflow coefficients for subsurface and ground runoff, respectively.

Unit hydrograph method is used for surface runoff routing and linear reservoir model is used for subsurface or ground runoff routing. Muskingum method is used for streamflow routing after runoff reaches river channels. Vertical movement of subsurface runoff is characterized by the Darcy law.

The Xinanjiang model has been widely used in humid and semihumid regions in China (Zhao 1992) and has great impact on later hydrological

Fig. 6 Structure of Sacramento model



model studies in China. Infiltration excess runoff was further introduced in the Xinanjiang model to enhance its application in arid regions (Bao 1995). The storage capacity curve was also used in the VIC model (Liang et al. 1994).

3.3 The Sacramento Model

Sacramento Soil Moisture Accounting (SAC-SMA) model was initially developed by the US National Weather Service (Burnash 1995). In the Sacramento model, precipitation on the impervious area will be the direct runoff. In each basin, the pervious area is represented vertically by two zones (Fig. 6): (i) an upper zone for short-term storage and (ii) a lower zone for the longer ground water storage. In the upper zone, precipitation first supplies as tension water until it reaches the tension water capacity (UZTWM), the excess then is the effective precipitation (PAV), equal to $P + UZTWC + UZTWM$, where UZTWC is the current tension water storage in the upper zone.

Direct runoff from the impervious area is counted as a proportion of precipitation: $P \times PCTIM$, where PCTIM is the fraction of impervious area. Variable impervious area also consists of two zones but no free water in the zones. Runoff from the variable impervious area is estimated as $PAV \times \frac{(ADIMC - UZTWC)}{LZTWM}^2$, where ADIMC is the total tension water storage of the two zones, UZTWC is the tension water storage in the upper zones, LZTWM is the free water storage capacity of the lower zone.

Surface runoff generates on the pervious area when the free water in the upper zone reaches its capacity (UZFWM), and the effective precipitation (PAV) then is $P + UZFWC + UZFWM$, where UZFWC is the current free water storage. Surface runoff (ADSUR) is estimated as $PAV \times PAREA$, where PAREA is the fraction of pervious area in a basin.

The volume of water laterally moves through the soil in the upper zone to provide the interflow, which is supposed to be linearly related to storage: $UZFWC \times UZK \times PAREA$, where UZK is the outflow coefficient.

Ground water in the lower zone consists of supplemental baseflow which supplements the baseflow after a period of relatively recent rainfall, and primary baseflow which is very slow draining and providing baseflow over a long time. Both the two baseflows drain independently by subjecting to the Darcy's Law, and are assumed to be linearly related to storage in the lower zone. The daily supplemental baseflow is $LZFSC \times LZSK \times PAREA$, where $LZFSC$ is supplemental baseflow storage, and $LZSK$ is the outflow coefficient.

Potential evaporation (PE) is estimated with adjusted pan-evaporation. Evaporation in the Sacramento model is estimated based on PE, and the evaporable water first from the tension water in the upper zone (E1), then from the free water in the upper zone (E2), and finally from the tension water in the lower zone (E3). Evaporation from surface water is estimated according to water area and PE.

Percolation in the Sacramento model is estimated before the interflow computation. Horizontal interflow occurs only when precipitation rate is greater than percolation rate from the upper zone free water. The vertical water movement from the upper zone to lower zone is driven by the lower zone percolation demand. The percolation into the lower zone is equal to the runoff draining out from the lower zone if the lower zone is saturated. Percolation rate is the greatest when the upper zone is saturated and the lower zone is dry. When the upper zone is not saturated, the percolation is controlled by the storage in upper zone free water. Percolation into the lower zone will first separate as tension water and free water, and the latter then divide into the supplemental baseflow and primary baseflow.

Direct runoff and surface runoff drain into river channel directly, while draining of interflow and baseflow into river is simulated by linear reservoir method. Flow routing in river channel is usually simulated by dimensionless unit hydrograph. The Muskingum method is recommended for complex river channels conditions.

References

- E.A. Anderson, *A Point Energy and Mass Balance Model of a Snow Cover* (U.S. Government printing office, Washington, DC, 1976), pp. 6–22
- R.G. Allen, L.S. Pereira, D. Raes, M. Smith, Crop evapotranspiration – Guidelines for computing crop water requirements, in *FAO Irrigation and Drainage Paper 56*, (Food and Agriculture Organization of the United Nations, Rome, 1998)
- W.M. Bao, *Conceptual Watershed Runoff and Sediment Model and Its Application in Loess Regions* (Hohai University Press, Nanjing, 1995)
- K.J. Bevin, Robert E. Horton's perceptual model of infiltration processes. *Hydrol. Process.* **18**, 3447–3460 (2004)
- R.J.C. Burnash, The NWS River Forecast System – Catchment Modeling, in *Computer models of watershed hydrology*, rev edn., ed. by V.P. Singh (Ed) (Water Resources Publications, Highlands Ranch, 1995). <http://www.wrpllc.com/books/cmwh.html>
- V.T. Chow, D.R. Maidment, L.W. Mays, *Applied Hydrology* (McGraw-Hill, New York, 1988)
- A.J. Clemmens, Infiltration equations for border irrigation models. In: Advances in infiltration, Proceedings of the National Conference, Chicago, 1983
- D. Hillel, *Environmental Soil Physics* (Academic Press, San Diego, 1998)
- R. Hock, Temperature index melt modelling in mountain areas. *J. Hydrol.* **282**, 104–115 (2003)

- H.N. Holtan, A concept for infiltration estimates in watershed engineering. USDA, Agricultural Research Service Publication, 41–51 (1961)
- R.E. Horton, An approach towards a physical interpretation of infiltration capacity. *Soil Science Society of America* **5**, 399–417 (1940)
- V. Koren, The potential for improving lumped parameter models using remotely sensed data. 13th Conference on Weather Analysis and Forecasting, Vienna, 1993, pp 397–400
- A.N. Kostiakov, On the dynamics of the coefficients of water percolation in soils. Six Comm. Int. Soc. Soil Sci. **Part A**, 15–21 (1932)
- X. Liang, D.P. Lettenmaier, E.F. Wood, S.J. Burges, A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.* **99**(D7), 14415–14428 (1994)
- M.R. Lewis, The rate of infiltration of water in irrigation practice. *Trans. Am. Geophys. Union* **18**, 361–368 (1937)
- D.R. Maidment, *Handbook of Hydrology* (McGraw-Hill, New York, 1993)
- G.T. McCarty, The unit hydrograph and flood routing, in *US Army Corps of Engineers, Proceeding of conference of North Atlantic Division* (US Engineers Office, 1938)
- V.J. Mezencev, Theory of formation of the surface runoff [Russian]. *Meteorologia e Hidrologia* **3**, 33–40 (1948)
- V. Novotny, H. Olem, *Water Quality: Prevention, Identification, and Management of Diffuse Pollution* (Van Nostrand Reinhold, New York, 1994)
- H.L. Penman, Natural evaporation from open water, bare soil, and grass. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **193**(1032), 120–145 (1948)
- A.G. Price, T. Dunne, Energy balance computations of snowmelt in a subarctic area. *Water Resour. Res.* **12**(4), 686–694 (1976)
- J.A. Ramírez, Prediction and Modeling of flood hydrology and hydraulics, in *Land Flood Hazards*, 1st edn., (Cambridge University Press, Cambridge, 2000), pp. 293–333
- W.J. Rawls, L.R. Ahuja, D.L. Brakensiek, A. Shirmohammadi, Infiltration and soil water movement, in *Handbook of Hydrology*, ed. by D.R. Maidment (Ed), (McGraw-Hill, New York, 1993)
- L.A. Richards, Capillary conduction through porous mediums. *Physics* **1**, 313–318 (1931)
- J.C. Schaake, V.I. Koren, Q.Y. Duan, K. Mitchell, F. Chen, Simple water balance model for estimating runoff at different spatial and temporal scales. *J. Geophys. Res. Atmos.* **101**(D3), 7461–7475 (1996)
- T. Strelkoff, *Modified Puls routing in Chuquatonchee Creek* (Hydrologic Engineering Center, US Army Corps of Engineers, 1980), pp. 12–18
- A.H. Thiessen, Precipitation averages for large areas. *Mon. Weather Rev.* **39**, 1082–1084 (1911)
- C.W. Thornthwaite, An approach toward a rational classification of climate. *Geogr. Rev.* **38**(1), 55–94 (1948)
- WMO (World Meteorological Organization), Tank model, HOMS component J04.1.01 (1981) <http://www.wmo.int/pages/prog/hwrp/homs/Components/English/j04101.htm>
- R.J. Zhao, X.R. Liu, V. Singh, The Xinanjiang model, in *Computer Models of Watershed Hydrology*, ed. by V.P. Singh (Ed), (Water Resources Publications, Colorado, 1995), pp. 215–232
- R.J. Zhao, Y.L. Zhuang, L.R. Fang, X.R. Liu, Q.S. Zhao, The Xinanjiang model, hydrological forecasting proceedings Oxford symposium. *IAHS* **129**, 351–356 (1980)
- R.J. Zhao, The Xinanjiang model applied in China. *J. Hydrol.* **135**(1–4), 371–381 (1992)



Distributed Hydrological Models

Yangbo Chen

Contents

1	Introduction	414
2	Model Structures	415
2.1	Basin Division	416
2.2	Terrain Property Data Preparation	416
2.3	Flow Network	417
2.4	Hydrological Processes	417
3	Methodologies for Calculating Hydrological Processes	418
3.1	Interception and Evapotranspiration	418
3.2	Runoff Formation and Movement	419
3.3	Runoff Routing	421
4	Parameter Determinations	423
4.1	Parameter Classification	424
4.2	Scalar Method	424
4.3	Automated Parameter Optimization	424
5	Case Study	425
5.1	Studied Basin and Hydrological Data	425
5.2	Terrain Property Data	425
5.3	Liuxihe Model Construction	426
5.4	Determination of the Initial Parameter Values	427
5.5	Automated Parameter Optimization	430
5.6	Model Validation	431
6	Conclusion	434
	References	434

Y. Chen (✉)

Department of Water Resources and Environment, Sun Yat-sen University, Guangzhou, Guangdong Province, China

e-mail: escyb@mail.sysu.edu.cn

Abstract

Physically based distributed hydrological models (PBDHMs), the development of which has been facilitated by advancements in GIS and remote sensing, meteorology, computer science and engineering, and other related science and engineering disciplines, divide the terrain of a basin into fine-resolution cells and calculate the hydrological processes at both the cell and basin scales. Numerous PBDHMs have been proposed. Because PBDHMs can model hydrological processes at a fine resolution and physically derive model parameters from the properties of the terrain, they have the potential to simulate/predict hydrologic processes more effectively, and they can be employed within ungauged basins. Following a brief review of the development of PBDHMs, this chapter introduces the general structures and methodologies of currently utilized PBDHMs. The basin division method, the sources of terrain property data used to construct PBDHMs, and the flow network delineation method are summarized, and the hydrological processes within watersheds, including interception, evapotranspiration, runoff formation and movement, and runoff routing, are discussed. The methodologies most commonly employed by PBDHMs are then introduced, including those used to calculate the interception, evaporation, runoff formation, and runoff routing. Parameter determination methods are discussed, three of which are introduced in detail: the physically based method, the scalar method, and the automated optimization method. Finally, a case study is presented that demonstrates the entire procedure of constructing a Liuxihe model for a river basin flood simulation/prediction to provide the reader with a complete example of the application of a PBDHM to a real-world problem.

Keywords

Distributed hydrological model · Hydrological process · Flow network · Terrain property · GIS · Remote sensing · Uncertainty · Parameter optimization · Flood forecasting · Liuxihe model · Particle swarm optimization

1 Introduction

A distributed hydrological model is a type of hydrological model that divides the terrain of an entire studied basin into numerous cells and then characterizes the hydrological processes, including interception and evapotranspiration, snowmelt, infiltration, and runoff formation and movement at both the cell and basin scales. Distributed hydrological models were modified from lumped hydrological models, which regard the entirety of a basin as uniform, i.e., all of the hydrological processes occur at the basin scale. The development of distributed hydrological models from lumped models was facilitated by advancements in GIS and remote sensing, meteorology, computer sciences and engineering, and other related science and engineering disciplines. Distributed hydrological models are usually physically based, and

thus, they are also known as physically based distributed hydrological models (PBDHMs) (Chen et al. 2011). PBDHMs assign different model parameters to different cells in consideration of terrain property impacts on hydrological processes at the cell scale. Thus, PBDHMs better represent basin characteristics and hydrological processes, and they have the potential to simulate hydrologic responses more effectively (Ambroise et al. 1996). PBDHMs are physically based models, which means that they can physically derive the model parameters from the properties of the terrain. Consequently, there is no need to calibrate the model parameters using long series of observed data, and PBDHMs can be utilized for ungauged basins.

The blueprints for PBDHMs were initially published by Freeze and Harlan (1969), but the first complete PBDHM (i.e., the SHE model) was not published until 1987 (Abbott et al. 1986a, b). Subsequently, due to their rapid development, many PBDHMs have been proposed, including the WATERFLOOD model (Kouwen 1988), VIC model (Xu et al. 1994), DHSVM model (Wigmota et al. 1994), CASC2D model (Julien et al. 1995), WetSpa model (Wang et al. 1997), GBHM model (Yang et al. 1997), WEP-L model (Jia et al. 2001), Vflo model (Vieux and Vieux 2002), tRIBS model (Vivoni et al. 2004), WEHY model (Kavvas et al. 2004), and the Liuxihe model (Chen et al. 2011).

This chapter, which is divided into four main sections, seeks to introduce the general structures and methodologies of currently utilized PBDHMs by providing a synthesis of typical PBDHMs. Section 2 introduces the model structures (i.e., the methodologies) used to construct typical PBDHMs, including the basin division method, the sources of terrain property data for the model construction, the flow network delineation method, and the hydrological processes considered in most PBDHMs. Section 3 introduces the methods employed by most PBDHMs to calculate those hydrological processes, including interception and evapotranspiration, runoff formation, and runoff routing. Section 4 introduces three of the methods that are used to determine the model parameters employed by most PBDHMs: the physically based method, the scalar method, and the automated parameter optimization method. Section 5 presents a case study that exemplifies the procedure in establishing the Liuxihe model for a river basin flood simulation/prediction to provide the reader with a complete example of how to utilize a PBDHM to solve a real-world scenario. Finally, Sect. 6 concludes this chapter. Owing to page limitations, neither the individual PBDHMs nor the methods employed by the specific PBDHMs are introduced; rather, only the most popular models and methods are discussed. Interested readers are directed to specific publications if they are interested in a specific PBDHM.

2 Model Structures

The model structure refers to the methodologies used to subdivide basins and calculate the hydrological processes. Different PBDHMs clearly have different model structures.

2.1 Basin Division

The first step in the establishment of a PBDHM is to divide the basin into both horizontal and vertical cells. There are two ways to divide a basin into horizontal cells. The first method, which is known as gridded division, divides the basin into grid cells with equivalent dimensions using a gridded digital elevation model (DEM). The second method, which is known as triangular division, divides the basin into triangular cells that have different sizes. There are three advantages of gridded division: the DEM that is used for the terrain division can be easily acquired, the division can be performed automatically, and the code for running the model can be easily operated. The computational requirements for PBDHMs are immense, and thus, executing them on a large computer is necessary. For this reason, gridded division is adopted by most PBDHMs (except for the tRIBS model). Triangular division is advantageous because it considers the local topography when subdividing the terrain, and thus, the number of cells can be reduced compared with the gridded division method. However, as this division cannot be performed automatically, and because some manual operations are required, the division itself cannot be conducted easily. Additionally, since the sizes of the cells are not the same, the code is not as simple as that for gridded division.

The cells are further divided into vertical layers, and a three-layer division method is the most popular. The three layers are known herein as the upper layer, middle layer, and lower layer, respectively, although different models may adopt their own nomenclatures. The upper layer usually extends from the top of the canopy to the land surface, while the middle layer, which is also called the unsaturated zone in some models (i.e., since the soil content in this layer is time-dependent), is usually the layer below the land surface to a particular depth. The lower layer is below the middle layer. Alternative models (e.g., the SHE model) may further subdivide the lower layer into additional zones, but it is usually difficult to practically model multiple lower layers as the necessary data are very difficult to acquire.

In some PBDHMs, the cells are categorized into different types. For example, in the Liuxihe model, the cells are categorized into hillslope cells, river channel cells, and reservoir cells. For different types of cells, alternative calculations are employed for the different hydrological processes.

2.2 Terrain Property Data Preparation

After dividing the basin into a grid, the terrain property data, which mainly consist of a DEM, a soil type, and a land use/cover type, must be prepared for every cell. Currently, with the development of satellite remote sensing techniques and joint international efforts, global-scale terrain property data are widely available and can be accessed and downloaded freely via the Internet, which greatly facilitates the development and application of PBDHMs. For example, high-resolution DEMs with resolutions of approximately 90 m by 90 m and 30 m by 30 m can be downloaded from the Shuttle Radar Topography Mission (SRTM) DEM database (Falorni et al.

2005, Sharma and Tiwari 2014), which has proven effective in mountainous basins and has been widely used worldwide. The US Geological Survey (USGS) land use type database (Loveland et al. 1991, 2000) and the Food and Agriculture Organization of the United Nations (FAO) soil type database (<http://www.isric.org>) are popular databases with resolutions of 1000 m by 1000 m, and they can also be downloaded freely. The terrain property data for the example basin given in Sect. 5 were downloaded from these databases and worked well for constructing the model.

2.3 Flow Network

In a PBDHM, runoff is first produced in a cell, after which it is routed into adjacent cells until it reaches the basin outlet. The runoff is routed from one cell to the next, which constitutes the runoff flow network for the entire basin, is known as the flow direction.

The D8 flow direction method (O'Callaghan and Mark 1984; Jensen and Domingue 1988) is widely used in PBDHMs to derive a flow network. As there are eight neighboring cells for every cell in a gridded division scheme, the D8 method assumes that there are eight possible flow directions for every cell. These possible flow directions, which represent the direct runoff routing from the center of a cell to the center of a neighboring cell, are labeled East, Southeast, South, Southwest, West, Northwest, North, and Northeast, and they are denoted by one of eight integers: 1, 2, 4, 8, 16, 32, 64, or 128. However, during real-time runoff routing, a cell takes a single flow direction that represents flow to a neighboring cell with the lowest adjacent elevation.

The flow network inclusively determines the runoff routing for the entire basin, and it is further subdivided into hillslope runoff routing and river runoff routing. The runoff routing in a river channel is calculated using the river channel routing method, while the runoff routing on a hillslope is calculated using the hillslope routing method. The methods employed to calculate runoff routing for different models are variable and will be discussed in the next section.

The flow accumulation is usually employed to extract a river channel from a flow network. Given a threshold value for the flow accumulation (e.g., FA0), the flow direction in a cell is regarded as a river channel if the flow accumulation in the cell is larger than FA0. Clearly, the value of FA0 plays a key role in extracting river channels from flow networks, and different FA0 thresholds will result in different river channels. Therefore, during practical modeling endeavors, an extracted river channel should be compared with an available river channel system, and the results should be contrasted with those derived using different FA0 values in order to choose an appropriate FA0 threshold in consideration of trade-offs among the results.

2.4 Hydrological Processes

The hydrological processes in PBDHMs are further subdivided into several sub-hydrological processes and can usually be categorized as interception and evapotranspiration, snowmelt, and runoff formation and movement. Runoff

formation processes can be further subdivided into infiltration, surface runoff formation, subsurface runoff formation, and underground runoff formation, while runoff movement is usually divided into hillslope routing and river routing.

3 Methodologies for Calculating Hydrological Processes

3.1 Interception and Evapotranspiration

3.1.1 Interception

Falling precipitation will first be intercepted by the vegetation canopy. Specified vegetation types exhibit definitive storage capacities; after this storage capacity is reached, the precipitation will then pass through the canopy layer to the land surface, which is known as the net precipitation. This assumption is widely adopted by PBDHMs, and the storage capacity is usually calculated with the following equation (Dickinson 1984):

$$W_m = K \cdot \text{LAI} \quad (1)$$

where W_m is the canopy storage capacity, LAI is the leaf area index, and K is a constant, which is recommended to be set at 0.2 mm (Xu et al. 1994). The primary purpose of some models is to study flood runoff. Since the volume of interception is considerably small relative to that of the total runoff, interception is usually neglected within such models (e.g., the Vflo and Liuxihe models).

In PBDHMs, the canopy interception is calculated at the cell scale; if there is more than one type of vegetation within a single cell, then the interception quantity is calculated according to each type of vegetation.

3.1.2 Evaporation from the Canopy

Stored water within the canopy is evaporated through canopy evaporation. In PBDHMs, water intercepted by the canopy is believed to evaporate according to a potential evaporation capacity that is determined by the vegetation type.

The canopy evaporation is also calculated at the cell scale. If the PBDHM has a very fine resolution, then only one vegetation type is considered within a single grid cell; however, if the resolution is coarser, then the evaporation must be calculated for each of the several vegetation types and then summed for the grid cell.

3.1.3 Canopy Transpiration

After the intercepted water within the canopy is evaporated, the vegetation will undergo transpiration. Water stored in the soil, i.e., the middle layer, is taken up by the roots of the vegetation and depleted via canopy transpiration. This is a continuous process that partly causes the soil water to change dynamically, which is why the middle layer is also known as the unsaturated zone.

The most common method used to calculate the canopy transpiration in a PBDHM is the Penman-Monteith equation (Monteith 1965; Abbott et al. 1984b):

$$E_a = \frac{R_n \Delta + \frac{Q C_p \delta_e}{r_a}}{\lambda \left[\Delta + \gamma \left(1 + \frac{r_c}{r_a} \right) \right]} \quad (2)$$

where E_a is the actual transpiration, R_n is the net radiation, Δ is the rate of increase with temperature of the saturation vapor pressure of water at air temperature, Q is the density of air, r_a is the aerodynamic resistance to water vapor transport, λ is the latent heat of the evaporation of water, γ is the psychrometric constant, and r_c is the canopy resistance to water transport.

As stated above, if the model is primarily utilized for flood forecasting purposes, the canopy transpiration may not be calculated, or it may be calculated simply (e.g., in the Vflo model, the evaporation and transpiration quantities are typically not calculated for flood forecasting). Meanwhile, in the Liuxihe model, a comparatively simple method that requires fewer vegetation and soil property data is employed to calculate the evapotranspiration (Chen et al. 2011) as follows:

$$\begin{aligned} E &= \lambda E_p \text{ if } \theta > \theta_{fc} \\ E &= \lambda E_p \frac{\theta - \theta_w}{\theta_{fc} - \theta_w} \text{ if } \theta_w < \theta \leq \theta_{fc} \\ E &= 0 \text{ if } \theta < \theta_w \end{aligned} \quad (3)$$

where E is the actual evaporation, θ_{fc} is the soil water content under field conditions, θ_w is the soil water content under wilting conditions, θ is the current soil water content, E_p is the potential evaporation, and λ is the evaporation coefficient that is determined by the vegetation type. For river and reservoir cells, the evaporation coefficient is 1, while for other vegetation types, it takes a value between 0 and 1.

3.2 Runoff Formation and Movement

Precipitation passes through the canopy, reaches the land surface, and then infiltrates into the soil to compensate for the soil water deficit in the middle layer (i.e., the unsaturated zone). Runoff exists in three forms: surface runoff, subsurface runoff, and underground runoff.

3.2.1 Infiltration and Surface Runoff Routing

There are two mechanisms that govern infiltration processes. The first is the excess-runoff mechanism that assumes the existence of an infiltration capacity. When the precipitation that reaches the land surface exceeds the infiltration capacity, the precipitation will then infiltrate into the soil at the infiltration capacity to compensate for the soil water deficit, and the surplus precipitation (i.e., exceeding the infiltration capacity) will comprise the surface runoff. The most commonly used method to estimate the infiltration capacity is the Green and Ampt equation, which is written as follows (Rawls et al. 1983; Julien et al. 1995):

$$f = K \left(1 + \frac{H_f M_d}{F} \right) \quad (4)$$

where f is the infiltration rate; K is the hydraulic conductivity; H_f is the capillary pressure head at the wetting front; M_d is the soil moisture deficit, which is equal to $(\theta_e - \theta_i)$; θ_e is the effective porosity, which is equal to $(\phi - \theta_r)$; ϕ is the initial soil porosity; θ_r is the residual saturation; θ_i is the initial soil moisture content; and F is the total infiltration depth.

The second mechanism is the saturation-runoff mechanism, which assumes that the infiltration capacity is sufficiently large insomuch that all precipitation will infiltrate into the soil before it is saturated, and thus, no surface runoff will be formed. Following the saturation of the soil, a constant infiltration capacity is used to represent the soil, and all precipitation is converted into surface runoff except for the precipitation that infiltrated into the soil according to the constant infiltration capacity.

Surface runoff is routed toward the river outlet in two forms: hillslope runoff routing and river runoff routing. Hillslope runoff routing is the surface runoff that flows along hillslopes before entering a river channel, while river runoff routing is the surface runoff flowing within the river channel. Different methods are employed to calculate these two forms of surface runoff routing, and they will be described in detail in the following subsection.

3.2.2 Subsurface Runoff and Movement

Subsurface runoff is the water stored within the middle layer (i.e., the unsaturated zone) and is controlled by infiltration, evapotranspiration, and subsurface runoff flow. In most PBDHMs, only vertical subsurface flow is considered. For example, only vertical flow is assumed in the SHE model, and it is modeled using the one-dimensional Richards equation:

$$C \frac{\partial \psi}{\partial t} = \frac{\partial}{\partial z} \left(K \frac{\partial \psi}{\partial z} \right) + \frac{\partial k}{\partial z} - S \quad (5)$$

where ψ is the soil moisture tension or pressure head, t denotes the time, z is the vertical spatial coordinate (positive upward), C is the soil water capacity, θ is the volumetric water content, $K(\theta, z)$ is the hydraulic conductivity, and S is the source/sink term for both the root extraction and soil evaporation.

Some models also consider horizontal subsurface flow when the soil water content is high. For example, in the Liuxihe model, a constant vertical water flow and a horizontal flow are modeled if the water content in the soil layer reaches a minimum value.

3.2.3 Underground Runoff and Movement

Water in the lower layer (i.e., the saturated layer) is considered underground runoff. In most PBDHMs, only horizontal flow is considered (e.g., in the SHE model), and each cell is modeled using the nonlinear Boussinesq equation:

$$S \frac{\partial h}{\partial t} = \frac{\partial}{\partial x} \left(K_x H \frac{\partial h}{\partial x} \right) + \frac{\partial}{\partial y} \left(K_y H \frac{\partial h}{\partial y} \right) + R \quad (6)$$

where $S(x,y)$ is the specific yield; $h(x,y,t)$ is the phreatic surface level; $K_x(x,y)$ and $K_y(x,y)$ are the saturated hydraulic conductivities in the x and y directions, respectively; $H(x,y,t)$ is the saturated thickness; t denotes the time; x and y are the Cartesian coordinates in the horizontal plane; and $R(x,y,t)$ is the instantaneous vertical recharge into the saturated zone.

Large quantities of data are required to solve the above equation due to its complexity. As a consequence, many PBDHMs do not consider underground runoff in practice, or they simply utilize a lumped underground flow model.

3.3 Runoff Routing

3.3.1 Hillslope Routing Methods

Theoretically, runoff flow is governed by the Saint-Venant equations of continuity and momentum with three-dimensional hydrodynamic flow; however, since the Saint-Venant equations are partial differential equations, fully solving these equations with the hydrodynamic method takes an enormous amount of computational time, particularly since the runoff routing must be performed cell-by-cell. Therefore, to ensure the method is practical, the complete Saint-Venant equations must be simplified, after which some effective algorithms are necessary to solve the simplified equations. Two types of simplifications are currently applied in PBDHMs for surface runoff routing. The first type, as is adopted by the SHE model, treats surface runoff flow as a two-dimensional flow; meanwhile, for the second type, all of the other PBDHMs treat surface runoff flow as a one-dimensional flow.

For the SHE model, the two-dimensional flow Saint-Venant equations are written as follows (Abbott et al. 1984b):

$$\frac{\partial h}{\partial t} + \frac{\partial(uh)}{\partial x} + \frac{\partial(vh)}{\partial y} = q \quad (7)$$

$$\frac{\partial h}{\partial x} = S_{0x} - S_{fx} \quad (8)$$

$$\frac{\partial h}{\partial y} = S_{0y} - S_{fy} \quad (9)$$

where $h(x,y)$ is the local water depth, t denotes the time, x and y are the Cartesian coordinates in the horizontal plane, $u(x,y)$ and $v(x,y)$ are the flow velocities in the x and y directions, $q(x,y,t)$ is the net precipitation minus infiltration, $S_{0x}(x,y)$ and $S_{0y}(x,y)$ are hillslopes along the x and y directions, and $S_{fx}(x,y)$ and $S_{fy}(x,y)$ are friction slopes along the x and y directions.

In the SHE model, the above equations are solved using the explicit procedure described by Preissmann and Zaoui (1979). In practice, this method is very time

consuming and exhibits a problem with convergence, and thus, it is currently only used in the SHE model.

For all other PBDHMs, surface runoff is treated as a one-dimensional flow along the slope. To greatly reduce the calculation time required, the kinematic wave approximation is employed to solve the equation, which is simplified below:

$$\frac{\partial Q}{\partial x} + L \frac{\partial h}{\partial t} = q \quad (10)$$

$$S_f = S_0 \quad (11)$$

where Q is the surface flow, h is the surface flow depth, q is the lateral flow, L is the length of the cell, S_0 is the hillslope, and S_f is the friction slope.

The above equations can be solved using different methods. For example, the finite element algorithm is employed in the Vflo model, while the equation in the Liuxihe model is solved using the Newton iteration algorithm.

Even with the abovementioned simplifications, the surface runoff routing calculation is still very complex and time consuming, and thus, surface runoff routing is neglected in some PBDHMs. Consequently, the produced runoff is regarded as directly flowing into a river channel without runoff routing. Meanwhile, some other PBDHMs employ lumped surface runoff routing methods.

3.3.2 River Runoff Routing

River runoff flow is treated as a one-dimensional flow in all PBDHMs, and the diffusive wave approximation is employed to solve the Saint-Venant equations, which are simplified below:

$$\frac{\partial Q}{\partial x} + L \frac{\partial h}{\partial t} = q \quad (12)$$

$$\frac{\partial h}{\partial x} = S_0 - S_f \quad (13)$$

The meanings of the variables in these equations are the same as previously defined. The above equations can also be solved using a variety of methods, but almost all PBDHMs employ the same algorithm to solve for river channel runoff routing with the exception of the SHE model, which employs the MIKE 11 modeling package to conduct full dynamic routing calculations.

In addition to enormous computational requirements, PBDHMs also face challenges during river runoff routing regarding measurements of the shapes of river channels and their cross-sectional sizes. These measurements are particularly difficult or are otherwise impossible to acquire within mountainous watersheds, and thus, PBDHMs are not applicable in many cases due to an absence of essential river channel information. To solve this problem, the river channel shape is assumed to be a trapezoid in the Liuxihe model. Consequently, the river channel cross-sectional size can be measured using two indices for the bottom width and the side slope.

Furthermore, the river channel is divided into virtual sections, and the cross-sectional sizes of the river channel within the same virtual sections are assumed to be equivalent, which greatly simplifies the river channel runoff routing calculation and increases the computational efficiency of the Liuxihe model. In addition, the river channel bottom width is estimated by referencing satellite remote sensing data, and thus, this method can be utilized in regions without field survey data.

4 Parameter Determinations

Since most model parameters cannot be measured directly, they must be estimated through specific estimation techniques (Laloy et al. 2010; Leta et al. 2015). PBDHMs are physically based models, and thus, their model parameters also possess physical meanings. As a consequence, these parameters can be directly derived from the properties of the terrain. Currently, there are no widely accepted references for deriving these parameters from terrain data; users usually determine the model parameters from alternative references or limited experimental or field results, which are “point” based. This method is herein known as the physically based method. For example, in the Liuxihe model, the parameters are divided into unadjustable and adjustable parameters. The flow direction and slope are two unadjustable parameters that are derived from the DEM, and they remain unchanged. The flow direction is determined using the D8 method introduced above, while the slope is the hill slope along the flow direction that can also be calculated using the DEM. The other parameters are all adjustable parameters that can be adjusted further to improve the model performance. The evaporation capacity is a climate-type parameter, the value of which is set by referencing observations in a watershed and is usually set to 5 mm/day. The evaporation coefficient and roughness are land use-type parameters; the former is a less-sensitive parameter in the Liuxihe model and is usually set to 0.7. The roughness parameter is derived from alternative references or local experimental data. Among the different soil-type parameters, a value of 2.5 is recommended for the parameter b , and the soil water content under wilting conditions is set to 30% of the soil water content under field conditions. The values of the other soil-type parameters are calculated using a hydraulic properties calculator for soil water characteristics (Arya and Paris 1981), which calculates the soil water content under saturated and field conditions and the hydraulic conductivity under saturated conditions based on the soil texture, organic matter, gravel content, salinity, and compaction. These calculations can be performed using the program developed by Keith E. Saxton, which can be downloaded from <http://hydrolab.arsusda.gov/soilwater/Index.htm>.

Due to the lack of experimental and field validation research, it has been found that model parameters determined in this manner have high uncertainties in practice. Thus, parameter optimizations are needed to reduce their uncertainties (Gupta et al. 1998; Madsen 2003; Vieux and Moreda 2003; Reed et al. 2004; Smith et al. 2004; Pokhrel et al. 2012; Chen et al. 2016). The scalar method (Vieux et al. 2003), which represents the first effort toward this goal, was proposed to adjust the Vflo model

parameters and was able to improve the model performance. Since this method is performed manually, it is generally very tedious and time consuming. Subsequently, automated parameter optimization methods were developed that improved the efficiencies and capabilities of parameter optimization techniques (Madsen 2003; Shafii and Smedt 2009; Chen et al. 2016).

4.1 Parameter Classification

The parameter values in PBDHMs are related to terrain properties, including the topography, soil type, and vegetation type, and thus, these values can be determined through the properties of the terrain (Chen et al. 2016). The model parameters of all PBDHMs are usually classified into one of four different types: climate-related parameters, topography-related parameters, vegetation (land use)-related parameters, and soil-related parameters. With this classification, the parameters in different cells will have the same values if they have the same terrain properties, and thus, they can be determined using a physically based method.

4.2 Scalar Method

The scalar method was first utilized for the Vflo model (Vieux et al. 2003). In the scalar method, every parameter value that is derived using a physically based method is manually adjusted with a factor or a multiplier (scalar). The scalars for the same parameter categories among different cells are taken to have the same values, so only a few parameters must be adjusted. The scalar method is simple, but it must be performed manually, and it is consequently tedious and time consuming. However, it has been proven capable of improving the model performance.

4.3 Automated Parameter Optimization

For the SHE model, an automatic parameter optimization method using a shuffled complex evolution (SCE) algorithm (Duan et al. 1994) was employed to simulate the catchment runoff (Madsen 2003) in consideration of two objectives: fitting the surface runoff at the catchment outlet and minimizing the error in the simulated underground water level at different wells. To simulate the runoff processes within a medium-sized catchment with the WetSpa model, a multi-objective genetic algorithm was used to optimize the model parameters (Shafii and Smedt 2009). An automated parameter optimization method based on a particle swarm optimization (PSO) algorithm was proposed for the Liuxihe model and was found to be effective (Chen et al. 2016). The PSO algorithm has three steps. The first step is to classify all of the parameters into a few independent parameters (i.e., which will subsequently be optimized) using the same classification method as described above. The second

step is to initialize and normalize the parameters. The parameter initialization is employed to derive the initial model parameter values using the physically based method described above. Then, the parameters are normalized with their initial values as follows:

$$X_i = X'_i/X_{i0} \quad (14)$$

where X'_i is the original value of a parameter i , x_{i0} is the initial value of a parameter i , and x_i is the normalized value of a parameter i . Every parameter becomes unit-less variable during the normalization process.

The third step is to automatically optimize the independent parameters using an optimization algorithm (i.e., PSO algorithm). The objective function for the optimization algorithm minimizes the peak flow relative error of the catchment discharge at the outlet.

To reduce the computational cost for a large watershed, the parameters sensitive to the model performance may be identified first, after which only relatively sensitive parameters are optimized.

5 Case Study

In this section, a case study is introduced that employs the Liuxihe model to simulate the flood processes of a river basin. The purpose of this case study is to demonstrate the procedure for constructing a PBDHM to simulate/predict the hydrological processes in a river basin; the procedure for which is the same for other PBDHMs.

5.1 Studied Basin and Hydrological Data

The studied basin is the Taiping Watershed, which is a second-order tributary of the Ganjiang River Basin in Jiangxi Province. The Taiping Watershed is 51.6 km long with a drainage area of 445 km². It is a mountainous watershed with frequent flash flooding events. Figure 1 displays a sketched map of the Taiping Watershed.

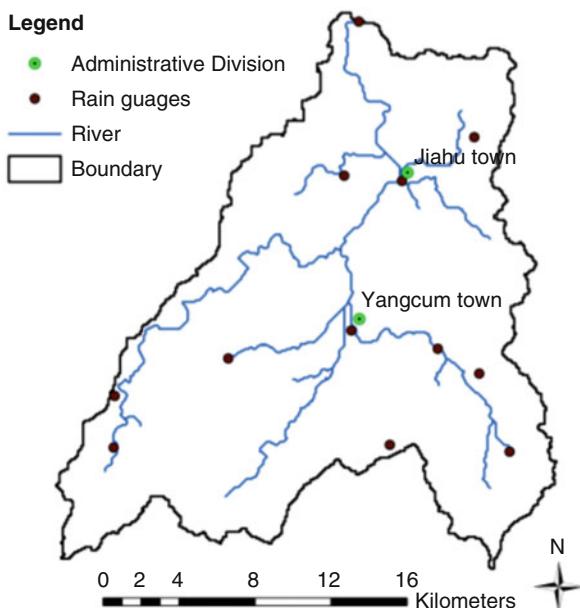
There are 12 rain gauges installed throughout the watershed that automatically collect precipitation. One river gauge is installed at the outlet of the watershed, and it is capable of measuring the discharge continuously. The locations of the rain gauges and river gauge are shown in Fig. 1.

Hydrological data of six flood events observed over the past years, including precipitation and river discharge data, have been collected for this case study.

5.2 Terrain Property Data

The terrain property data used for the construction of the Liuxihe model in this case study constitute a DEM in addition to land use types and soil types. These data for the

Fig. 1 Sketched map of the Taiping Watershed



studied watershed were downloaded from open access databases. The DEM was downloaded from the SRTM database, the land use type data were downloaded from the USGS land cover database, and the soil type data were downloaded from <http://www.isric.org>. The downloaded DEM has a spatial resolution of 90 m by 90 m, but the other two datasets have spatial resolutions of 1000 m by 1000 m. Consequently, they are rescaled to a spatial resolution of 90 m by 90 m. Figure 2 exhibits the terrain property data (i.e., the DEM, land use types, and soil types) of the Taiping Watershed.

5.3 Liuxihe Model Construction

To construct the Liuxihe model for the studied watershed, the whole basin of the Taiping Watershed is divided into 55,221 grid cells using the DEM with a grid cell size of 90 m by 90 m (as previously prepared), after which the grid cells are categorized into reservoir cells, river channel cells, and hillslope cells. As there are no significant reservoirs, no reservoir cells are derived during this process.

In this study, different FA0 thresholds are used to derive the river channel cells, the results of which are shown in Fig. 3.

To compare the results of this process with the natural river system of the Taiping Watershed, and to make it possible to estimate the river cross-sectional size using remote sensing data, a third-order river system is adopted. With this division, 1133 river channel cells and 54,088 hillslope cells are produced. Furthermore, 12 nodes are established within the Taiping Watershed, and the river channel system is divided

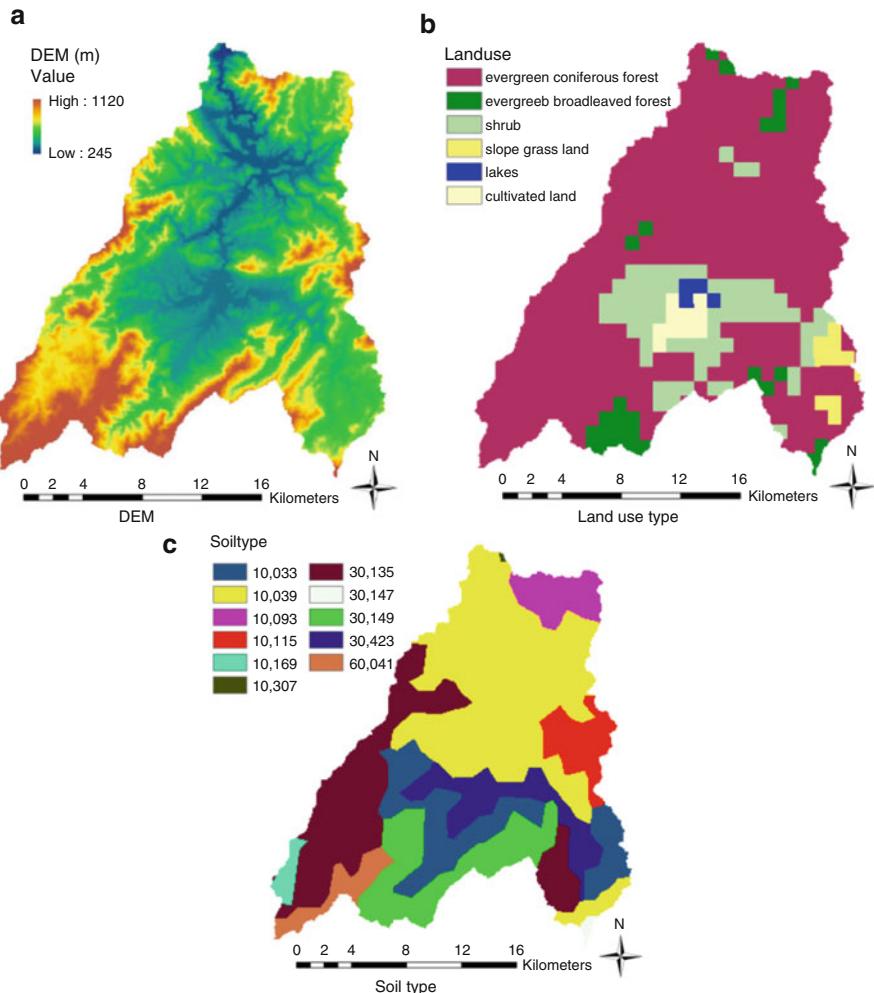


Fig. 2 Property data of the Taiping Watershed. (a) DEM. (b) Land use type. (c) Soil type

into 29 virtual sections, the cross-sectional sizes of which are estimated through a reference with satellite remote sensing imagery. The Liuxihe model structure of the Taiping Watershed is shown in Fig. 4.

5.4 Determination of the Initial Parameter Values

Based on the DEM shown in Fig. 2a, the flow directions and slopes of all of the cells are derived and are shown in Figs. 5 and 6, respectively.

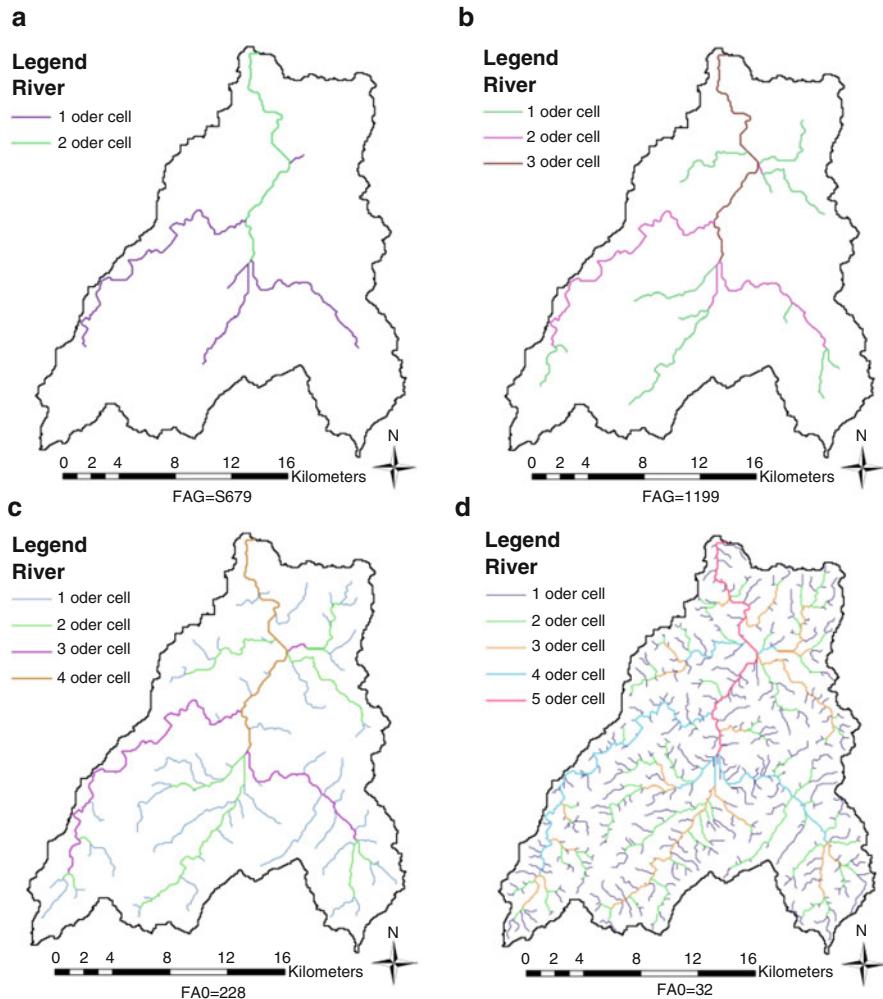


Fig. 3 River cells classification with different FA0 thresholds. (a) FA0=S679. (b) FA0 = 1199. (c) FA0 = 228. (d) FA0 = 32

The initial value of the evaporation capacity is set to 5 mm/day, and the initial value of the evaporation coefficient is set to 0.7. Meanwhile, the initial value of the roughness is derived based on a reference (Wang et al. 1997). The initial values of the parameters are listed in Table 1.

For the soil-type parameters, the variable b is set at a value of 2.5, and the soil water content under wilting conditions is taken as 30% of the soil water content under field conditions. The initial values of the soil water content under saturated and field conditions and the hydraulic conductivity under saturated conditions are determined using Keith E. Saxton's simulator based on the soil texture, organic matter,

Fig. 4 Liuxihe model structure of the Taiping Watershed

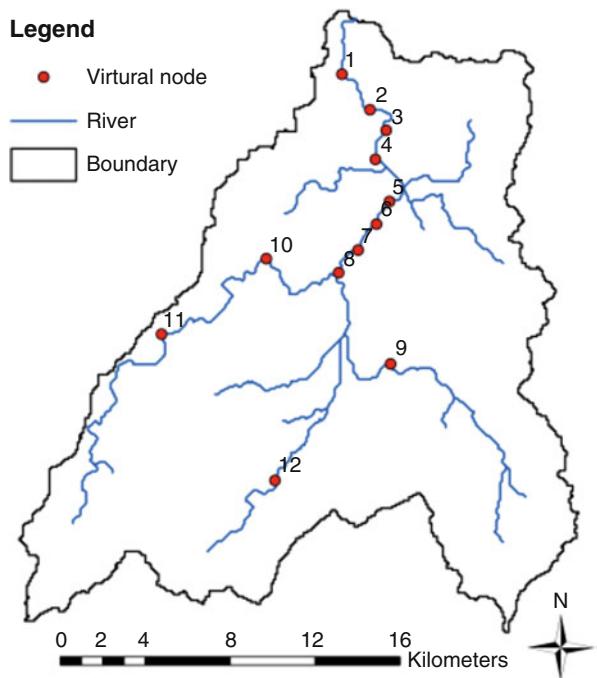


Fig. 5 Flow direction of the Taiping Watershed

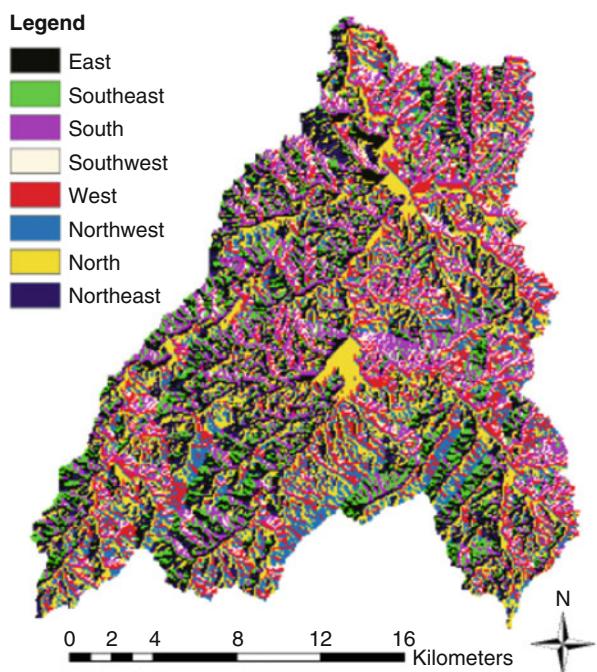


Fig. 6 Slope of the Taiping Watershed

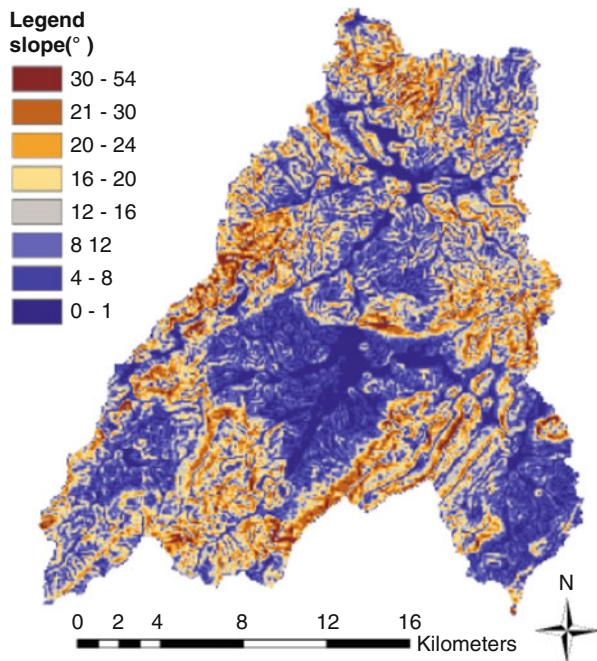


Table 1 Initial values of the land use-type parameters

ID	Name	Evaporation coefficient	Roughness coefficient
2	Evergreen coniferous forest	0.7	0.4
3	Evergreen broad-leaved forest	0.7	0.6
5	Shrub	0.7	0.4
8	Slope grassland	0.7	0.2
10	Lakes	0.7	0.045
15	Cultivated land	0.7	0.15

gravel content, salinity, and compaction of the soil type. These parameter initial values are listed in Table 2.

5.5 Automated Parameter Optimization

The flood event flood2006071409 is used to automatically optimize the model parameters with the initial values, and the particle number used is 20. Figure 7 shows both the objective and parameter evolution processes of the parameter optimization.

Table 2 Initial values of soil-type parameters

Soil type	Thickness (mm)	Water content at saturated conditions	Water content at field conditions	Hydraulic conductivity at saturated conditions (mm/h)	b
CN10033	1000	0.466	0.354	3.5	2.5
CN10039	600	0.515	0.422	1.95	2.5
CN10093	1000	0.454	0.144	74.49	2.5
CN10115	700	0.5	0.377	4.89	2.5
CN10169	1000	0.438	0.192	35.15	2.5
CN10307	1000	0.451	0.315	6.28	2.5
CN30135	1000	0.435	0.207	28.33	2.5
CN30147	1000	0.443	0.262	14.88	2.5
CN30149	1300	0.429	0.211	24.13	2.5
CN30423	670	0.446	0.24	21.87	2.5
CN60041	870	0.511	0.451	0.51	2.5

During the evolution process, the objective function rapidly decreases and converges to its optimal value (Fig. 7). After only 5 evolutions, most of the parameters nearly converged to their optimal values; after 12 evolutions, most of the parameters had converged to their optimal values. These results demonstrate that the PSO algorithm exhibits effective convergence capabilities.

Figure 7 also shows that the optimal parameter values of several of the parameters are quite different from the initial parameters, while those of others are relatively unchanged, which implies that the initial model parameters determined using the physically deriving method possess high uncertainties. A parameter optimization algorithm could reduce these uncertainties.

5.6 Model Validation

The other observed flood events of the Taiping Watershed are simulated using the model with the parameters that were optimized above to validate the model performance for flood forecasting. To analyze the effects of the parameter optimization on the model performance improvement, Fig. 8 shows three of the simulated hydrographs.

These results reveal that the model with initial parameter values is unable to simulate the observed flood events satisfactorily, i.e., the uncertainties are high. In addition, the simulated hydrographs with the optimized model parameters fit the observed hydrographs well, particularly with regard to the simulated peak flow. These results imply that the parameter uncertainties have been reduced through the model parameter optimization.

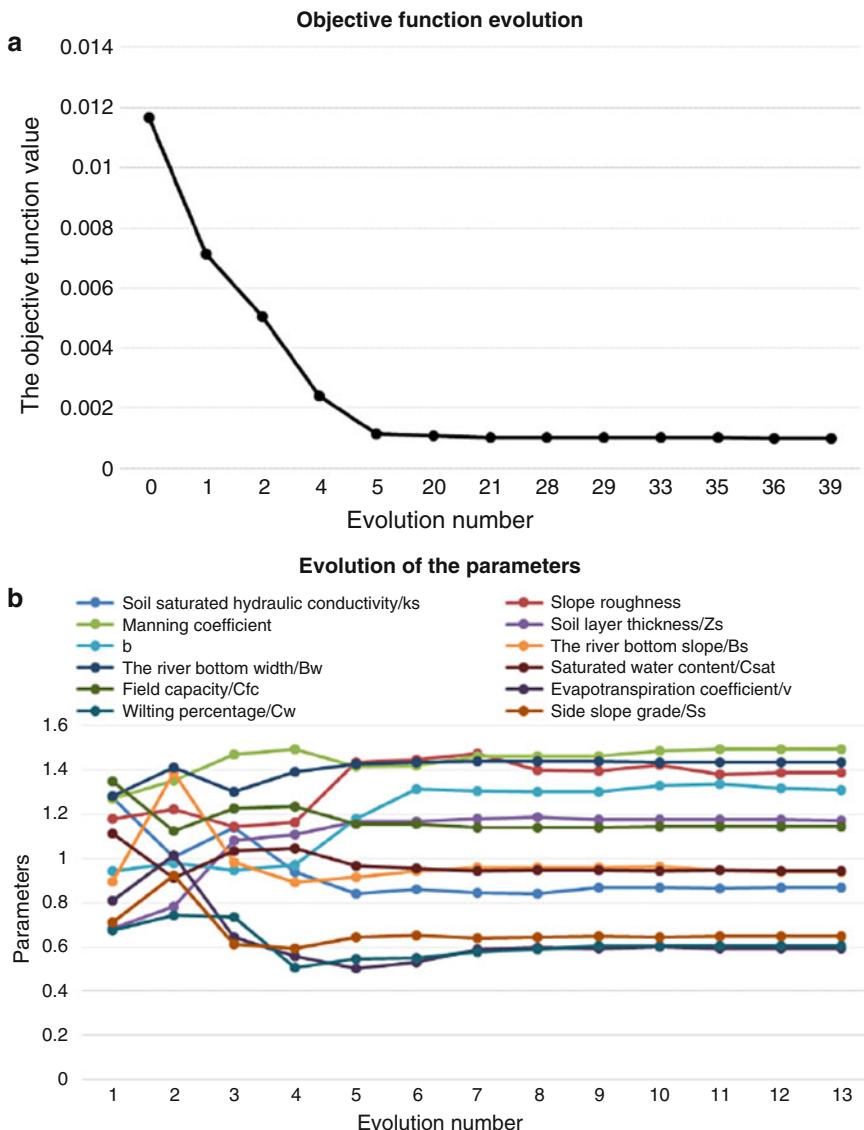


Fig. 7 The evolution processes of the parameter optimization. (a) Objective function evolution. (b) Evolution of the parameters

The above results suggest that using a PSO parameter optimization algorithm can improve the Liuxihe model performance for watershed flood forecasting in the Taiping Watershed and that optimizing the parameters of the Liuxihe model is necessary.

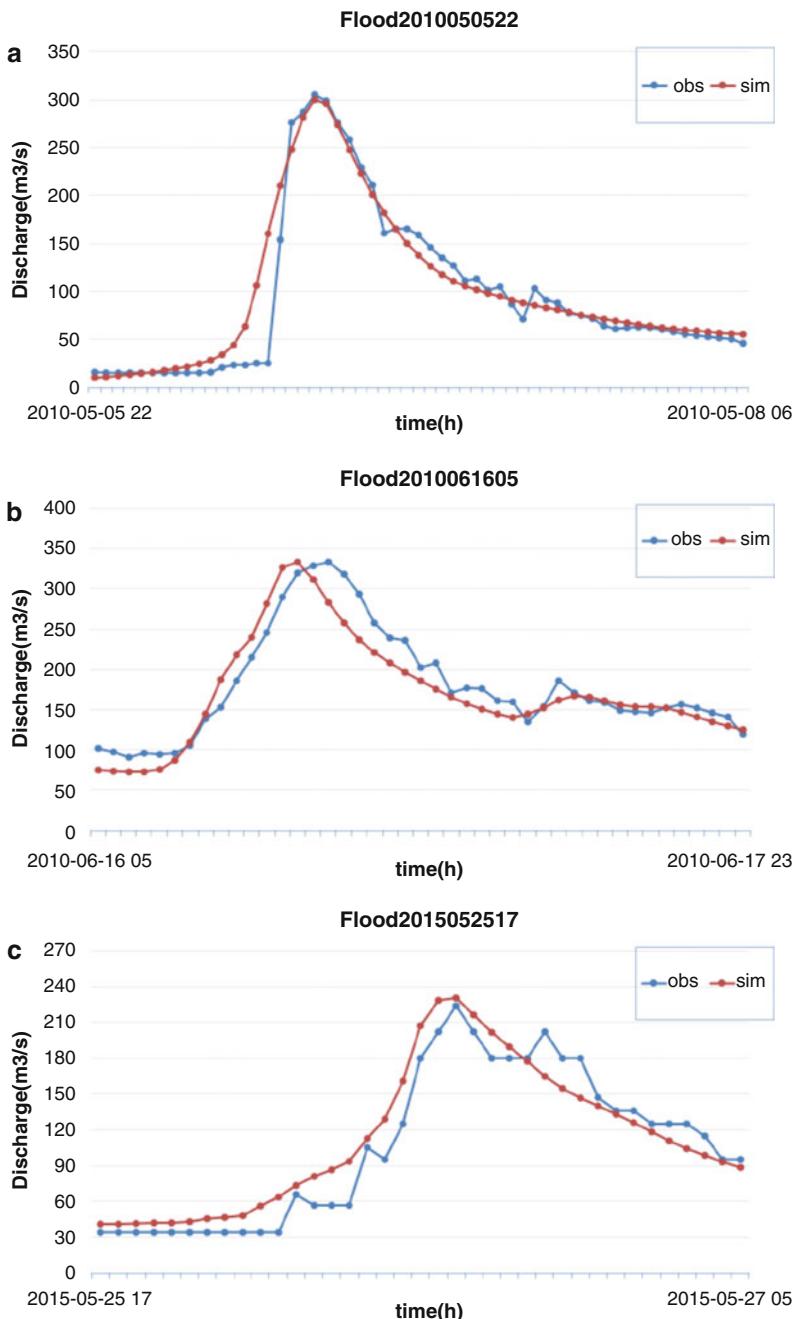


Fig. 8 Simulated flood events with optimized model parameters. (a) Flood2010050522. (b) Flood2010061605. (c) Flood2015052517

6 Conclusion

This chapter introduced the general structures and methodologies of currently employed PBDHMs, and a case study of a watershed in China was presented using the Liuxihe model to demonstrate the complete procedure of constructing a PBDHM for a river basin flood simulation/prediction experiment. The following conclusions can be summarized:

1. A number of PBDHMs have been proposed, and many successful applications of these models have been reported. Scientifically sound PBDHM structures and methodologies have been presented and implemented to satisfy the requirements of these applications. New PBDHMs are not urgently needed, but in order to improve the accuracies of existing PBDHMs, new algorithms for multiple hydrological processes, including infiltration and runoff routing, must be developed based on the advancement of hydrological principle research.
2. Data for the construction of PBDHMs with global coverage are widely available and can be accessed and freely downloaded via the Internet. These data could be satisfactorily utilized to construct PBDHMs for both scientific studies and real-time applications, particularly in mountainous watersheds.
3. Model uncertainties are still high presently, particularly with regard to the determinations of parameters that require high hydrological process simulation/prediction accuracies, e.g., for flood simulation/forecasting applications. Parameter optimization methods, which are necessary for applications requiring high accuracies, have been proposed, and they have proven very useful in improving the model performances. Validation studies in this field still need to be strengthened.
4. Algorithms with higher computational efficiencies still need to be explored. Furthermore, the development of standardized software tools is urgent, and public infrastructures, which could be shared by general users to test new PBDHMs for both scientific studies and real-time applications, should be established.

References

- M.B. Abbott et al., An introduction to the European hydrologic system-system hydrologue European, 'SHE', a: history and philosophy of a physically-based, distributed modelling system. *J. Hydrol.* **87**, 45–59 (1986a)
- M.B. Abbott et al., An introduction to the European hydrologic system-system hydrologue European, 'SHE', b: structure of a physically based, distributed modeling system. *J. Hydrol.* **87**, 61–77 (1986b)
- B. Ambroise, K. Beven, J. Freer, Toward a generalization of the TOPMODEL concepts: topographic indices of hydrologic similarity. *Water Resour. Res.* **32**, 2135–2145 (1996)
- L.M. Arya, J.F. Paris, An empirical model to predict the soil moisture characteristic from particle-size distribution and bulk density data. *Soil Sci. Soc. Am. J.* **45**, 1023–1030 (1981)
- Y. Chen, Q.W. Ren, F.H. Huang, H.J. Xu, I. Cluckie, Liuxihe model and its modeling to river basin flood. *J. Hydrol. Eng.* **16**, 33–50 (2011)
- Y. Chen, J. Li, H. Xu, Improving flood forecasting capability of physically based distributed hydrological model by parameter optimization. *Hydrol. Earth Syst. Sci.* **20**, 375–392 (2016)

- Dickinson, R. E., Modelling evapotranspiration for three-dimensional global climate models, in *Climate Processes and Climate Sensitivity*, Geophys. Monogr. Ser., ed. J. E. Hansen, T. Takahashi, vol. 29 (AGU, Washington, DC, 1984)
- Q. Duan, S. Sorooshian, V.K. Gupta, Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.* **158**, 265–284 (1994)
- G. Falorni, V. Teles, E.R. Vivoni, R.L. Bras, K.S. Amaralunga, Analysis and characterization of the vertical accuracy of digital elevation models from the Shuttle Radar Topography Mission. *J. Geophys. Res. F: Earth Surf.* **110**(2), F02005 (2005)
- R.A. Freeze, R.L. Harlan, Blueprint for a physically-based, digitally simulated, hydrologic response model. *J. Hydrol.* **9**, 237–258 (1969)
- R.B. Grayson, I.D. Moore, T.A. McMahon, Physically based hydrologic modeling: 1.A terrain-based model for investigative purposes. *Water Resour. Res.* **28**, 2639–2658 (1992)
- H.V. Gupta, S. Sorooshian, P.O. Yapo, Toward improved calibration of hydrological models: multiple and non-commensurable measures of information. *Water Resour. Res.* **34**(4), 751–763 (1998)
- S.K. Jensen, J.O. Domingue, Extracting topographic structure from digital elevation data for geographic information system analysis. *Photogramm. Eng. Remote Sens.* **54**(11), 1593–1600 (1988)
- Y. Jia, G. Ni, Y. Kawahara, Development of WEP model and its application to an urban watershed. *Hydrolog. Process.* **15**, 2175–2194 (2001)
- P.Y. Julien, B. Saghafian, F.L. Ogden, Raster-based hydrologic modeling of spatially- varied surface runoff. *Water Resour. Bull.* **31**, 523–536 (1995)
- M. Kavvas, Z. Chen, C. Dogrul, J. Yoon, N. Ohara, L. Liang, H. Aksoy, M. Anderson, J. Yoshitani, K. Fukami, T. Matsuura, Watershed environmental hydrology (WEHY) model based on upscaled conservation equations: hydrologic module. *J. Hydrol. Eng.* **6**(450), 450–464 (2004). [https://doi.org/10.1061/\(ASCE\)1084-0699\(2004\)9](https://doi.org/10.1061/(ASCE)1084-0699(2004)9)
- N. Kouwen, WATFLOOD: a micro-computer based flood forecasting system based on real-time weather radar. *Can. Water Resour. J.* **13**, 62–77 (1988)
- E. Laloy, D. Fasbender, C.L. Bielders, Parameter optimization and uncertainty analysis for plot-scale continuous modeling of runoff using a formal Bayesian approach. *J. Hydrol.* **380**(1–2), 82–93 (2010)
- O.T. Leta, J. Nossent, C. Velez, N.K. Shrestha, A. Griensven, W. Bauwens, Assessment of the different sources of uncertainty in a SWAT model of the River Senne (Belgium). *Environ. Model. Softw.* **68**, 129–146 (2015)
- L. Xu, D.P. Lettenmaier, E.F. Wood, S.J. Burges, A simple hydrologically based model of land surface water and energy fluxes for general circulation models. *J. Geophys. Res.* **99**, 14415–14428 (1994)
- T.R. Loveland, J.W. Merchant, D.O. Ohlen, J.F. Brown, Development of a land cover characteristics data base for the conterminous U.S. *Photogramm. Eng. Remote. Sens.* **57**(11), 1453–1463 (1991)
- T.R. Loveland, B.C. Reed, J.F. Brown, D.O. Ohlen, J. Zhu, L. Yang, J.W. Merchant, Development of a global land cover characteristics database and IGBP DISCover from 1km AVHRR data. *Int. J. Remote Sens.* **21**(6–7), 1303–1330 (2000)
- H. Madsen, Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Adv. Water Resour.* **26**, 205–216 (2003)
- J. L. Monteith, Evaporation and environment, in *The State and Movement of Water in Living Organisms. Proceedings of 15th Symposium Society for Experimental Biology*, Swansea (Cambridge University Press, London, 1965)
- J. O'Callaghan, D.M. Mark, The extraction of drainage networks from digital elevation data. *Comput. Vis. Graph. Image Process* **28**(3), 323–344 (1984)
- P. Pokhrel, K.K. Yilmaz, H.V. Gupta, Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. *J. Hydrol.* **418–419**, 49–60 (2012)
- A. Preissmann, J. Zaoui, Le Module “coulement de surface” du Systme Hydrologique Europen (SHE), in *Proceedings of 18th Congress International Association for Hydraulic Research*, Cagliari, 5 (1979), pp. 193–199
- W.J. Rawls, D.L. Brakensiek, N. Miller, Green-Ampt infiltration parameters from soils data. *J. Hydraul. Eng. ASCE* **109**(1), 62–70 (1983)

- S. Reed, V. Koren, M. Smith, Z. Zhang, F. Moreda, D.-J. Seo, DMIP participants: overall distributed model intercomparison project results. *J. Hydrol.* **298**(1–4), 27–60 (2004)
- M. Shafii, F.D. Smedt, Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. *Hydrol. Earth Syst. Sci.* **13**, 2137–2149 (2009)
- A. Sharma, K.N. Tiwari, A comparative appraisal of hydrological behavior of SRTM DEM at catchment level. *J. Hydrol.* **519**, 1394–1404 (2014)
- M.B. Smith, D.-J. Seo, V.I. Koren, S. Reed, Z. Zhang, Q.-Y. Duan, S. Cong, F. Moreda, R. Anderson, The distributed model intercomparison project (DMIP): motivation and experiment design. *J. Hydrol.* **298**(1–4), 4–26 (2004)
- B.E. Vieux, F.G. Moreda, Ordered physics-based parameter adjustment of a distributed model, in *Advances in Calibration of Watershed Models*, Water Science and Application Series, ed. by Q. Duan, S. Sorooshian, H.V. Gupta, A.N. Rousseau, R. Turcotte, vol. 6 (American Geophysical Union, Washington, DC, 2003), pp. 267–281. ISBN:0-87590-335-X (Chapter 20)
- B. E. Vieux, J. E. Vieux, VfloTM: a real-time distributed hydrologic model[A], in *Proceedings of the 2nd Federal Interagency Hydrologic Modeling Conference*. 28 July–1 Aug, Las Vegas, Nevada. Abstract and paper on CD-ROM 2002
- E.R. Vivoni, V.Y. Ivanov, R.L. Bras, D. Entekhabi, Generation of triangulated irregular networks based on hydrological similarity. *J. Hydrol. Eng.* **9**(4), 288–302 (2004)
- Z. Wang, O. Batelaan, F. De Smedt, A distributed model for water and energy transfer between soil, plants and atmosphere (WetSpa). *J. Phys. Chem. Earth* **21**, 189–193 (1997)
- M.S. Wigmosta, L.W. Vai, D.P. Lettenmaier, A distributed hydrology-vegetation model for complex terrain. *Water Resour. Res.* **30**, 1665–1669 (1994)
- D. Yang, S. Herath, K. Musiake, Development of a geomorphologic properties extracted from DEMs for hydrologic modeling. *Ann. J. Hydraul. Eng. JSCE* **47**, 49–65 (1997)



Land Surface Hydrological Models

Michael B. Ek

Contents

1	Introduction	438
2	Atmospheric Forcing Data	438
3	Land Data Sets	440
4	Land-Surface Model	444
4.1	Surface Fluxes	444
4.2	Surface Turbulent Exchange Coefficients	446
4.3	Prognostic Land States	448
4.4	Solution of Surface Energy Budget	449
4.5	Surface Temperature	455
4.6	Soil Hydraulics	456
4.7	Soil Thermodynamics	457
4.8	Cold Season Processes	459
5	Land-Atmosphere Interaction	463
5.1	Near-Surface Land-Atmosphere Interaction (NSLAI)	464
5.2	Land-ABL Interaction	471
6	Summary	474
	References	475

Abstract

The details of land-surface models (LSMs) are presented here from the perspective of providing the proper boundary condition to and interaction with a “parent” atmospheric model. Topics include atmospheric forcing to LSMs, land data sets, surface-layer turbulence, surface fluxes and energy and water budgets,

M. B. Ek (✉)

National Center for Atmospheric Research, Boulder, CO, USA

e-mail: ek@ucar.edu

land-surface physics, and the role of the land states and surface fluxes in local land-atmosphere interaction. Connections of LSMs with hydrological models (e.g., saturated zone or groundwater, and streamflow or river-routing) and land data assimilation are outside the scope of this chapter.

Keywords

Land-surface model · Land-atmosphere interaction

1 Introduction

Traditionally, from the perspective of a Numerical Weather Prediction (NWP) or a coupled atmosphere-ocean-land-ice seasonal and longer time-scale climate model, the role of a land-surface model (LSM) is to provide surface quantities as boundary conditions to the atmosphere, which includes the surface fluxes and land states (and other land-surface properties) needed to calculate the surface fluxes. In providing these necessary surface boundary conditions, the land model closes the surface energy and water budgets. The land provides predictability in weather and climate models, where land states, especially soil moisture, vegetation, and snow, can provide predictability in the window between deterministic (weather) and seasonal and longer climate (e.g., ocean-atmosphere) time scales (Fig. 1). The Noah land model (Ek et al. 2003) is described here as a useful example of those classes of land models developed in conjunction with atmospheric numerical prediction models (Fig. 2). Although not covered in this chapter, the LSM also provides the upper boundary conditions in the form of runoff that is passed to a hydrology model that accounts for the movement of water deeper into the groundwater and lateral connections that end up as stream/riverflow with an ultimate connection to the ocean.

To provide proper boundary conditions, a land model must have atmospheric forcing to drive the LSM; appropriate physics to represent land-surface processes (for the relevant temporal and spatial scales, including the correct land-atmosphere interactions) and associated LSM parameters; corresponding land data sets, e.g., land use/land cover (vegetation type), soil type, surface albedo, snow cover, surface roughness, etc.; and proper initial land states, analogous to initial atmospheric conditions, though land states may carry more memory, especially, e.g., in deep soil moisture, similar to ocean temperatures and corresponding ocean heat content.

2 Atmospheric Forcing Data

A land model is forced by incoming solar and longwave radiation; precipitation; pressure; and wind, temperature, and humidity. Atmospheric forcing may be from a parent atmospheric model (analysis or reanalysis) and/or from in situ or remotely sensed observations, where precipitation is quite important for LSMs since this affects soil moisture which in turn affects both heat and moisture flux, with solar

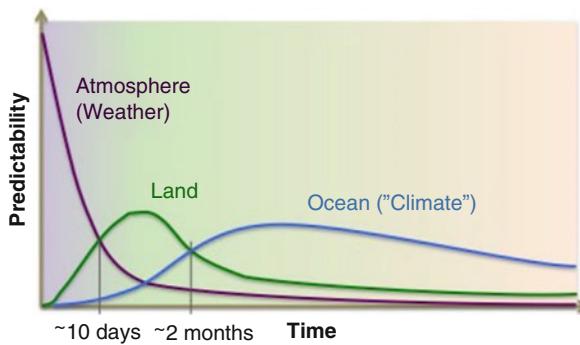


Fig. 1 Land predictability lies between atmosphere and ocean-atmosphere interaction. (Courtesy Paul Dirmeyer, George Mason University.)

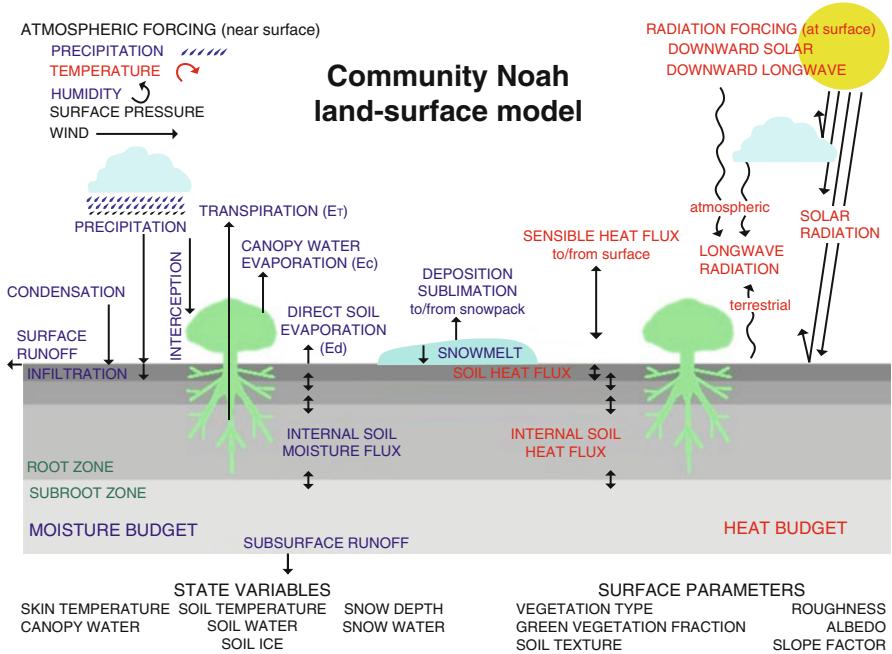


Fig. 2 Noah land-surface model. See text for details

radiation incident at the surface an additional important forcing as this drives the available energy for surface fluxes. The downward longwave radiation can be important especially during the night when small changes in this radiative forcing may affect the surface energy budget such that the surface may become “decoupled” from the atmosphere, especially over snow cover which may be “insulated” from energy from the soil beneath the snowpack. This will be affected by patchy snow

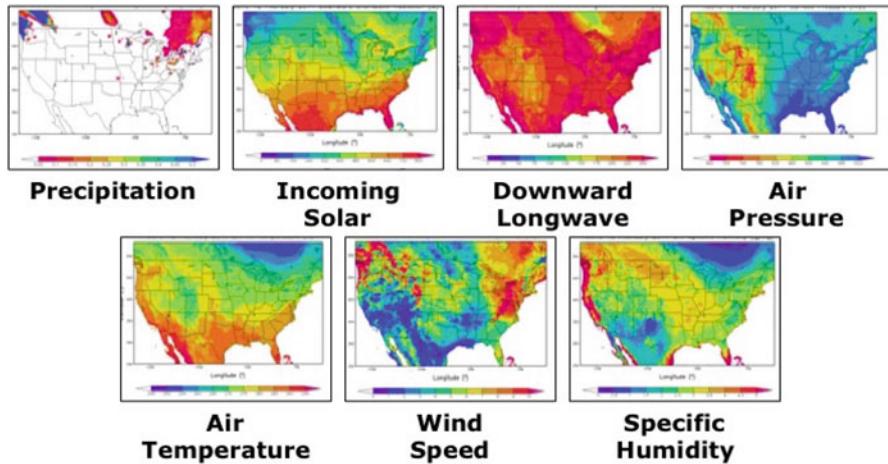


Fig. 3 Example atmospheric forcing to LSM. See text for details

cover, where the soil may then more properly “communicate” with the surface. The near-surface atmospheric variables (wind, temperature, and humidity) affect the vegetation response as well as near-surface gradients in these quantities, thereby affecting surface fluxes. This forcing data set (e.g., Fig. 3) may be spatial in manner at a particular time for an LSM run “offline” in 2-D mode, or coupled with a parent atmospheric model. Alternately, a time series of a forcing data set for a particular point can be used to provide a location-specific LSM run, i.e., 1-D offline, thereby allowing for long runs of the LSM for analysis and land model development, with minimal compute cost. Inclusion of a land data assimilation feature to ingest some or all of these forcing terms constitute a land data assimilation system, whether in an offline LSM-only or coupled with a parent atmospheric model mode.

3 Land Data Sets

Land models depend on a number of land data sets in order to properly specify the surface characteristics necessary for the execution of the LSM physics. Land-use class (or vegetation type) and soil type (or soil texture class) may be specified for a given site, and generally come from global data sets, often from a satellite product (e.g., Fig. 4; Hansen et al. 2000 for land-use, and Schwarz and Alexander 1995, and USDA Soil Survey 1995 for soils). These quantities are generally treated as static, especially soil type, but for land-use, there may be year-to-year changes (and on even shorter time scales, especially at higher resolution), e.g., in the case of urbanization, deforestation, and desertification.

Albedo is a diurnally and seasonally varying quantity, where we make use of monthly climatologies of mid-day albedo based on remotely sensed Moderate Resolution Imaging Spectroradiometer (MODIS) data for snow-free albedo

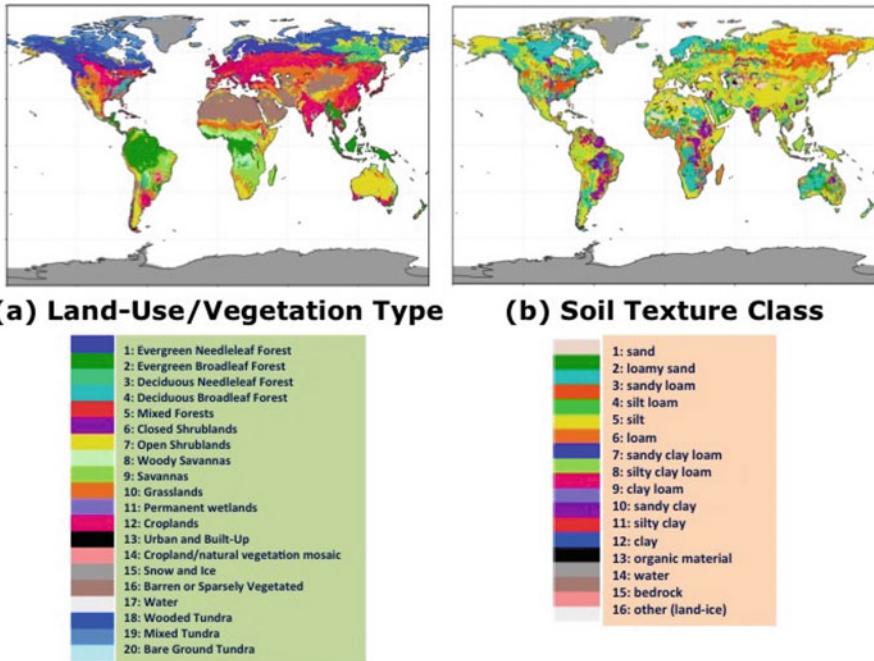


Fig. 4 (a) 1-km IGBP-MODIS land-use class/vegetation type, (b) 1-km STATSGO-FAO soil type

(Friedl et al. 2010; Fig. 5a). Over deep snow, the albedo of the surface is higher and in LSMs previously was often set to some uniformly large value (e.g., 0.70), however, this can vary greatly depending on the surface character. For example, a conifer forest may have a lower albedo due to darker treetops sticking through a brighter (deep) snowpack (depending on the vegetation fraction), compared with a higher albedo for a completely snow-covered grassland. Maximum snow albedo may be given as a function of the land-use class and vegetation fraction, or again from MODIS data as annual maximum snow albedo climatology (e.g., Barlage et al. 2005; Fig. 5b). Note the differences between the North American boreal forests with lower maximum snow albedos due to more shading of the snowpack under the canopy, compared to the U.S. Great Plains grasslands with higher maximum snow albedos due to more open ground and exposed snow cover. Albedo may also be a calculated quantity by the LSM which depends on soil moisture and texture/color, as well as solar zenith angle (as determined by, e.g., a solar elevation calculation or from a radiation code from a coupled parent atmospheric model to adjust the mid-day albedo to a value for a particular time of day).

The initial inclusion of seasonally and spatially varying vegetation climatology in LSMs was an important feature in order to more properly represent the surface energy partition (including the calculation of evapotranspiration). Unless the LSM has a prognostic calculation of vegetation phenology (vegetation cover (*green vegetation fraction* or GVF) and vegetation density (*leaf area index* or LAI)),

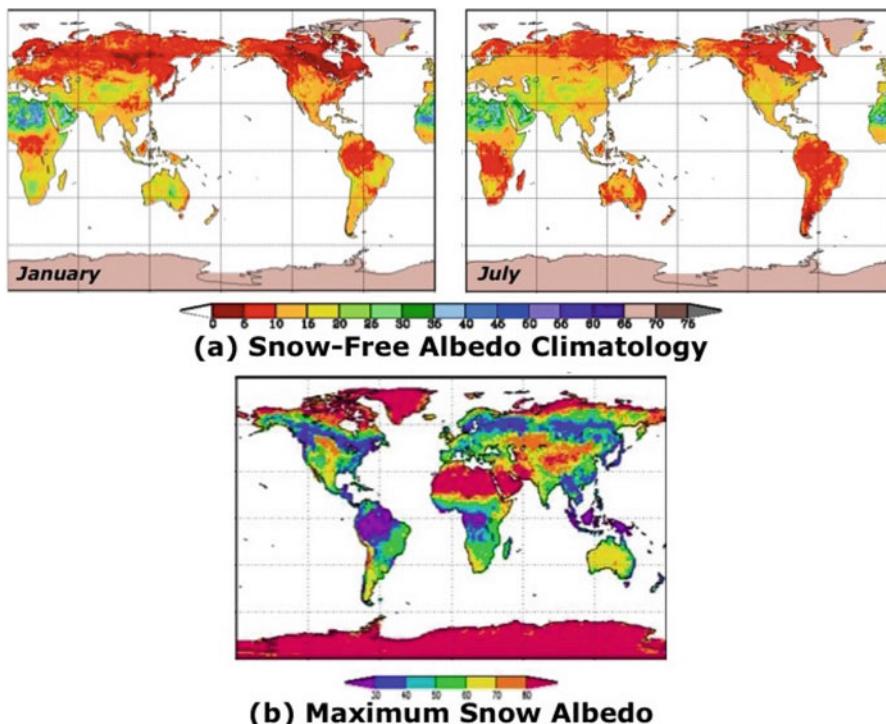


Fig. 5 (a) 1-km monthly, Boston Univ.-MODIS monthly snow-free albedo, (b) 5-km Univ. Arizona-MODIS maximum snow albedo

these quantities are provided in the form of 2-D maps or given as functions of the vegetation type, or scaled based on the seasonal phenology of GVF in the case of LAI. These quantities may come from remotely sensed data sets, e.g., long-term climatology data sets (Fig. 6a–b; Gutman 1999 for GVF) or via near-realtime observations (e.g., Fig. 6c). These data sets were generated from Advanced Very-High-Resolution Radiometer (AVHRR), MODIS, and Visible Infrared Imager Radiometer Suite (VIIRS) satellite imagery.

Snow cover is a seasonal feature for many high-latitude regions, and more ephemeral in nature at locations equatorward during the cold season. Snow data sets are often updated daily (e.g., Fig. 7) or even subdaily depending on availability of remotely sensed and in situ data. LSMs determine the onset, evolution, and ablation of snowpacks, but updating snow information based on observations is necessary, again to determine the proper surface energy and water budgets, and especially important for numerical weather prediction models that are run multiple times in a day. Snowpack physics will be discussed further below.

Longwave surface emissivity is another land data set that can be determined from remotely sensed data or specified from other land data sets, e.g., vegetation

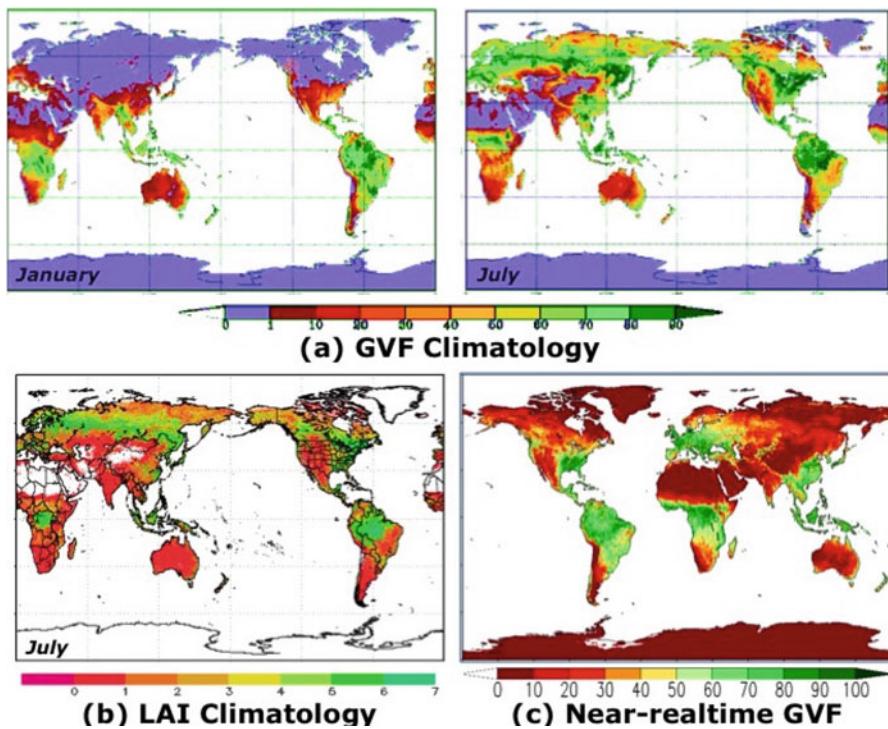


Fig. 6 (a) 16-km green vegetation fraction (GVF) multi-year climatology from AVHRR, (b) 4-km leaf area index (LAI) climatology from MODIS, (c) 4-km near-realtime (15 May 2016) GVF from VIIRS

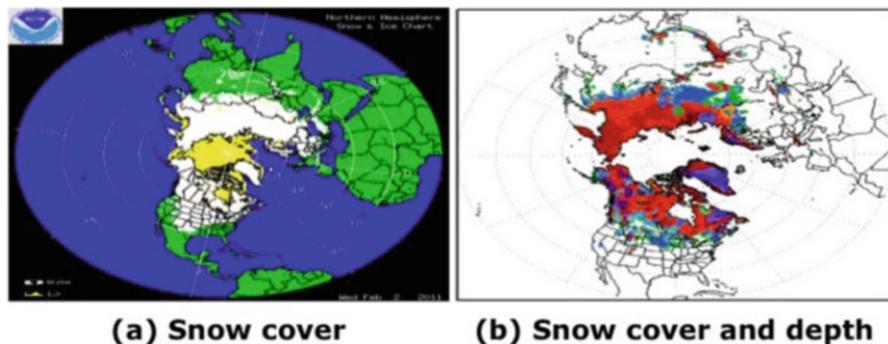


Fig. 7 (a) 4-km snow cover from the National Ice Center, (b) 16-km snow depth and cover from the US Air Force 557th Weather Wing (formerly Air Force Weather Agency, AFWA)

and soil type. Fresh snow may have a surface emissivity approaching unity (1.0), while a sandy soil with high quartz content may be less than 0.9. Surface emissivity is included in equations for solving the surface energy budget, and as a general approximation is often set to 1.0.

Other surface characteristics include soil depth and slope, all important for vegetation processes and the movement of water in the soil, with connections to runoff and groundwater hydrology.

4 Land-Surface Model

The Noah LSM has many features that are similar to other LSMs and has been previously described in Mahrt and Pan (1984) and Pan and Mahrt (1987), with updates described in Chen et al. (1996), and Ek et al. (2003). More recent updates to the Noah LSM are described in Niu et al. (2011) and Yang et al. (2011), and among these include an explicit vegetation layer and subcanopy (requiring separate energy budgets), CO₂-based photosynthesis (for canopy conductance in the calculation of transpiration, mentioned further below), a prognostic calculation of vegetation phenology (cover and density), and a multilayer snowpack, with this updated Noah LSM called “Noah-MP.” Our focus here is on the “classic” Noah LSM.

4.1 Surface Fluxes

The land-surface model determines the surface fluxes for the surface and water budgets, and the surface momentum flux, i.e.,:

$$R_n = H + LE + G, \quad (1)$$

$$\Delta S = P - R - E, \quad (2)$$

$$\tau = \rho C_m U_a^2, \quad (3)$$

where Eqs. (1) and (2) are the surface energy and water budgets, respectively, and Eq. (3) is the surface momentum flux. In the surface energy budget, the net radiation (R_n) is partitioned between the turbulent sensible (H) and latent heat (LE) fluxes to or from the surface, respectively, and the soil heat flux (G), the flux into or out of the soil.

The latent heat flux (or evapotranspiration) is composed of evaporation of canopy-intercepted water, soil, or direct (non-vegetative) evaporation and plant transpiration, and in the presence of snowpack, an additional latent heat flux associated with snow phase change. The L in the LE may be the latent heat of vaporization (L_v) or of sublimation (L_s), depending on a liquid or frozen surface (Evapotranspiration is discussed further below).

In the surface water budget, ΔS includes the change in land-surface water, i.e., soil moisture, snowpack (cold season), and canopy water (dewfall or frostfall and intercepted precipitation, that are small, but not always negligible) and is balanced by precipitation (P), runoff (R), and evapotranspiration (E), where $P - R$ is infiltration of moisture into the soil.

Finally, the surface momentum flux (3) is a function of wind speed and surface drag that depends on surface characteristics and the near-surface turbulent exchange (discussed further below). Surface characteristics include the land-use or vegetation type, vegetation density, and cover (e.g., patchy grassland vs. a dense forest canopy, presence of snow, etc.), while the surface turbulent exchange is in the form of a surface drag coefficient depends on the near-surface atmospheric stability, where unstable (stable) conditions are characterized by stronger (weaker) turbulent exchange with the near surface atmosphere, and U_a is the horizontal wind speed at some reference height in the atmosphere, (z_a) defined as

$$U_a = \sqrt{u_a^2 + v_a^2}, \quad (4)$$

where u_a and v_a are the horizontal wind components at z_a , e.g., as determined from a parent atmospheric model, or alternately via observations. Note that the grid-averaged wind speed in a model may vanish, but surface heat and moisture fluxes can be non-zero (due to subgrid horizontal motions), hence the surface flux parameterizations needs to account for free-convection conditions, e.g., by adding a minimum wind speed ($< 1 \text{ ms}^{-1}$) or the convective velocity scale (w_*) to the mean wind speed.

The surface energy budget may be broken down further as:

$$\begin{aligned} R_n &= S \downarrow -S \uparrow + L \downarrow -L \uparrow, \\ &= S \downarrow (1 - \alpha_s) + L \downarrow -\epsilon_s \sigma T_s^4, \end{aligned} \quad (5)$$

$$H = \rho c_p C_h U_a (T_s - T_a), \quad (6)$$

$$LE = \rho L_v C_q U_a (q_s - q_a), \quad (7)$$

where R_n is comprised of the incoming solar radiation ($S \downarrow$), outgoing or reflected solar radiation ($S \uparrow$), incoming atmospheric longwave radiation ($L \downarrow$), and emitted longwave radiation ($L \uparrow$). The reflected solar radiation can be expressed as the incoming solar radiation multiplied by a surface albedo (α_s , a function of surface properties such as vegetation and soil type, soil moisture, snow cover, etc.), while the emitted longwave radiation is a function of the surface emissivity (ϵ_s , also a function of surface properties, with a value near unity), the Stefan-Boltzmann constant ($\sigma = 5.67 \times 10^{-8}$), and the surface skin temperature (T_s). The sensible and latent heat fluxes depend on U_a , and the near-surface gradients in the temperature ($T_s - T_a$) and specific humidity ($q_s - q_a$), respectively, specific heat (c_p , $1004.5 \text{ J kg}^{-1} \text{ K}^{-1}$) and latent heat (L_v , assumed constant as $2.5 \times 10^6 \text{ J kg}^{-1}$, or L_s as noted further above), respectively, and the surface turbulent exchange coefficients for heat (C_h) and moisture (C_q), respectively. (Here we adopt the usual convention that $C_q = C_h$.) Note that the approach of a single effective surface temperature, encompassing surface, canopy, and snow, has been adopted which is then used in the calculation of surface fluxes. See Fig. 8.

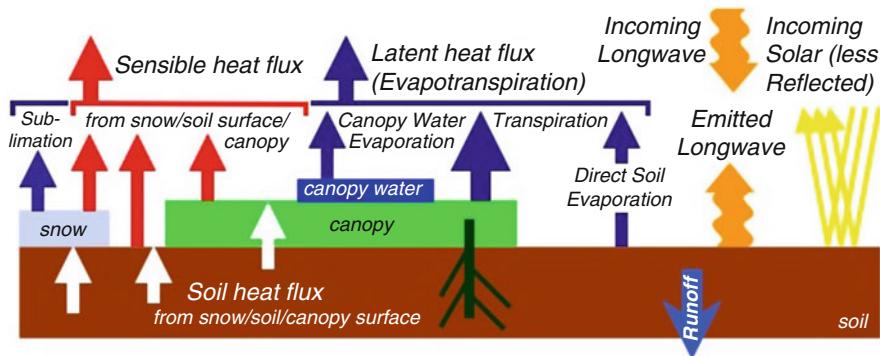


Fig. 8 Surface fluxes for different surfaces in Noah LSM

4.2 Surface Turbulent Exchange Coefficients

Surface exchange coefficients (and thus surface fluxes) are calculated by iterating an implicit formula of the Monin-Obukhov stability-dependent profile functions based on surface-layer similarity theory. This is an alternative to using the explicit approach by Louis (1979) and Louis et al. (1982) based on the near-surface bulk Richardson number. There are limitations in the Louis formulations for cases where the ratio of the momentum to heat roughness is large, as demonstrated in Holtslag and Ek (1996). These limitations have been further addressed in van den Hurk and Holtslag (1997), who suggest more accurate explicit functions based on the bulk Richardson number. We present both implicit and explicit approaches here; see Beljaars and Holtslag (1991) for a more detailed comparison.

4.2.1 Implicit Exchange Coefficients Calculation

The surface exchange coefficients for momentum and heat (and moisture) using the implicit formulations are (respectively)

$$C_m = \frac{k^2}{[\ln(z_a/z_{0m}) - \Psi_m(z_a/L) + \Psi_m(z_{0m}/L)]^2}, \quad (8)$$

$$C_h = \frac{k^2}{[\ln(z_a/z_{0m}) - \Psi_m(z_a/L) + \Psi_m(z_{0m}/L)][\ln(z_a/z_{0h}) - \Psi_h(z_a/L) + \Psi_h(z_{0h}/L)]}, \quad (9)$$

where k is the von Kármán constant (taken as 0.40), z_a is the atmospheric reference height (e.g., first atmospheric model level height), z_{0m} and z_{0h} are the roughness lengths for momentum and heat, respectively, L is the Obukhov length, and $\Psi_{m,h}$ are the stability profile functions for momentum and heat. (As with the exchange coefficient, we assume that $\Psi_q = \Psi_h$.) The profile functions for unstable conditions (following Paulson 1970) are

$$\Psi_m = 2\ln[(1+x)/2] + \ln[(1+x^2)/2] - 2\tan^{-1}(x) + \pi/2, \quad (10)$$

$$\Psi_h = 2\ln[(1+x^2)/2], \quad (11)$$

where

$$x = (1 - 16z/L)^{1/4}, \quad (12)$$

and for stable conditions (following Webb 1970) are

$$\Psi_m = \Psi_h = -5z/L. \quad (13)$$

The Webb (1970) profile functions are fairly consistent with most data for $0 < z/L < 0.5$ (see Beljaars and Holtslag 1991). Alternate profile functions for stable conditions follow Holtslag and De Bruin (1988) for up to $z/L \approx 7$

$$-\Psi_m = a\frac{z}{L} + b\left(\frac{z}{L} - \frac{c}{d}\right)\exp\left(-d\frac{z}{L}\right) + \frac{bc}{d}, \quad (14)$$

$$\Psi_h = \Psi_m, \quad (15)$$

where $a = 0.7$, $b = 0.75$, $c = 5$, and $d = 0.35$. This expression behaves like (13) for small z/L values and approaches $-\Psi_m \approx a(z/L)$ for large z/L (very stable conditions).

Typically, $z_{0m} \approx 0.01 - 0.10m$ and $z_{0m} \gg z_{0h}$ for bare soil and short vegetation (e.g., croplands, grasslands), while $z_{0m} \approx 1.0m$ and $z_{0m}/z_{0h} \approx O(1-10)$ for taller vegetation (e.g., forests) and is affected by vegetation cover, e.g., weighted in some manner by the fraction of vegetation versus bare soil.

4.2.2 Explicit Exchange Coefficients Calculation

Following Louis (1979) and Louis et al. (1982), the surface exchange coefficients for momentum and heat (and moisture) using the explicit formulations are (respectively)

$$C_m = k^2 \frac{F_m}{[\ln(z_a/z_{0m})]^2}, \quad (16)$$

$$C_h = \left(\frac{k^2}{R}\right) \frac{F_h}{\ln(z_a/z_{0m})\ln(z_a/z_{0h})}, \quad (17)$$

where R , estimated as 1.0, is the ratio of the drag coefficients for momentum and heat in the neutral limit and is taken from Businger et al. (1971). Here, C_m and C_h are functions of $F_{m,h}$ instead of $\Psi_{m,h}$. For unstable condition (modified by Holtslag and Beljaars 1989), $F_{m,h}$ are defined as

$$F_m = 1 - \frac{10Ri_b}{1 + 75k^2[\ln(z_a/z_{0m})]^{-2}[-Ri_b(z_a/z_{0m})]^{1/2}}, \quad (18)$$

$$F_h = 1 - \frac{15Ri_b}{1 + 75k^2[\ln(z_a/z_{0m})]^{-1}[\ln(z_a/z_{0h})]^{-1}[-Ri_b(z_a/z_{0m})]^{1/2}}, \quad (19)$$

and for stable conditions (modified by Holtslag and Beljaars 1989)

$$F_m = F_h = \frac{1}{1 + 10Ri_b(1 + 8Ri_b)}, \quad (20)$$

where Ri_b is the near-surface bulk Richardson number, defined as

$$Ri_b = \frac{gz_a(\theta_{av} - \theta_{sv})}{\theta_{av}U_a^2}, \quad (21)$$

where g is gravity, $\theta_{av} - \theta_{sv}$ is the virtual potential temperature gradient between the air θ_{av} at z_a and the surface θ_{sv} .

By itself, the usual similarity theory under stable conditions leads to a significant overestimation of surface cooling. This is due to (a) failure to consider subgrid-scale spatial variability where vertical fluxes can occur in part of the grid even with large (bulk) Richardson number (Ri_b) based on grid averaged variables (Mahrt 1987), (b) poor vertical resolution where turbulence may occur in thinner layers, perhaps intermittently, even when Ri_b over the model layer is large, (c) neglect of clear air radiative cooling, (d) neglect of gravity wave momentum transport, and (e) use of a surface skin temperature from the surface energy balance (as is done, instead of temperature at the roughness height) to compute the near-surface bulk Richardson number.

To compensate for such inadequacies, various mechanisms have been employed (and are often unreported) which include capping the allowable value of the Richardson number or specifying a minimum wind speed. An alternative to Eq. (20) that leads to noted improvement in model performance in the nocturnal boundary layer is the area-averaged exchange coefficient relationship of Mahrt (1987) where for stable conditions $F_{m,h}$ are defined as

$$F_m = F_h = \exp(-\alpha_m Ri_b), \quad (22)$$

where α_m is nominally set equal to 1.0. However, α_m is expected to depend on, e.g., (a) model vertical resolution, (b) wind speed, and (c) subgrid characteristics such as standard deviation of subgrid surface skin temperature, terrain height, or some other measure of the surface inhomogeneity.

4.3 Prognostic Land States

The prognostic variables that must be determined in order to calculate surface fluxes are the volumetric soil moisture content (Θ_{soil_n} , for soil layer n), and soil temperature (T_{soil_n}), and the canopy water content (C_w).

4.3.1 Soil Moisture Tendency

Soil moisture is modeled with the prognostic equation for the volumetric water content (Θ) as

$$\frac{\partial \Theta}{\partial t} = \frac{\partial K_\Theta}{\partial z} + \frac{\partial}{\partial z} \left(D_\Theta \frac{\partial \Theta}{\partial z} \right) + F(\Theta), \quad (23)$$

where K_Θ is hydraulic conductivity and D_Θ is the soil water diffusivity, both highly nonlinear functions of the soil water content (Θ), varying by several orders of magnitude from dry to wet soil conditions, and $F(\Theta)$ is the soil water source/sink term representing evapotranspiration, infiltration, and runoff. Infiltration is the ability of a soil of a given wetness to absorb water at a given rate, and runoff is amount of soil moisture that leaves the soil column, at the surface due to precipitation in excess of the infiltration rate, laterally due to soil moisture in excess of saturation for a given layer, and through the bottom due to gravitational drainage, where infiltration and runoff are functions of soil type and soil water content (see Chen et al. 1996, Schaake et al. 1996, and Koren et al. 1999 for further details). Evapotranspiration is discussed further in Sect. 4.4.4, and K_Θ and D_Θ in Sect. 4.6.

4.3.2 Soil Temperature Tendency

Soil heat transfer is treated with a prognostic equation for soil temperature (T) such that

$$C_\Theta \frac{\partial T}{\partial t} = \frac{\partial}{\partial z} \left(\lambda_T \frac{\partial T}{\partial z} \right), \quad (24)$$

where C_Θ is the *volumetric* heat capacity of moist soil and λ_T is the soil thermal conductivity, both functions of the soil water content (Θ). C_Θ is linearly related to Θ , whereas λ_T is a nonlinear function of Θ and increases by several orders of magnitude from dry to wet soil conditions; C_Θ and λ_T are discussed further in Sect. 4.7.

4.3.3 Canopy Water Tendency

The canopy water content (C_w) changes as

$$\frac{\partial C_w}{\partial t} = \sigma_f PD \downarrow - E_c, \quad (25)$$

where σ_f is the plant shading factor ($0 \leq \sigma_f \leq 1$). $PD \downarrow$ is precipitation + dewfall which increases C_w , while canopy water evaporation (E_c) decreases C_w (Precipitation is a forcing quantity provided to the land model from e.g., observations or from a parent atmospheric model.)

4.4 Solution of Surface Energy Budget

In order to determine the surface values of temperature and moisture necessary to calculate surface heat and moisture fluxes, it is necessary to solve the surface energy balance where we begin by evaluating the surface energy balance for the reference

state of the surface that is in a saturated condition in order to determine the potential evaporation. We determine the potential evaporation closely following the derivation in Mahrt and Ek (1984), except that the usual Penman (1948) potential evaporation relationship is modified (as discussed below) since the surface temperature is needed to compute net radiation. This surface energy balance for a saturated surface condition is:

$$(1 - \alpha)S \downarrow + L \downarrow - \epsilon_s \sigma T_s^4 = H + LE_p + G, \quad (26)$$

where potential evaporation (LE_p) and the other terms have been previously defined. Here T_s and H are their values corresponding to the potential evaporation LE_p . (The left hand side of Eq. (26) is simply the net radiation under this saturated condition.) In determining potential evaporation, G is determined using variables from the previous model time step and is updated later. The outgoing longwave radiation, σT_s^4 , is *linearized* as

$$\sigma T_s^4 \approx \sigma T_a^4 \left[1 + 4 \left(\frac{T_s - T_a}{T_a} \right) \right], \quad (27)$$

where T_a is the air temperature at the first model level in the atmosphere. Here the sensible heat flux uses a saturated surface temperature appropriate for the potential evaporation and is defined as

$$H = \rho c_p C_h U (T_s - T_a), \quad (28)$$

where T_s is the saturated surface temperature, and the other terms have been previously defined.

4.4.1 Soil Heat Flux

Soil heat flux (G) is formulated (e.g., described in McCumber and Pielke 1981) as

$$G = -\lambda_T \frac{\partial T_{s_1}}{\partial z}, \quad (29)$$

where λ_T is the soil thermal conductivity and $\partial T_{s_1} / \partial z$ is the soil temperature gradient in the upper soil layer. The finite difference form of Eq. (29) is

$$G = -\lambda_T \frac{T_s - T_{s_1}}{\Delta z}, \quad (30)$$

where T_s and T_{s_1} are the surface and upper soil layer temperatures, respectively, and Δz is the mid-point of the upper soil layer. As with the sensible heat flux (28), for the purpose of calculating potential evaporation, G is determined using values (e.g., actual T_s , etc.) from the previous model time step, but is updated later.

In the presence of a vegetation layer, soil heat flux is reduced because of reduced heat conductivity through vegetation (see Fig. 8). This has been demonstrated by

Viterbo and Beljaars (1995) in the ECMWF model land-surface scheme (TESSEL, van den Hurk et al. 2000). They suggest a simple parameterization to deal with this effect where G is computed as the product of an empirical coefficient (appropriate to the surface concerned) and the temperature difference between the surface and the center of the upper soil layer (3.5 cm in the TESSEL scheme, at that time), i.e.,

$$G = \Lambda_T \Delta T, \quad (31)$$

where Λ_T is a fixed constant thermal conductivity *function* (e.g., $7 \text{ W m}^{-2} \text{ K}^{-1}$ for a grassland site at Cabauw, Netherlands). This formulation draws upon earlier work by van Ulden and Holtslag (1985), and implicitly accounts for the reduction of soil heat flux in the presence of vegetation. Van den Hurk et al. (1995), van den Hurk and Beljaars (1996), and van den Hurk et al. (2000) describe refinements to this approach where the value of Λ_T varies depending on land-surface classification, e.g., bare ground, sparse vegetation, etc. λ_T and alternatives to the soil heat flux formulation in the TESSEL scheme are discussed further in Sect. 4.7.

4.4.2 Linearized Surface Energy Balance

We further define

$$F_n = (1 - \alpha)S \downarrow + L \downarrow - \epsilon_s \sigma T_a^4 - G, \quad (32)$$

and substitute into Eq. (26) to obtain

$$F_n - 4\sigma T_a^4 \left(\frac{T_s - T_a}{T_a} \right) = H + L_v E_p. \quad (33)$$

Substituting Eq. (28) into Eq. (33) we obtain

$$F_n - 4\sigma T_a^4 \left(\frac{T_s - T_a}{T_a} \right) = \rho c_p C_h U [(T_s - T_a) - (\theta_a - T_a)] + L_v E_p. \quad (34)$$

4.4.3 Potential Evaporation

To determine the surface evapotranspiration, we begin by calculating the Penman (1948) potential evaporation that is defined as evaporation from a surface with no “resistance” to evaporation (e.g., free evaporation from an open water surface) and formulated as

$$\begin{aligned} L_v E_p &= \rho c_p C_q U \left(q_{s,sat} - q_a \right) \\ &= \rho c_p C_h U \left[\frac{dq_s}{dT} (T_s - T_a) + \left(q_{a,sat} - q_a \right) \right], \end{aligned} \quad (35)$$

where we make the usual assumption that the exchange coefficients for moisture and heat are equal ($C_q = C_h$). dq_s/dT is the slope of the saturation specific humidity with temperature, $q_{s,sat}$ is the surface saturation specific humidity, and $q_{a,sat}$ and q_a

are the saturation and actual specific humidities at the first atmospheric model level, respectively. To explicitly eliminate T_s in our expression for potential evaporation, we solve for $T_s - T_a$ in Eq. (35), where

$$T_s - T_a = \left[\frac{L_v E_p}{\rho L_v C_h U} - (q_{a,sat} - q_a) \right] \left(\frac{dq_s}{dT} \right)^{-1}. \quad (36)$$

Substituting for $T_s - T_a$ in Eq. (34) using Eq. (36), and after some rearranging, we solve for potential evaporation

$$L_v E_p = \rho c_p C_h U \left(\frac{\Delta \left[\frac{F_n}{\rho c_p C_h U} + (\theta_a - T_a) \right] + A(r+1)}{\Delta + r + 1} \right), \quad (37)$$

where

$$\begin{aligned} \Delta &= \frac{dq_s L_v}{dT c_p}, \\ A &= \frac{L_v}{c_p} (q_{a,sat} - q_a), \\ r &= \frac{4\sigma T_a^4 R_d}{p_s c_p C_h U}. \end{aligned}$$

4.4.4 Surface Evapotranspiration

The total surface moisture flux, or evapotranspiration (E), has contributions from five sources: evaporation of water from an open water source (E_w) (e.g., lakes), plant canopy (E_c), direct evaporation from the soil (E_d), plant transpiration (E_t), and sublimation from the snow or ice surfaces (E_s ; see Sect. 4.8.4 under Cold Season Processes), e.g., Fig. 9, so the total is

$$E = E_w + E_d + E_c + E_t + E_s. \quad (38)$$

The total evaporation cannot exceed the potential evaporation (E_p) defined in Eq. (37). Evaporation from an open water source, E_w , is simply the potential evaporation (E_p). The sum of the evaporative terms is aggregated into a single flux that is then passed to the atmosphere (Fig. 9).

4.4.5 Canopy Evaporation

The canopy evaporation of free water (E_c) is formulated as

$$E_c = \sigma_f \left(\frac{C_w}{S_w} \right)^n E_p, \quad (39)$$

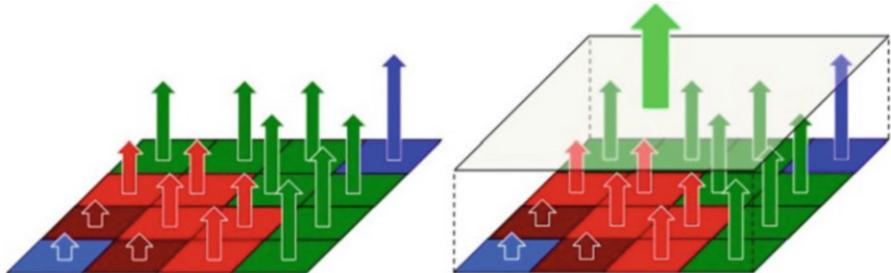


Fig. 9 Example model grid box showing (left) surface fluxes from a number of sources, i.e., open water, canopy water, plant transpiration, direct evaporation, and snow/ice sublimation, and (right) the aggregated flux that is passed to the atmosphere, e.g., a parent atmospheric model

where σ_f is the vegetation fraction (or plant shading factor, a fraction between 0 and 1), C_w and S_w are the actual and saturated water contents, respectively, for a canopy surface (a function of plant type), and $n = 0.5$, following Pan and Mahrt (1987) who cite earlier studies. The canopy water is filled by precipitation or dewfall, and when saturated, all additional water is assumed to fall through to the ground surface.

4.4.6 Direct Soil Evaporation

To determine direct evaporation (E_d) at the air-soil interface, it is necessary to determine the rate at which the soil can provide moisture to the surface to evaporate. We assume that evaporation may proceed at the potential rate until the surface soil moisture content decreases to an “air-dry” value, Θ_d (see Mahrt and Pan 1984, Chang et al. 1999, and references therein). As a first step, we demand that the evaporation be at the potential rate in which case

$$E_p = \left[D_\Theta \left(\frac{\partial \Theta}{\partial z} \right) + K_\Theta \right] (1 - \sigma_f), \quad (40)$$

where D_Θ is the soil water diffusivity and K_Θ is the soil hydraulic conductivity (D_Θ and K_Θ will be discussed further below). The finite-difference form of Eq. (40) is

$$E_p = \left[D_\Theta \left(\frac{\Theta_s - \Theta_1}{\Delta z / 2} \right) + K_\Theta \right] (1 - \sigma_f), \quad (41)$$

where D_Θ and K_Θ are the values averaged between the surface and upper soil layer, Θ_s and Θ_1 are the volumetric soil moisture contents at the surface and upper soil model layer, respectively, and $\Delta z / 2$ is the mid-point of the upper soil layer. The direct soil evaporation can proceed at a potential rate when the apparent soil moisture at the surface (obtained by solving for Θ_s in Eq. (41)) is greater than the air-dry value (Θ_d),

that is, when the soil is sufficiently wet (*demand control stage*). When the soil dries out, the evaporation can only proceed at the rate by which the soil can diffuse water upward from below (*flux control stage*) in which case $\Theta_s = \Theta_d$ and $E_d < E_p$. Then the direct soil evaporation (in finite difference form) is

$$E_d = \left[D_\Theta \left(\frac{\Theta_d - \Theta_1}{\Delta z / 2} \right) + K_\Theta \right] (1 - \sigma_f). \quad (42)$$

4.4.7 Plant Transpiration and Canopy Resistance

Plant transpiration (E_t) is calculated as

$$E_t = \sigma_f k_v \left[1 - \left(\frac{C_w}{S_w} \right)^n \right] E_p, \quad (43)$$

where k_v is the “plant coefficient” (a fraction between 0 and 1) and can be related to the commonly used expression of “canopy resistance,” r_c (sometimes called “surface resistance” if the surface is not fully covered with vegetation). The canopy resistance (r_c) accounts for the reduction in transpiration due to plant stomatal control and has been often expressed in the meteorological land-surface modeling community as a function of environmental variables, most commonly: incoming solar radiation, air temperature, specific humidity deficit of the air, and soil moisture availability. The plant coefficient (k_v) may be related to r_c by equating the expression for transpiration used in the Noah LSM (43) with the usual Penman-Monteith expression for transpiration (Monteith 1965). The following relation is then obtained for k_v

$$k_v = \frac{(r + 1 + \Delta + \delta_\theta)}{(r + 1 + \delta_\theta)(1 + r_c C_h U) + \Delta}, \quad (44)$$

where terms have been defined above. The canopy resistance itself may follow the Jarvis-Stewart “big leaf” approach (Jarvis 1976; Stewart 1988), where (r_c) is a function of a number of empirical coefficient based on environmental conditions (atmospheric and soil); r_c is then given as

$$r_c = r_{cmin} (r_{cs} r_{cT} r_{cq} r_{csoil})^{-1}, \quad (45)$$

where r_{cmin} is the minimum canopy resistance, and r_{cs} , r_{cT} , r_{cq} , and r_{csoil} are the irradiance, temperature, specific humidity deficit, and soil moisture availability factors, respectively, all affecting the canopy resistance, where all terms here are a function of vegetation type and time of year. (An additional factor is the soil temperature, i.e., r_{cST} , that could affect the canopy resistance, especially in the spring with “vegetation green-up” and seed germination when soil temperatures are increasing.) The description here closely follows the canopy resistance formulation described in Noilhan and Planton (1989), i.e.,

$$r_{cs} = \frac{a_{s1}S \downarrow a_{s2}LAI + \frac{r_{smin}}{r_{smax}}}{a_{s3} + a_{s1}S \downarrow a_{s2}LAI}, \quad (46)$$

where LAI is the leaf area index, a_{s1} , a_{s2} , and a_{s3} are coefficients, and r_{smin} is the minimum stomatal resistance ($r_{smin} = r_{cmin} LAI$), and $S\downarrow$ is the incoming solar radiation.

$$r_{cT} = 1 - a_{T1} (T_{cref} - T_a)^2, \quad (47)$$

where a_{T1} is a coefficient, T_{cref} is a reference temperature, and T_a is the air temperature at the first model level in the atmosphere.

$$r_{cq} = 1 - a_{q1} (q_{a,sat} - q_a), \quad (48)$$

where a_{q1} is a coefficient, and $q_{a,sat}$ and q_a are the saturation and actual specific humidities, respectively, at the first model level in the atmosphere.

$$r_{csoil}(\Theta_i) = \begin{cases} 0, & \Theta_i \leq \Theta_{wilt} \\ \frac{\Theta_i - \Theta_{wilt}}{\Theta_{fc} - \Theta_{wilt}}, & \Theta_{wilt} < \Theta_i \leq \Theta_{fc} \\ 1, & \Theta_{fc} < \Theta_i \end{cases}, \quad (49)$$

where $r_{csoil}(\Theta_i)$ is for a given soil layer Θ_i . Θ_{fc} is the *field capacity*, the volumetric soil moisture content above which plants are no longer water stressed, while Θ_{wilt} is the *permanent wilting point*, the volumetric soil moisture content at which transpiration ceases. The total r_{csoil} is then

$$r_{csoil} = \sum_{i=1}^n r_{csoil}(\Theta_i) g_i \frac{\Delta z_i}{\Delta z}, \quad (50)$$

where n is the number of soil layers, g_i is the root density function for the i th soil layer, and Δz_i and Δz are the thicknesses of the i th soil layer and total soil column, respectively. (g_i is nominally set to unity for each soil layer in the Noah LSM, that is, an equal root density with depth. However, observations suggest that the root density varies with depth, perhaps higher nearer the surface or in a soil layer with episodically higher soil moisture content.)

Alternates to “Jarvis-Stewart,” e.g., following Niu et al. (2011) and Yang et al. (2011) in the “Noah-MP” use the more physically based Ball-Berry CO₂-based photosynthesis approach to describe canopy conductance and how this depends on environmental factors. See Niu et al. (2011) and Yang et al. (2011) for further details.

4.5 Surface Temperature

To determine surface temperature (T_s) we start with the surface energy balance similar to Eq. (26) except now we use the actual evaporation

E calculated from Eq. (38) instead of the potential evaporation E_p . Note that actual evaporation can be expressed as $E = \beta E_p$ where β is a factor multiplied by the potential evaporation to get the actual evaporation; β absorbs all influences that reduce the potential evaporation to the actual. The surface energy balance then becomes

$$(1 - \alpha)S \downarrow + L \downarrow - \epsilon_s \sigma T_s^4 = H + \beta L_v E_p + G. \quad (51)$$

Using Eqs. (27) and (28), we can rewrite this surface energy balance as

$$\begin{aligned} F - 4\sigma T_a^4 - 4\sigma T_a^4 \left(\frac{T_s - T_a}{T_a} \right) &= \rho c_p C_h U [(\theta_s - T_a) - (\theta_a - T_a)] + \beta L_v E_p \\ &\quad + G, \end{aligned} \quad (52)$$

where $F = (1 - \alpha)S \downarrow + L \downarrow$. Using the definition of the soil heat flux (G) from Eq. (30), and r from Eq. (38), we can solve for T_s as

$$T_s = \frac{\Delta z \rho c_p C_h U [T_a(r+1) + (\Theta_a - T_a)] + \Delta z (F - \sigma T_a^4 - \beta L_v E_p) + \lambda_T T_{s1}}{\Delta z \rho c_p C_h U (r+1) + \lambda_T}. \quad (53)$$

After updating the soil moisture content, and soil and surface temperatures, an updated soil heat flux (G) can be found by re-evaluating Eq. (29). Similarly, the sensible heat flux (H) is updated using Eq. (6).

4.6 Soil Hydraulics

4.6.1 Clapp and Hornberger

Hydraulic conductivity (K_Θ) and soil water diffusivity (D_Θ) used in Eq. (23) are nonlinear functions of soil moisture (Θ) and change by several orders of magnitude from dry to wet soil conditions (see Ek and Cuenca 1994). They follow Clapp and Hornberger (1978) (and Cosby et al. 1984) and are defined as

$$K_\Theta = K_{\Theta_s} \left(\frac{\Theta}{\Theta_s} \right)^{2b+3}, \quad (54)$$

$$D_\Theta = \left(\frac{b K_{\Theta_s} \psi_s}{\Theta_s} \right) \left(\frac{\Theta}{\Theta_s} \right)^{b+2}, \quad (55)$$

where K_{Θ_s} is the saturation hydraulic conductivity, Θ_s is the saturation volumetric soil moisture content, b is an empirically derived coefficient, and ψ_s is the saturation soil moisture potential (all a function of soil type), where the actual soil moisture potential, ψ , is defined as

$$\psi = \psi_s \left(\frac{\Theta}{\Theta_s} \right)^{-b}, \quad (56)$$

where Eq. (56) is also from Clapp and Hornberger (1978).

4.6.2 van Genuchten

An alternate to Clapp and Hornberger is the approach by van Genuchten (1980) where

$$K_\Theta = K_{\Theta_s} S_e^l \left[1 - \left(1 - S_e^{1/m} \right)^m \right]^2, \quad (57)$$

$$D_\Theta = K_\Theta (\partial \Theta / \partial \psi), \quad (58)$$

where l and m are fitting parameters (functions of soil type and soil density), and S_e is the effective soil moisture saturation fraction defined as

$$S_e = (\Theta - \Theta_r) / (\Theta_s - \Theta_r), \quad (59)$$

where the Θ_r is the residual volumetric soil moisture content and the other terms have been defined above. The soil moisture potential is defined as

$$\psi = \frac{1}{\alpha_s} \left[S_e^{-1/m} - 1 \right]^{1/n}, \quad (60)$$

where α_s and n are also fitting parameters, and $m = 1 - 1/n$. See Cuenca et al. (1996) and Beljaars and Bosveld (1997) for further information on the van Genuchten formulation.

4.7 Soil Thermodynamics

The thermal conductivity (λ_T) used in Eq. (24) is a nonlinear function of the soil moisture content (Θ), changing by a few orders of magnitude from dry to wet soil conditions, and in the absence of vegetation is the “bare soil” thermal conductivity λ_{T0} . Following Al Nakshabandi and Kohnke (1965), λ_{T0} is expressed as

$$\lambda_{T0} = \begin{cases} 420 \exp([- \log_{10}(100|\psi|)] + 2.7), & \log_{10}(100|\psi|) \leq 5.1 \\ 0.1722, & \log_{10}(100|\psi|) > 5.1 \end{cases}, \quad (61)$$

where ψ is soil moisture potential.

An alternative to Al Nakshabandi and Kohnke is the formulation by Johansen (1975) described in Peters-Lidard et al. (1998), where λ_{T0} is a less nonlinear function of soil moisture content and yields more (*less*) thermal conductivity for drier (*moister*) soils. As noted in Marshall et al. (2003) and Ek et al. (2003), this then yields greater (*lesser*) soil heat flux, that in turn leads to a more damped (*amplified*)

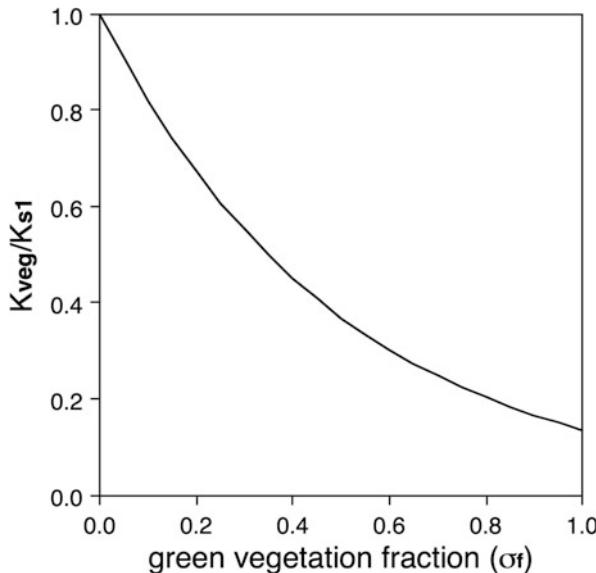


Fig. 10 Ratio of soil thermal conductivity under vegetation to bare soil thermal conductivity (K_{veg}/K_{s1}) as a function of green vegetation fraction

diurnal signal in the surface skin and near-surface (e.g., 2-m) air temperatures, and was found to improve the land-surface response in mesoscale model performance.

Soil heat flux is reduced in the presence of a vegetation canopy because of reduced heat conductivity through vegetation and is included implicitly in the soil heat flux formulation in the ECMWF TESSEL land-surface scheme (van den Hurk et al. 2000). An alternative is described in Peters-Lidard et al. (1997) where the effect of vegetation is explicitly included, so that soil thermal conductivity is reduced by an exponential function of vegetation as

$$\lambda_T = \lambda_{T0} e^{-\beta LAI}, \quad (62)$$

where LAI is the leaf area index, and β is an empirical coefficient equal to 0.5. Alternatively, the vegetation fraction ($0 \leq \sigma_f \leq 1$) may be used instead of LAI , where

$$\lambda_T = \lambda_{T0} e^{-\beta' \sigma_f}, \quad (63)$$

and β' is an empirical coefficient, nominally equal to 2.0 (Ek et al. 2003). See Fig. 10.

The volumetric heat capacity of moist soil (C_Θ) used in Eq. (24) includes contributions from the mineral soil itself, as well as from air, water, and ice in the soil and is linearly related to soil moisture (Θ) as

$$C_\Theta = (1 - \Theta_{sat})c_{soil} + (\Theta_{sat} - \Theta)c_a + (\Theta - \Theta_{ice})c_w + \Theta_{ice}c_{ice}, \quad (64)$$

where c_{soil} is the soil heat capacity (a function of soil type, but chosen as $1.26 \times 10^6 \text{ J m}^{-3} \text{ K}^{-1}$), c_a is the heat capacity of air in the soil ($1250 \text{ J m}^{-3} \text{ K}^{-1}$, which assumes an air density of $\approx 1.24 \text{ kg m}^{-3}$), c_w is the heat capacity of water in the soil ($4.2 \times 10^6 \text{ J m}^{-3} \text{ K}^{-1}$), and c_i is the heat capacity of ice in the soil ($2.1 \times 10^6 \text{ J m}^{-3} \text{ K}^{-1}$), and Θ_{ice} is the frozen soil moisture (discussed in Cold Season Processes below).

4.8 Cold Season Processes

Cold season processes are important in the evolution of the land-surface for a large portion of the earth during many cold season months. In the presence of snow cover, albedo increases, surface roughness is often reduced, and the exchange of heat and moisture between land-surface and atmosphere is diminished, while subsurface freezing reduces the movement of heat and moisture within the soil. All of these processes affect the surface energy budget and thus the surface fluxes (and snow melting), so it is necessary to include these effects in LSMs used in weather and climate models, i.e., in the Noah and other LSMs (e.g., Viterbo et al. 1999; Smirnova et al. 2000; Boone et al. 2000; Boone and Etchevers 2001). These processes are described further below. Refer also to Koren et al. (1999) and Ek et al. (2003).

4.8.1 Snowpack Evolution

Snowpack accumulates due to falling snow, with the depth determined by the precipitation amount (the snow water equivalent), and the density of the snow which uses a parameterization based on air temperature. Snow density may be as high as 10:1 (snow depth:snow water equivalent) or higher for newly fallen cold, dry snow, or may have a much smaller ratio approaching 2.5:1 in the Noah LSM after compaction, important in the seasonal snowpack evolution, where the compaction is determined by snow temperature and snow age. The Noah LSM has a single-layer bulk snowpack, where newly-fallen snow is added to the snowpack, increasing depth, and then the density is “homogenized” into single value for the snowpack. On the otherhand, the Noah-MP has a multilayer (up to 3) snowpack (See Niu et al. 2011; Yang et al. 2011) and carries separate temperatures and densities in each snow layer.

4.8.2 Fractional Snow Cover

A fractional snow cover treatment allows for patchy snow cover if the snow depth is below some threshold, and hence allows exposed ground, a lower albedo, more energy absorption, and the aggregate (e.g., model gridbox) surface skin temperature may rise above freezing. As such the surface sensible heat flux may be greater than a completely snow-covered surface, with a corresponding increase in low-level air temperature. The subgrid patchiness is related to the depth of the snow and surface characteristics, e.g., for a “smoother” surface such as grassland, a smaller snow depth threshold is required for 100% snow cover compared to a forest (Fig. 11).

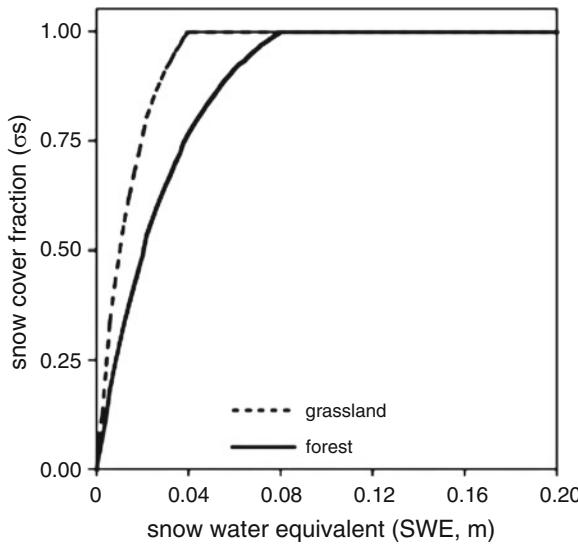


Fig. 11 Snow cover fraction as a function of snow water equivalent (SWE) for forest (thick solid line) and grassland (thick dashed line) vegetation classes

4.8.3 Albedo over Snow

In the presence of snow cover, the surface albedo may be markedly increased due to the high albedo of snow (depending on vegetation cover). Even over deep snow, however, the albedo can vary greatly depending on the surface characteristics. For example, a conifer forest may have a lower albedo due to darker treetops sticking through a brighter (deep) snowpack, compared with a higher albedo for completely snow-covered grassland. In conditions of shallow snowpack when snow first accumulates at the start of snowfall or diminishes due to snow sublimation or snow melt, there will be patchy snow-covered areas, e.g., in a model gridbox. To account for this patchiness effect, the surface albedo is formulated as a composite of a snow-covered and non-snow-covered surface as

$$\alpha = \alpha_0 + (1 - \sigma_f)\sigma_s(\alpha_s - \alpha_0) \quad (65)$$

where α , α_0 , and α_s are the actual, snow-free, and maximum snow surface albedo (from Sect. 3), respectively, σ_f is the green vegetation fraction ($0 \leq \sigma_f \leq 1$), and σ_s is the snow cover fraction, as illustrated in Fig. 12. As snow depth becomes zero, the albedo becomes the snow-free albedo ($\alpha = \alpha_0$). When the snow depth exceeds a threshold value (dependent on land-use class, e.g., vegetation type), snow cover is 100% ($\sigma_s = 1$) and $\alpha = \alpha_s$, the maximum snow albedo.

4.8.4 Snow Sublimation

Snow sublimation is calculated from the potential evaporation, except that the latent heat of sublimation rather than latent heat of vaporization is used in calculating the

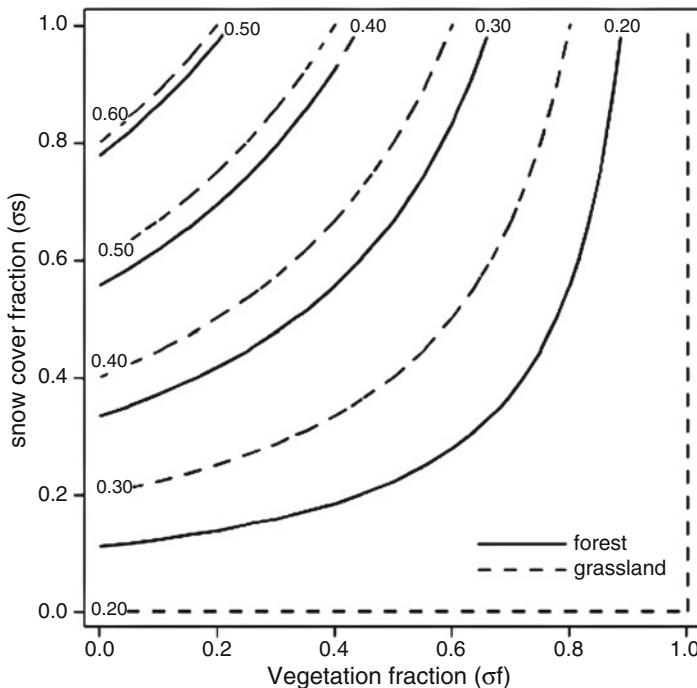


Fig. 12 Surface albedo contours as a function of snow cover fraction versus green vegetation fraction with “typical” forest (grassland) values for snow-free albedo, $\alpha_0 = 0.15$ ($\alpha_0 = 0.20$) and maximum snow albedo, $\alpha_s = 0.60$ ($\alpha_s = 0.70$)

surface moisture flux in energy terms. For patchy snow cover, the snow sublimation is simply proportionally weighted by the snow cover fraction (snow sublimation, E_s) and the non-snow-covered fraction (a sum of the other evaporative terms, $E_w + E_d + E_c + E_t$), similarly proportionally weighted, i.e.,

$$E = \sigma_s E_s + (1 - \sigma_s)(E_w + E_d + E_c + E_t). \quad (66)$$

The depth of the snowpack is then decreased by the loss of snow mass (SWE) corresponding to E_s .

4.8.5 Snow Melt

The solution of the surface energy budget yields a surface temperature, and when that temperature remains below freezing, if any snowpack is present, no melting takes place. When that temperature is above freezing, then the surface energy budget is reevaluated, but holding the surface temperature at freezing, with the residual energy then used to melt snow, where the depth of the snowpack is then decreased by the loss of snow mass (SWE) corresponding to the snow melt. In the case of a fractional snow cover, a portion of this residual energy is used to melt snow, with the

remaining portion heating the non-snow-covered fraction of the surface, hence yielding an aggregate (e.g., model gridbox) surface skin temperature above freezing. In the Noah LSM, any melted snow is immediately added to the soil water, while the Noah-MP can carry melted snow liquid water in the snowpack, including the re-freezing of this liquid water (e.g., at night when temperatures drop).

4.8.6 Soil Heat Flux Under Snow

As the snowpack becomes very thin, it is difficult to estimate the large near-surface temperature gradients in the snow and upper soil layer. As such, the soil heat flux formulation in the Noah LSM includes the effect of heat flow through thin patchy snow cover by considering the thermal conductivity of a snowpack-plus-upper-soil layer following a method described in Lunardini (1981), where heat flow can be in parallel, in series, or intermediate between the two. Here parallel heat flow through the snowpack-plus-upper-soil-layer is assumed which yields a larger thermal conductivity (than say, series), implicitly accounting for the nonuniform nature of snowpack cover. The effective thermal conductivity for the surface is then determined via a linear weighting between the snow-covered and non-snow-covered fractions (e.g., of a model gridbox), where

$$\begin{aligned} K_T &= \Delta Z_s K_s + \Delta Z_{s1} K_{s1}, \\ K_{\text{eff}} &= \sigma_s K_T + (1 - \sigma_s) K_{s1}, \end{aligned} \quad (67)$$

where K_{s1} , K_T , K_{eff} are the thermal conductivities of the upper soil layer, snow-plus-upper-soil-layer, and patchy snow-covered surface (Fig. 13), respectively, ΔZ_{s1} is the upper soil layer depth, and σ_s is the snow cover fraction ($0 \leq \sigma_s \leq 1$). The soil heat flux through the patchy snow-covered surface is then formulated as

$$G = \frac{K_{\text{eff}}(T_s - T_{s1})}{\Delta Z_s + \Delta Z_{s1}} \quad (68)$$

In this formulation, the thermal conductivity remains robustly defined even in the extremes of vanishing snow cover ($\Delta Z_s = 0$, $\sigma_s = 0$, $K_{\text{eff}} = K_{s1}$), or for a very deep snowpack ($\Delta Z_s \gg \Delta Z_{s1}$, $\sigma_s = 1$, $K_{\text{eff}} \rightarrow K_s$), which is quite important for numerical stability. Patchy snow cover must be accounted for since it increases the heat flux between the surface and atmosphere (especially at smaller snow cover fractions) because of the typically larger thermal conductivity of soil compared to snow.

4.8.7 Frozen Soil Thermodynamics and Hydraulics

In freezing conditions, a portion of the soil moisture may undergo freezing, that is,

$$\Theta = \Theta_{\text{liq}} + \Theta_{\text{ice}} \quad (69)$$

where Θ is the total soil moisture, Θ_{liq} is the liquid soil moisture, and Θ_{ice} is the soil ice (frozen soil moisture, a function of both soil temperature and soil moisture), and soil physics must accommodate this condition. Soil moisture can be

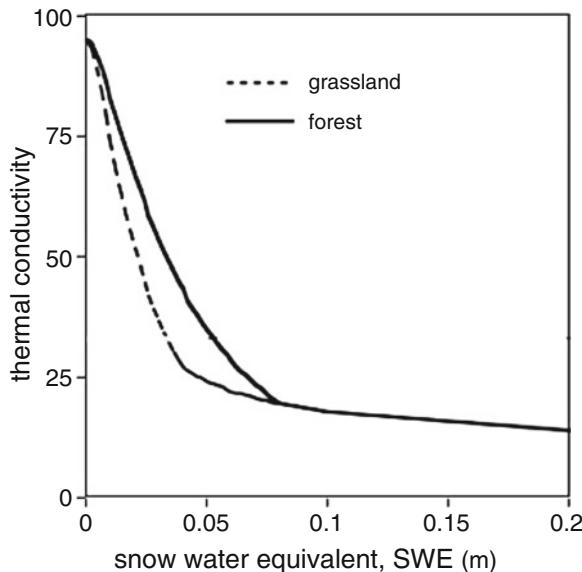


Fig. 13 Thermal conductivity (K_{eff}) through patchy snow cover versus snow water equivalent (SWE) for forest (thick solid line) and grassland (thick dashed line) vegetation classes, with the same patchiness corresponding to Fig. 11

super-cooled such that not all water is frozen even if the soil layer is below freezing, so that the flow of *liquid* water portion in the soil follows the soil moisture tendency equation (23). Also, to account for the phase change of soil moisture, an additional source/sink term, $\rho_w L_s \partial \Theta_{ice} / \partial t$, is added to the soil temperature tendency equation (24), where ρ_w is water density, and L_s is latent heat of fusion. Additionally, the thermal conductivity for frozen soil is adjusted, depending on soil type and soil moisture content; similarly there is a reduction in the infiltration of water (from unfrozen precipitation or snowmelt) into frozen soils. The impact of freezing soils provides a thermal inertia at the freezing point due to freezing and thawing of soils, reducing the amplitude of diurnal and seasonal temperature cycle in the soil.

5 Land-Atmosphere Interaction

Land-atmosphere coupling involves the interactions between the land-surface and the atmospheric boundary layer (ABL), and in turn with the free atmosphere above. The role of soil moisture in the evolution of surface fluxes and atmospheric boundary layer (ABL) development, including ABL clouds (i.e., fair-weather cumulus) involves a complex interaction of surface and atmospheric processes (see Fig. 14). We examine *local* land-atmosphere interaction or coupling in a “two-legged” approach (see Dirmeyer et al. 2018; Santanello et al. 2017) by inspecting the soil moisture-surface flux (“terrestrial leg”) and surface flux-ABL (“atmospheric leg”) relationships.

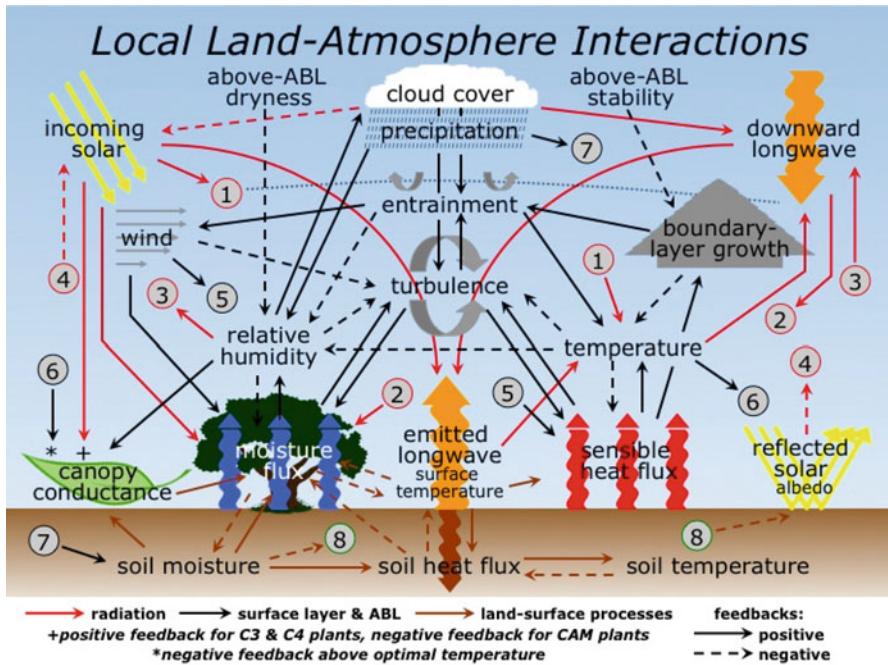


Fig. 14 Important interactions between the land-surface and atmospheric boundary layer for conditions of daytime surface heating. Solid arrows indicate the direction of feedbacks that are normally positive (leading to an increase of the recipient variable). Dashed arrows indicate negative feedbacks. Two consecutive negative feedbacks make a positive feedback

5.1 Near-Surface Land-Atmosphere Interaction (NSLAI)

The *terrestrial leg* of land-atmospheric interaction determines the coupling between soil moisture and the evolution of surface fluxes and is posed in terms of soil moisture changes and the effect on the surface evaporative fraction. Evaporative fraction is the fraction of available energy at the surface that goes into surface moisture flux (versus the energy that goes into surface sensible heat flux, soil or ground heat flux, and outgoing longwave radiation), and is a combination of plant transpiration, bare soil (direct) evaporation, and evaporation of canopy-intercepted water. The change in evaporative fraction with changing soil moisture is an indicator of the strength of coupling between the surface and the atmosphere and depends on a number of different processes. These include to what degree the surface is vegetated, how plant transpiration and soil hydraulic and thermal processes change with changing soil moisture, as well as the low-level surface-layer turbulence. For strong (*weak*) near-surface land-atmosphere coupling, a given change in soil moisture yields a large (*small*) change in evaporative fraction.

5.1.1 NSAI: Potential Evaporation

We begin with the potential evaporation (LE_p) since it has a role in calculating the total surface moisture flux. Theoretically, LE_p is the evaporation from a well-watered or wet surface with unstressed vegetation (Penman 1948) (and has also been applied to evaporation from open water and bare soil). The corresponding evaporative fraction for potential evaporation (ef_p) is

$$\begin{aligned} LE_p &= \frac{s(R_n - G) + \rho c_p g_a \delta e}{s + \gamma}, \\ ef_p &= \frac{s + \frac{\rho c_p g_a \delta e}{R_n - G}}{s + \gamma}, \end{aligned} \quad (70)$$

where s is the slope of the saturation vapor pressure (with temperature), R_n is net radiation, G is soil heat flux, ρ is air density, c_p is specific heat of air, g_a is aerodynamic conductance (a measure of atmospheric turbulence), δe is the atmospheric vapor pressure deficit (a measure of atmospheric humidity), and γ is the psychometric “constant.” s and γ are

$$\begin{aligned} s &= \frac{de_s}{dT} = \frac{L_v}{R_v} \frac{e_s}{T^2}, \\ \gamma &= \frac{c_p p}{\epsilon L_v}, \end{aligned} \quad (71)$$

where L_v is latent heat, R_v is the gas constant for water vapor, e_s is saturation vapor pressure, T is air temperature, p is surface air pressure, and ϵ is the ratio of the molecular weight of water vapor to dry air (0.622). Soil heat flux is

$$G = \frac{\lambda_T (T_{sfc} - T_{ns})}{\delta z} = \frac{\lambda_T \delta T_{ns}}{\delta z}, \quad (72)$$

where λ_T is soil thermal conductivity, T_{sfc} and T_{ns} are the surface skin and near-surface soil temperatures, respectively (δT_{ns} is the near-surface soil temperature gradient), and δz is the nominal thickness of the near-surface soil layer. Soil moisture (matric) potential (ψ , following Clapp and Hornberger 1978 and Cosby et al. 1984) and soil thermal conductivity (λ_T , following Al Nakshabandi and Kohnke 1965) are, respectively

$$\begin{aligned} \psi &= \psi_{sat} \left(\frac{\Theta_{ns}}{\Theta_{sat}} \right)^{-\beta}, \\ \lambda_T &= a \exp[-\log_{10}(c\psi) + d] = a \exp[-b \ln(c\psi) + d], \end{aligned} \quad (73)$$

where ψ_{sat} is soil moisture potential at saturation, Θ_{ns} and Θ_{sat} are the near-surface and saturation (porosity) soil moisture values, respectively, and β is a coefficient, and $a = 420$, $b = \log(e)$, $c = 100$, and $d = 2.7$; ψ_{sat} , Θ_{sat} , and β are functions of soil type. Alternate functions for soil thermal conductivity may be used, e.g., Johansen (1975) as discussed in Peters-Lidard et al. (1998).

The changes in ψ , λ_T , and G with changing soil moisture are

$$\begin{aligned}\frac{\partial\psi}{\partial\Theta} &= -\frac{\beta\psi}{\Theta_{ns}}, \\ \frac{\partial\lambda_T}{\partial\Theta} &= \frac{b\beta\lambda_T}{\Theta_{ns}}, \\ \frac{\partial G}{\partial\Theta} &= \frac{b\beta G}{\Theta_{ns}}.\end{aligned}\tag{74}$$

The change in *potential* evaporative fraction with changing soil moisture is then

$$\frac{\partial \ln ef_p}{\partial \Theta} = \frac{1}{\Theta_{ns}} \left[\frac{s(R_n - G)}{\rho c_p g_a \delta e} + 1 \right]^{-1} \frac{b\beta G}{R_n - G}.\tag{75}$$

Potential Evaporation Case, Strong Land-Atmosphere Coupling. Conditions that lead to a large change in ef_p with changing Θ_{ns} (i.e., strong land-atmosphere coupling) include strong surface-layer turbulence ($g_a \gg 0$), very dry air ($\delta e \gg 0$) and dry soil (Θ_{ns} small), and a large soil heat flux to available energy ratio ($G/(R_n - G)$) which yield

$$\frac{\partial \ln ef_p}{\partial \Theta} \rightarrow \frac{b\beta}{\Theta_{ns}} \frac{G}{(R_n - G)}.\tag{76}$$

Potential Evaporation Case, Weak Land-Atmosphere Coupling. Conversely, conditions that lead to a small change in ef_p with changing Θ_{ns} (i.e., weak land-atmosphere coupling) include weak surface-layer turbulence ($g_a \rightarrow 0$), very humid air ($\delta e \rightarrow 0$) and moist soil (Θ_{ns} large), and a small soil heat flux to available energy ratio ($G/(R_n - G)$) which yield

$$\frac{\partial \ln ef_p}{\partial \Theta} \rightarrow 0.\tag{77}$$

5.1.2 NSLAI: Transpiration

Transpiration by vegetation (LE_t), using the “Penman-Monteith” approach (Monteith 1965) and the evaporative fraction for transpiration (ef_t) are

$$\begin{aligned}LE_t &= \frac{s(R_n - G) + \rho c_p g_a \delta e}{s + \gamma \left(1 + \frac{g_a}{g_c} \right)}, \\ ef_t &= \frac{s + \frac{\rho c_p g_a \delta e}{R_n - G}}{s + \gamma \left(1 + \frac{g_a}{g_c} \right)},\end{aligned}\tag{78}$$

where g_c is canopy conductance. (Note that as $g_c \rightarrow \infty$, $LE_t \rightarrow LE_p$ and $ef_t \rightarrow ef_p$.) Following Jarvis (1976, and others), canopy conductance can be written as

$$g_c = g_{s_{max}} LAI g_{s\downarrow} g_T g_{\delta e} g_\Theta, \quad (79)$$

where $g_{s_{max}}$ is maximum stomatal conductance, LAI is leaf area index (vegetation density), and $g_{s\downarrow}$, g_T , $g_{\delta e}$ and g_Θ are transpiration factors accounting for the effects of incoming solar radiation, air temperature, atmospheric humidity deficit and soil moisture availability, respectively, all functions of vegetation type and environmental conditions. (Note that $g_c = 1/r_c$, as described in Eq. (45).) Soil moisture availability is defined as

$$\begin{aligned} g_\Theta &= \frac{\Theta_{rz} - \Theta_{wilt}}{\Theta_{ref} - \Theta_{wilt}}, \\ &= \frac{\delta\Theta_{rz}}{\Theta_{ref} - \Theta_{wilt}}, \end{aligned} \quad (80)$$

where Θ_{rz} is root zone soil moisture, Θ_{wilt} is soil moisture wilting point below which transpiration ceases, and Θ_{ref} is the soil moisture reference value above which transpiration is not soil moisture limited ($\delta\Theta_{rz}$ is root zone volumetric soil moisture availability). The change in g_c with changing soil moisture is

$$\frac{\partial g_c}{\partial \Theta} = \frac{g_c}{\delta\Theta_{rz}}. \quad (81)$$

Using Eqs. (78) and (81), the change in *transpiration* fraction with changing soil moisture is then

$$\frac{\partial \ln ef_t}{\partial \Theta} = \frac{1}{\delta\Theta_{rz}} \left\{ \left[\left(\frac{s + \gamma}{\gamma} \right) \frac{g_c}{g_a} + 1 \right]^{-1} + \left[\frac{s(R_n - G)}{\rho c_p g_a \delta e} + 1 \right]^{-1} \frac{\delta\Theta_{rz}}{\Theta_{ns}} \frac{b\beta G}{(R_n - G)} \right\}. \quad (82)$$

Strictly speaking, Eq. (82) applies to the change in evaporative fraction with the change in *root zone* soil moisture, while the second term on the right hand side of Eq. (82) is with respect to *near-surface* soil moisture. But here we assume that $\Theta_{rz} \approx \Theta_{ns}$ so that Eq. (82) is still generally valid.

The relationship in Eq. (82) is described in Jacobs et al. (2008) (although without the second term on the right hand side) which follows Jarvis and McNaughton (1986) who define a “decoupling” parameter (Ω) as

$$\Omega = \left[\left(\frac{\gamma}{s + \gamma} \right) \frac{g_a}{g_c} + 1 \right]^{-1}, \quad (83)$$

where $\Omega \rightarrow 0$ ($\Omega \rightarrow 1$) indicates strong (*weak*) land-atmosphere coupling. Instead, a “coupling” parameter, ω ($= 1 - \Omega$) is defined as

$$\omega = \left[\left(\frac{s + \gamma}{\gamma} \right) \frac{g_c}{g_a} + 1 \right]^{-1}, \quad (84)$$

where $0 \leq \omega \leq 1$, and $\omega \rightarrow 1$ ($\omega \rightarrow 0$) indicates strong (*weak*) land-atmosphere coupling. Further, the second term on the right hand side of Eq. (82) is an additional coupling parameter defined as

$$\omega_G = \left[\frac{s(R_n - G)}{\rho c_p g_a \delta e} + 1 \right]^{-1} \frac{\delta \Theta_{rz}}{\Theta_{ns}} \frac{b\beta G}{(R_n - G)}, \quad (85)$$

where $0 \leq \omega_G < \approx O(1)$, and $\omega_G \gg 0$ ($\omega_G \rightarrow 0$) indicates strong (*weak*) land-atmosphere coupling. ω_G is typically much smaller than ω and is included in the coupling parameter to account for “communication” between the soil and surface through the soil heat flux (G) and also depends on atmospheric turbulence (g_a), humidity (δe), and the available energy ($R_n - G$).

Using Eqs. (84) and (85), Eq. (82) may then be expressed simply as

$$\frac{\partial \ln e f_t}{\partial \Theta} = \frac{\omega + \omega_G}{\delta \Theta_{rz}}. \quad (86)$$

Vegetated Surface, Strong Land-Atmosphere Coupling. For very strong surface-layer turbulence ($g_a \gg 0$), and very strong stomatal control ($g_c \rightarrow 0$) (due to vegetation with strong stomatal control (small $g_{s_{max}}$), low vegetation density (small LAI), nonoptimal solar insolation ($g_{s\downarrow} \rightarrow 0$), nonoptimal air temperature ($g_T \rightarrow 0$), very dry air ($\delta e \gg 0$ and $g_{\delta e} \rightarrow 0$), dry soil (small $\delta \Theta_{rz}$, $g_\Theta \rightarrow 0$, and small Θ_{ns})), i.e., $\omega \rightarrow 1$, and for a large soil heat flux to available energy ratio ($G/(R_n - G)$), then

$$\frac{\partial \ln e f_t}{\partial \Theta} \rightarrow \frac{1}{\delta \Theta_{rz}} + \frac{b\beta G}{\Theta_{ns}(R_n - G)}. \quad (87)$$

This is the case of strong land-atmosphere coupling, so for a given change in soil moisture, there is a large change in transpiration, e.g., an evergreen forest in dry conditions.

Vegetated Surface, Weak Land-Atmosphere Coupling. On the other hand, for very weak surface-layer turbulence ($g_a \rightarrow 0$), and very weak stomatal control ($g_c \gg 0$) (due to vegetation with weak stomatal control (large $g_{s_{max}}$), high vegetation density (large LAI), optimal solar insolation ($g_{s\downarrow} \rightarrow 1$), optimal air temperature ($g_T \rightarrow 1$), very humid air ($\delta e \rightarrow 0$ and $g_{\delta e} \rightarrow 1$), moist soil (large $\delta \Theta_{rz}$, $g_\Theta \rightarrow 1$, and large Θ_{ns})), i.e., $\omega \rightarrow 0$, and for a small soil heat flux to available energy ratio ($G/(R_n - G)$), then

$$\frac{\partial \ln e f_t}{\partial \Theta} \rightarrow 0. \quad (88)$$

This is the case of weak land-atmosphere coupling, so for a given change in soil moisture, there is little change in transpiration, e.g., a short crop canopy or grassland in wet conditions.

5.1.3 NSLAI: Bare Soil Evaporation

Bare soil (or direct) evaporation (LE_d) (Mahrt and Pan 1984) and the evaporative fraction for bare soil (ef_d) are

$$\begin{aligned} LE_d &= \rho_w L_v \left[\left(\frac{\Theta_{ns} - \Theta_{dry}}{\delta z} \right) D_\Theta + K_\Theta \right], \\ ef_d &= \frac{\rho_w L_v}{R_n - G} \left[\frac{\delta \Theta_{ns}}{\delta z} D_\Theta + K_\Theta \right], \end{aligned} \quad (89)$$

where ρ_w is water density, L_v is latent heat, Θ_{dry} is the soil moisture air-dry value which is the lower limit on surface soil moisture where evaporation can remain at the potential rate via soil moisture supplied from the soil ($\delta \Theta_{ns}$ is the near-surface soil moisture availability), and D_Θ and K_Θ are soil water diffusivity and soil hydraulic conductivity, respectively, both functions of soil moisture and soil type. Following Clapp and Hornberger (1978) and Cosby et al. (1984), D_Θ and K_Θ are defined as

$$\begin{aligned} D_\Theta &= \frac{b K_{\Theta_{sat}} \psi_{sat}}{\Theta_{sat}} \left(\frac{\Theta_{ns}}{\Theta_{sat}} \right)^{\beta+2}, \\ K_\Theta &= K_{\Theta_{sat}} \left(\frac{\Theta_{ns}}{\Theta_{sat}} \right)^{2\beta+3}, \end{aligned} \quad (90)$$

where $K_{\Theta_{sat}}$ is saturated soil hydraulic conductivity, a function of soil type. Alternate functions for soil water diffusivity and soil hydraulic conductivity may be used, e.g., van Genuchten (1980).

When the value of soil moisture at the surface (Θ_{sfc}) is sufficiently wet such that $\Theta_{sfc} > \Theta_{dry}$, then evaporation proceeds at the potential rate ($E_{dir} = E_{pot}$: atmospheric demand control stage), so $ef_d = ef_p$. This corresponds to near-surface soil moisture (Θ_{ns}) where

$$\Theta_{ns} \geq \Theta_{dry} + \frac{\delta z}{D_\Theta} \left[\frac{ef_p(R_n - G)}{\rho_w L_v} - K_\Theta \right]. \quad (91)$$

But as the soil dries out, $\Theta_{sfc} = \Theta_{dry}$ and evaporation proceeds only at the rate that the soil can diffuse water upward from below, so that evaporation is less than the potential rate ($E_{dir} < E_{pot}$: soil moisture flux control stage), so $ef_d < ef_p$.

The changes in D_Θ and K_Θ with changing soil moisture are

$$\begin{aligned} \frac{\partial D_\Theta}{\partial \Theta} &= (\beta + 2) \frac{D_\Theta}{\Theta_{ns}}, \\ \frac{\partial K_\Theta}{\partial \Theta} &= (2\beta + 3) \frac{K_\Theta}{\Theta_{ns}}. \end{aligned} \quad (92)$$

For the atmospheric demand control stage ($ef_d = ef_p$), the change in bare soil evaporative fraction with changing soil moisture is given by Eq. (75). But for the

soil moisture flux control stage ($ef_d < ef_p$), the change in *bare soil* evaporative fraction with changing soil moisture is then

$$\frac{\partial \ln ef_d}{\partial \Theta} = \frac{1}{\Theta_{ns}} \left\{ \frac{[\Theta_{ns} + (\beta + 2)\delta\Theta_{ns}]s_\Theta + (2\beta + 3)}{1 + \delta\Theta_{ns}s_\Theta} + \frac{b\beta G}{R_n - G} \right\}, \quad (93)$$

where $s_\Theta = D_\Theta / (\delta z K_\Theta)$.

Bare Soil Surface, Strong Land-Atmosphere Coupling. Conditions that lead to a large change in ef_d with changing Θ_{ns} (i.e., strong land-atmosphere coupling) include dry soil (small Θ_{ns}) and a small value of s_Θ , and a large soil heat flux to available energy ratio ($G/(R_n - G)$) where

$$\frac{\partial \ln ef_d}{\partial \Theta} \rightarrow \frac{(2\beta + 3)}{\Theta_{ns}} + \frac{b\beta}{\Theta_{ns}} \frac{G}{(R_n - G)} \quad (94)$$

Bare Soil Surface, Weak Land-Atmosphere Coupling. Conversely, conditions that lead to a small change in ef_d with changing Θ_{ns} (i.e., weak land-atmosphere coupling) include moist soil (large Θ_{ns}) and a large value of s_Θ , and a small soil heat flux to available energy ratio ($G/(R_n - G)$) where

$$\frac{\partial \ln ef_d}{\partial \Theta} \rightarrow 0. \quad (95)$$

5.1.4 NSAI: Canopy Evaporation

Evaporation of canopy-intercepted water (LE_c) and the evaporative fraction for canopy evaporation (ef_c) are

$$\begin{aligned} LE_c &= \left(\frac{C_w}{S_w} \right)^n E_p, \\ ef_c &= \left(\frac{C_w}{S_w} \right)^n ef_p, \\ &= \left(\frac{C_w}{S_w} \right)^n \left(\frac{s + \frac{\rho c_p g_a \delta e}{R_n - G}}{s + \gamma} \right), \end{aligned} \quad (96)$$

where Eq. (70) has been used, and C_w (S_w) is the canopy water content (*storage capacity*), and $n = 0.5$ (following Pan and Mahrt 1987, who cite earlier studies). Using Eq. (75), the change in *canopy water* evaporative fraction with changing soil moisture is then

$$\frac{\partial \ln ef_c}{\partial \Theta} = \left(\frac{C_w}{S_w} \right)^n \left[\frac{s(R_n - G)}{\rho c_p g_a \delta e} + 1 \right]^{-1} \frac{b\beta G}{\Theta_{ns}(R_n - G)}. \quad (97)$$

Canopy Evaporation Case, Strong Land-Atmosphere Coupling. Conditions that lead to a large change in ef_c with changing Θ_{ns} (i.e., strong land-atmosphere coupling) include a wet canopy ($C_w \rightarrow S_w$), strong surface-layer turbulence ($g_a \gg 0$), very dry air ($\delta e \gg 0$) and dry soil (Θ_{ns} small), and a large soil heat flux to available energy ratio ($G/(R_n - G)$) which yield

$$\frac{\partial \ln ef_c}{\partial \Theta} \rightarrow \frac{b\beta G}{\Theta_{ns}(R_n - G)}, \quad (98)$$

which is the same as Eq. (76) for the potential evaporation case.

Canopy Evaporation Case, Weak Land-Atmosphere Coupling. Conversely, conditions that lead to a small change in ef_c with changing Θ_{ns} (i.e., weak land-atmosphere coupling) include a dry canopy ($C_w \rightarrow 0$), weak surface-layer turbulence ($g_a \rightarrow 0$), very humid air ($\delta e \rightarrow 0$) and moist soil (Θ_{ns} large), and a small soil heat flux to available energy ratio ($G/(R_n - G)$) which yield

$$\frac{\partial \ln ef_c}{\partial \Theta} \rightarrow 0. \quad (99)$$

5.1.5 NSAI: Total Evapotranspiration

The total evapotranspiration (LE) and *evapotranspirative fraction* (ef) may be determined using terms from Sects. 5.1.2, 5.1.3, and 5.1.4, weighted by the green vegetation fraction (σ_f), so

$$\begin{aligned} LE &= (1 - \sigma_f)E_d + \sigma_f \left[1 - \left(\frac{C_w}{S_w} \right)^n \right] E_t + \sigma_f E_c, \\ ef &= (1 - \sigma_f)ef_d + \sigma_f \left[1 - \left(\frac{C_w}{S_w} \right)^n \right] ef_t + \sigma_f ef_c, \end{aligned} \quad (100)$$

where $0 \leq \sigma_f \leq 1$. The corresponding change in *evapotranspirative fraction* with changing soil moisture is then

$$\frac{\partial \ln ef}{\partial \Theta} = (1 - \sigma_f) \frac{\partial \ln ef_d}{\partial \Theta} + \sigma_f \left[1 - \left(\frac{C_w}{S_w} \right)^n \right] \frac{\partial \ln ef_t}{\partial \Theta} + \sigma_f \frac{\partial \ln ef_c}{\partial \Theta}. \quad (101)$$

5.2 Land-ABL Interaction

The *atmospheric leg* of land-atmospheric interaction determines the coupling between the surface fluxes and ABL development. Ek and Mahrt (1994) and Ek and Holtslag (2004) examined the daytime evolution of ABL-top relative humidity which is expected to control ABL cloud development. They showed that the relative humidity tendency at

the ABL top involves a number of competing mechanisms, with relative humidity directly *increasing* due to surface evaporation and due to ABL growth (ABL-top temperature decreases), and relative humidity directly *decreasing* due to surface sensible heat flux and due to entrainment of warm and dry air into the ABL from above. The *indirect* role of surface evaporation is to reduce surface heating, thereby competing with ABL growth that is reduced due to reduced surface heating, although this diminishes ABL-top warm-and dry-air entrainment. In a similar type of study, De Bruin (1983) examined the effect of different land-surface and ABL processes on the Priestley-Taylor parameter (used in relating surface available energy to surface evaporation).

To further understand the role of soil moisture and other factors in ABL cloud development, we follow Ek and Holtslag (2004) and examine a useful equation for relative humidity tendency at the ABL top which assumes a well-mixed ABL, uses the Clausis-Clapeyron relationship, equation of state, and definition of potential temperature, as well as expressions for ABL growth ($\partial h / \partial t$) and dry air entrainment flux ($\overline{w'q'_h}$) from Tennekes (1973) and Betts (1973), respectively, i.e.,

$$\frac{\partial h}{\partial t} = \frac{\overline{w'\theta'_s}(1 + C_\theta)}{h\gamma_\theta}, \quad (102)$$

$$\overline{w'q'_h} = -\Delta q \frac{\partial h}{\partial t}.$$

where h is boundary-layer depth and t is time, $\overline{w'\theta'_s}$ is the surface sensible heat flux, C_θ is the (negative of the) ratio of ABL-top to surface sensible heat flux, γ_θ is the potential temperature lapse rate above the ABL, and Δq is the change in specific humidity across the ABL top (which is normally negative).

The RH tendency at the ABL-top can then be given as

$$\frac{\partial RH}{\partial t} = \left(\frac{R_n - G}{\rho L_v h q_s} \right) [e_f + ne(1 - e_f)], \quad (103)$$

where $R_n - G$ is available energy at the surface (R_n is net radiation and G is soil heat flux), ρ is air density, L_v is latent heat, h is ABL depth, and q_s is saturation specific humidity just below the ABL top. In Eq. (103), e_f is the surface evaporative fraction (of surface energy available for evaporation) defined as

$$e_f = \frac{LE}{R_n - G} = \frac{LE}{H + LE}, \quad (104)$$

where LE and H are the surface latent and sensible heat fluxes, respectively. Furthermore, $ne(1 - e_f)$ reflects the direct effects of nonevaporative processes on relative humidity tendency, where ne is given by

$$\begin{aligned}
ne &= L_v/c_p(1 + C_\theta) \left[\frac{\Delta q}{h\gamma_\theta} + RH \left(\frac{c_2}{\gamma_\theta} - c_1 \right) \right], \\
c_1 &= \frac{L_v}{R_v} \frac{q_s}{T^2} \left(\frac{p}{p_s} \right)^{R_d/c_p}, \\
c_2 &= \left(\frac{L_v}{R_v} \frac{q_s}{T^2} - \frac{c_p}{R_d} \frac{q_s}{T} \right) \frac{g}{c_p}.
\end{aligned} \tag{105}$$

where c_p is specific heat, and c_1 and c_2 are functions of surface pressure, temperature and pressure at the ABL top, and constants. ne consists of three terms (each multiplied by $L_v/c_p(1 + C_\theta)$): ABL-top dry-air entrainment ($\Delta q/h\gamma_\theta$, a negative contribution to ABL-top relative humidity tendency), boundary-layer growth (RHc_2/γ_θ , a positive contribution), and boundary layer heating through surface warming and ABL-top warm-air entrainment (RHc_1 , a negative contribution).

From Eq. (103) we see that the relative humidity tendency is proportional to available energy and inversely proportional to ABL depth and temperature (via saturation specific humidity), while the sign of the relative humidity tendency is determined by the sign of $e_f + ne(1 - e_f)$. Examining Eq. (103), it is apparent that the direct role of e_f is to increase the ABL-top relative humidity, while the indirect role of surface evaporation (via reduced surface heating and diminished ABL growth and entrainment) is found in the expression $ne(1 - e_f)$. Figure 15 shows how $e_f + ne(1 - e_f)$ depends on e_f versus ne , where $e_f + ne(1 - e_f)$ is the relative humidity tendency, $\partial RH/\partial t$, normalized by the available energy term, $(R_n - G)/(\rho\lambda_v hq_s)$.

When the above-ABL atmospheric stability is rather strong (larger γ_θ), or if the stability is rather weak and the above-ABL air is rather dry (larger Δq), then $ne < 1$ so that $\partial RH/\partial t$ increases as e_f increases, confirming intuition. (For the range $0 < ne < 1$, $\partial RH/\partial t > 0$ and increases with increasing e_f , while for $ne < 0$, $\partial RH/\partial t > 0$ only when e_f exceeds some threshold value which increases for increasingly negative values of ne). This is the *surface-moistening regime* where soil moisture acts to increase ABL-top relative humidity and thus increases the probability of ABL cloud development given a sufficient initial ABL relative humidity.

On the other hand, with weaker above-ABL stability (smaller γ_θ), boundary-layer growth is less restricted over drier soils than over moister soils compared to the case with stronger stability. So with above-ABL air not too dry, then $ne > 1$ so that $\partial RH/\partial t$ increases as e_f decreases, which is somewhat counter-intuitive. This is the *ABL-growth regime* where soil moisture acts to limit the increase of ABL-top relative humidity and thus decreases the probability of ABL cloud development. **Note that the largest values of $\partial RH/\partial t$ are achieved for $ne > 1$ suggesting that the greatest potential for ABL cloud development is not over moist soils, but rather over dry soils with weak stability and above-ABL air not too dry given a sufficient initial ABL relative humidity.**

From Eqs. (103), (104), and (105), note that with drier air above the ABL (increasingly negative Δq), the value of ne decreases, and that as the soil moisture

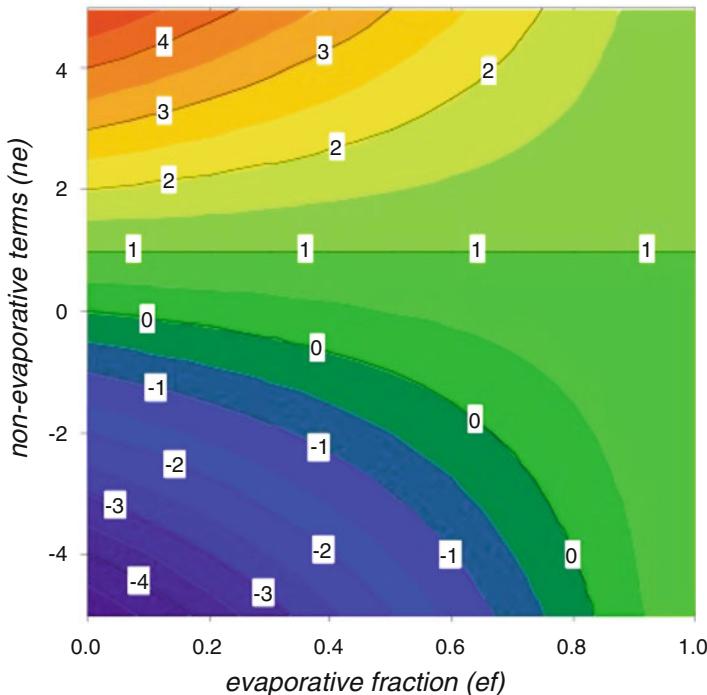


Fig. 15 ABL-top relative humidity tendency equation, $e_f + ne(1 - e_f)$ (normalized by the available energy term), as a function of evaporative fraction (e_f) versus nonevaporative processes (ne)

increases, generally e_f increases (that increase depending on the precise relationship between changing soil moisture and surface evaporation, i.e., as discussed in the Sect. 5.1; also see Wetzel and Chang 1987). But a change in stability above the ABL (γ_θ) affects both dry-air entrainment and boundary layer growth, two opposing processes in the ABL-top relative humidity tendency equation. So, only if the above-ABL specific humidity drop is greater (less negative) than some threshold $\Delta q > -RHc_2$ (at the ABL top), will ne increase with decreasing stability, which corresponds to $ne > -L_v/c_p(1 + C_\theta)RHc_1$. Note that this threshold value of Δq decreases (becomes *more* negative) for increasing RH , h , and c_2 (decreasing T). Finally, as $\Delta q \rightarrow 0$, $ne > 0$ for $\gamma_\theta < c_2/c_1 < \approx g/c_p \approx 1^\circ\text{C}/100\text{ m}$ (dry adiabatic lapse rate).

6 Summary

This review of land-surface modeling and land-atmosphere interaction provides a basis to examine these processes in future observational and modeling studies, and further conceptual and theoretical developments. Land processes and the complex interaction of land and ABL processes must be better understood in order to be

properly modeled. A number of efforts are addressing this need for process-level improvement in models, e.g., the international *Land Model Benchmarking* and *Local Land-Atmosphere Coupling (LoCo)* projects led by the Global Land/Atmosphere System Study (GLASS) panel in the Global Energy and Water Exchanges (GEWEX) project, where GEWEX is a core project of the World Climate Research Program (WCRP). See www.gewex.org.

References

- G. Al Nakshabandi, H. Kohnke, Thermal conductivity and diffusivity of soils as related to moisture tension and other physical properties. *Agric. Meteorol.* **2**, 271–279 (1965)
- M. Barlage, X. Zeng, H. Wei, K.E. Mitchell, A global 0.05 maximum albedo dataset of snow-covered land based on MODIS observations. *Geophys. Res. Lett.* **32**(17), 8851 (2005). <https://doi.org/10.1029/2005GL022881>
- A.C.M. Beljaars, F.C. Bosveld, Cabauw data for the validation of land surface parameterization schemes. *J. Clim.* **10**, 1172–1193 (1997)
- A.C.M. Beljaars, A.A.M. Holtslag, Flux parameterization over land surfaces for atmospheric models. *J. Appl. Meteorol.* **30**, 327–341 (1991)
- A.K. Betts, Non-precipitating cumulus convection and its parameterization. *Q. J. R. Meteorol. Soc.* **99**, 178–196 (1973)
- A. Boone, P. Etchevers, An inter-comparison of three snow schemes of varying complexity coupled to the same land-surface model: Local scale evaluation at an Alpine site. *J. Hydrometeorol.* **2**, 374–394 (2001)
- A. Boone, V. Masson, T. Meyers, J. Noilhan, The influence of the inclusion of soil freezing on simulations by a soil-vegetation-atmosphere transfer scheme. *J. Appl. Meteorol.* **39**, 1544–1569 (2000)
- J.A. Businger, J.C. Wyngaard, Y. Izumi, E.F. Bradley, Flux-profile relationships in the atmospheric surface layer. *J. Atmos. Sci.* **28**, 181–189 (1971)
- S. Chang, D. Hahn, C.-H. Yang, D. Norquist, M. Ek, Validation study of the CAPS model land surface scheme using the 1987 Cabauw/PLPS dataset. *J. Appl. Meteorol.* **38**, 405–422 (1999)
- F. Chen, K. Mitchell, J. Schaake, Y. Xue, H.-L. Pan, V. Koren, Q.Y. Duan, M. Ek, A. Betts, Modeling of land-surface evaporation by four schemes and comparison with FIFE observations. *J. Geophys. Res.* **101**, 7251–7268 (1996)
- R.B. Clapp, G.M. Hornberger, Empirical equations for some soil hydraulic properties. *Water Resour. Res.* **14**, 601–604 (1978)
- B.J. Cosby, G.M. Hornberger, R.B. Clapp, T.R. Ginn, A statistical exploration of the relationship of soil moisture characteristics to the physical properties of soils. *Water Resour. Res.* **20**, 682–690 (1984)
- R.H. Cuenca, M. Ek, L. Mahrt, Impact of soil water property parameterization on atmospheric boundary-layer simulation. *J. Geophys. Res.* **101**, 7269–7277 (1996)
- H.A.R. De Bruin, A model for the Priestley-Taylor parameter α . *J. Clim. Appl. Meteorol.* **22**, 572–578 (1983)
- P.A. Dirmeyer et al., Verification of land-atmosphere coupling in forecast models, reanalyses, and land surface models using flux site observations. *J. Hydrometeorol.* (2018). <https://doi.org/10.1175/JHM-D-17-0152.1>
- M. Ek, R.H. Cuenca, Variation in soil parameters: implications for modeling surface fluxes and atmospheric boundary-layer development. *Bound.-Layer Meteorol.* **70**, 369–383 (1994)
- M. Ek, A.A.M. Holtslag, Influence of soil moisture on boundary-layer cloud development. *J. Hydrometeorol.* **5**, 86–99 (2004)
- M. Ek, L. Mahrt, Daytime evolution of relative humidity at the boundary-layer top. *Mon. Weather Rev.* **122**, 2709–2721 (1994)

- M. Ek, K.E. Mitchell, Y. Lin, E. Rogers, P. Grunmann, V. Koren, G. Gayno, J.D. Tarpley, Implementation of Noah land-surface model advances in the NCEP operational mesoscale Eta model. *J. Geophys. Res.* **108**(D22), 8851 (2003). <https://doi.org/10.1029/2002JD003296>
- M.A. Friedl, D. Sulla-Menashe, B. Tan, A. Schneider, N. Ramankutty, A. Sibley, X. Huang, MODIS Collection 5 global land cover: algorithm refinements and characterization of new datasets. *Remote Sens. Environ.* **114**, 168–182 (2010). <https://doi.org/10.1016/j.rse.2009.08.016>
- G. Gutman, On the use of long-term global data of land reflectances and vegetation indices derived from the advanced very high resolution radiometer. *J. Geophys. Res.* **104**, 62416255 (1999). <https://doi.org/10.1029/1998JD200106>
- M.C. Hansen, R.S. DeFries, J.R.G. Townshend, R. Sohlberg, Global land cover classification at 1 km spatial resolution using a classification tree approach. *Int. J. Remote Sens.* **21**, 13311364 (2000)
- A.A.M. Holtslag, A.C.M. Beljaars, Surface flux parameterization schemes; developments and experiences at KNMI, in *Proceedings of Workshop on Parameterization of Fluxes and Land Surfaces*, 24–26 Oct 1988 (ECMWF, Reading, 1989), pp. 121–147. (Also available as KNMI Sci. Rep. 88-06, 27 pp, 1988, De Bilt, Netherlands.)
- A.A.M. Holtslag, H.A.R. de Bruin, Applied modeling of the night-time surface energy balance over land. *J. Appl. Meteorol.* **27**, 689–704 (1988)
- A.A.M. Holtslag, M. Ek, Simulation of surface fluxes and boundary layer development over the pine forest in HAPEX-MOBILHY. *J. Appl. Meteorol.* **35**, 202–213 (1996)
- C.M.J. Jacobs, E.J. Moors, H.W. Ter Maat, A.J. Teuling, G. Balsamo, K. Bergaoui, J. Ettema, M. Lange, B.J.J.M. van den Hurk, P. Viterbo, W. Wergen, Evaluation of European Land Data Assimilation system (ELDAs) products using in situ observations. *Tellus*. **60A**(5), 1023–1037 (2008)
- P.G. Jarvis, The interpretation of the variations in leaf water potential and stomatal conductance found in canopies in the field. *Philos. Trans. R. Soc. Lond. B* **273**, 593–610 (1976)
- P.G. Jarvis, K.G. McNaughton, Stomatal control of transpiration: scaling up from leaf to region. *Adv. Ecol. Res.* **15**, 1–49 (1986)
- O. Johansen, *Thermal Conductivity of Soils (in Norwegian)*, Ph.D. thesis, Publ. ADA 044002, Trondheim, 1975. (English translation 637, Cold Reg. Res and Eng. Lab., Hanover, N.H., 1977)
- V. Koren, J. Schaake, K. Mitchell, Q.-Y. Duan, F. Chen, J. Baker, A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. *J. Geophys. Res.* **104**(D16), 19,569–19,585 (1999)
- J.-F. Louis, A parametric model of vertical eddy fluxes in the atmosphere. *Bound.-Layer Meteorol.* **17**, 187–202 (1979)
- J.-F. Louis, M. Tiedke, J.F. Geleyn, A short history of the operational PBL-parameterization at ECMWF, in *Proceedings of the ECMWF Workshop on Planetary Boundary Layer Parameterisation*, European Centre for Medium-Range Weather Forecasts, Reading, 25–27 Nov 1981 (1982), pp. 59–80
- V.J. Lunardini, *Heat Transfer in Cold Climates* (Van Nostrand Reinhold Co., New York, 1981), 731 pp
- L. Mahrt, Grid-averaged surface fluxes. *Mon. Weather. Rev.* **115**, 1550–1560 (1987)
- L. Mahrt, M. Ek, The influence of atmospheric stability on potential evaporation. *J. Clim. Appl. Meteorol.* **23**, 222–234 (1984)
- L. Mahrt, H.-L. Pan, A two-layer model of soil hydrology. *Bound.-Layer Meteorol.* **29**, 1–20 (1984)
- C.H. Marshall, K.C. Crawford, K.E. Mitchell, D.J. Stensrud, The impact of the land surface physics in the operational NCEP Eta model on simulating the diurnal cycle: evaluation and testing using Oklahoma Mesonet data. *Weather Forecast.* **18**, 748–768 (2003)
- M.C. McCumber, R.A. Pielke, Simulation of the effects of surface fluxes of heat and moisture in a mesoscale numerical model. I. Soil layer. *J. Geophys. Res.* **86**(C10), 9929–9938 (1981)
- J.L. Monteith, Evaporation and environment. *Symp. Soc. Exp. Biol.* **19**, 205–234 (1965)
- G.-Y. Niu et al., The community Noah land surface model with multi-parameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *J. Geophys. Res.* **116**, D12109 (2011). <https://doi.org/10.1029/2010JD015139>

- J. Noilhan, S. Planton, A simple parameterization of land surface processes for meteorological models. *Mon. Weather Rev.* **117**, 536–549 (1989)
- H.-L. Pan, L. Mahrt, Interaction between soil hydrology and boundary-layer development. *Bound.-Layer Meteorol.* **38**, 185–202 (1987)
- C.A. Paulson, The mathematical representation of wind speed and temperature profiles in the unstable atmospheric surface layer. *J. Appl. Meteorol.* **9**, 857–861 (1970)
- H.L. Penman, Natural evaporation from open water, bare soil, and grass. *Proc. R. Soc. Lond.* **A193**, 120–146 (1948)
- C.D. Peters-Lidard, M.S. Zion, E.F. Wood, A soil-vegetation-atmosphere transfer scheme for modeling spatially variable water and energy balance processes. *J. Geophys. Res.* **102**(D4), 4303–4324 (1997)
- C.D. Peters-Lidard, E. Blackburn, X. Liang, E.F. Wood, The effect of soil thermal conductivity parameterization on surface energy fluxes and temperatures. *J. Atmos. Sci.* **55**, 1209–1224 (1998)
- J. Santanello, P.A. Dirmeyer, et al., Land-atmosphere interactions: the LoCo perspective. *Bull. Am. Meteorol. Soc.* (2017). <https://doi.org/10.1175/BAMS-D-17-0001.1>
- J.C. Schaake, V.I. Koren, O.-Y. Duan, K. Mitchell, F. Chen, Simple water balance model for estimating runoff at different spatial and temporal scales. *J. Geophys. Res.* **101**, 7461–7475 (1996)
- G.E. Schwarz, R.B. Alexander, *Soils Data for the Conterminous United States Derived from the NRCS State Soil Geographic (STATSGO) Data Base. Edition: 1.1* (U.S. Geological Survey, Reston, 1995). Publication Date: 19950901
- T.G. Smirnova, J.M. Brown, S.G. Benjamin, D. Kim, Parameterization of cold season processes in the MAPS land-surface scheme. *J. Geophys. Res.* **105** (D3) 4077–4086 (2000)
- J.B. Stewart, Modeling surface conductance of pine forest. *Agric. For. Meteorol.* **43**, 19–35 (1988)
- H. Tennekes, A model for the dynamics of the inversion above a convective boundary layer. *J. Atmos. Sci.* **30**, 558–567 (1973)
- USDA (United States Department of Agriculture), Natural Resources Conservation Service, Soil Survey Staff. Web Soil Survey (1995). Available online at <http://websoilsurvey.nrcs.usda.gov>
- B.J.J.M. van den Hurk, A.C.M. Beljaars, Impact of some simplifying assumptions in the new ECMWF surface scheme. *J. Appl. Meteorol.* **35**, 1333–1343 (1996)
- B.J.J.M. van den Hurk, A.A.M. Holtslag, On the bulk parameterization of surface fluxes for various conditions and parameter ranges. *Bound.-Layer Meteorol.* **82**, 119–134 (1997)
- B.J.J.M. van den Hurk, A. Verhoef, A.R. van den Berg, H.A.R. de Bruin, An intercomparison of three vegetation/soil models for a sparse vineyard canopy. *Q. J. R. Meteorol. Soc.* **121**, 1867–1889 (1995)
- B.J.J.M. van den Hurk, P. Viterbo, A.C.M. Beljaars, A.K. Betts, *Offline validation of the ERA40 surface scheme*, European Centre for Medium-Range Weather Forecasts, Technical memorandum No. 295 (ECMWF, Reading, 2000)
- M.Th. van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.* **44**, 892–898 (1980).
- A.P. van Ulden, A.A.M. Holtslag, Estimation of atmospheric boundary layer parameters for diffusion applications. *J. Clim. Appl. Meteorol.* **24**, 1198–1207 (1985)
- P. Viterbo, A.C.M. Beljaars, An improved land surface parameterization scheme in the ECMWF model and its validation. *J. Clim.* **8**, 2716–2748 (1995)
- P. Viterbo, A.C.M. Beljaars, J.-F. Mahfouf, J. Teixeira, The representation of soil moisture freezing and its impact on the stable boundary layer. *Q. J. R. Meteorol. Soc.* **125**, 2401–2426 (1999)
- E.K. Webb, Profile relationships: the log-linear range, and extension to strong stability. *Q. J. R. Meteorol. Soc.* **96**, 67–90 (1970)
- P.J. Wetzel, J.-T. Chang, Concerning the relationship between evaporation and soil moisture. *J. Clim. Appl. Meteorol.* **26**, 18–27 (1987)
- Z.-L. Yang et al., The community Noah land surface model with multi-parameterization options (Noah-MP): 2. Evaluation over global river basins. *J. Geophys. Res.* **116**, D12110 (2011). <https://doi.org/10.1029/2010JD015140>

Part V

Model Parameter Estimation and Uncertainty Analysis



Parameter Estimation and Predictive Uncertainty Quantification in Hydrological Modelling

Dmitri Kavetski

Contents

1	Introduction	482
2	Basic Concepts of Parameter Estimation	484
2.1	Basic Setup of the Calibration Problem	484
2.2	A Priori Estimation	485
2.3	Calibration	486
2.4	Manual Calibration	486
2.5	Goodness-Of-Fit Function as an Optimization Objective	488
2.6	Other Objective Functions: How Different Are They?	490
3	Automatic Calibration Via Optimization	491
4	Multi-Objective Optimization	494
5	Probabilistic/Statistical Uncertainty Quantification	494
5.1	Bayesian Inference: General Principles	495
5.2	Least Squares Techniques as Gaussian Error Models	497
5.3	Tools for Analyzing Bayesian Posteriors	499
5.4	Aggregational Methods	500
5.5	Decompositional Methods	504
5.6	Methods Other than Bayesian and Other than Probabilistic	505
6	Model Diagnostics as Part of Parameter Estimation	507
7	Practicalities	510
7.1	Parameter Transformations	510
7.2	Impact of Model Non-smoothness/Discontinuities	511
7.3	Initial Conditions: Estimate or Warm-Up	512
7.4	Estimation of Expensive Models	512

D. Kavetski (✉)

School of Civil, Environmental and Mining Engineering, University of Adelaide,
Adelaide, SA, Australia

School of Engineering, University of Newcastle, Callaghan, NSW, Australia

Department of Systems Analysis, Integrated Assessment and Modelling (SIAM), Eawag, Swiss
Federal Institute of Aquatic Science and Technology, Dübendorf, Switzerland
e-mail: dmitri.kavetski@adelaide.edu.au

8	Research Directions	513
8.1	Operational Improvements	513
8.2	Sparse-Data Problems	513
8.3	Recursive Estimation and Data Assimilation	514
9	Summary and Conclusion	514
	References	515

Abstract

The majority of hydrological and environmental models contain parameters that must be specified before the model can be used. Parameter estimation is hence a very common problem in environmental sciences and has received tremendous amount of research and industry attention. This chapter reviews some of the key principles of parameter estimation, with a focus on calibration approaches and uncertainty quantification. The distinct approaches of manual calibration, optimization, multi-objective optimization, and probabilistic approaches are described in terms of key theory and representative applications. Advantages and limitations of these strategies are listed and discussed, with a focus on their ability to represent parametric and predictive uncertainties. The role of posterior diagnostics to check calibration and model assumptions that impact on parameter estimation is emphasized. Auxiliary tricks and techniques are described to simplify the process of parameter estimation in practical applications. The chapter concludes with an outline of directions for ongoing and future research. It is hoped that this chapter will help hydrologists and environmental modellers get to the current state of research and practice in model calibration, parameter estimation, and uncertainty quantification.

Keywords

Hydrological model · Parameter estimation · Model calibration · Optimization · Bayesian inference · Uncertainty quantification

1 Introduction

Hydrological (rainfall-runoff) models are widely used in environmental sciences and engineering, including flood forecasting, water yield predictions, and so forth (e.g., Duan et al. 1992; Beven 1997; Lindstrom et al. 1997; Clark et al. 2008, 2015). In addition to being useful in their own right, predictions from hydrological models, particularly rainfall-runoff models, provide inputs to the planning and operation of water resource systems (Loucks et al. 1981). The scales of these applications vary from a single hillslope to entire continents (Archfield et al. 2015), and the prediction lead times vary from minutes in operational flood forecasting (Neal et al. 2012) to seasonal scales (Tuteja et al. 2011). Given the inherent uncertainty of environmental predictions, uncertainty quantification and risk assessment is another key aspect that is receiving increased attention in the literature (e.g., Vogel 2017; Reichert et al. 2015).

Hydrological models are often classified on a spectrum from black-box models to conceptual models to physical models. Typical black-box models are given by artificial neural networks (e.g., Govindaraju 2000; Kingston et al. 2008); physical “bottom-up” models can be defined as models based on contemporary understanding of physical laws (e.g., Freeze and Harlan 1969; Ivanov et al. 2004; Clark et al. 2015), with conceptual “top-down” models (e.g., Sivapalan et al. 2003a; Fenicia et al. 2011) somewhere in between these bookends. This classification provides useful guidance but is not always crisp, with most practical models not fitting neatly into a single category and instead exhibiting a mix of different modelling philosophies (e.g., Clark et al. 2011). Irrespective of their philosophy and mathematics, hydrological and environmental models almost always contain adjustable parameters, intended to describe the invariant properties of the system. Before a model can be used for simulation or prediction of a system of interest, its parameters must be specified.

This chapter deals with the problem of parameter estimation and uncertainty quantification. Broadly speaking, two types of estimation strategies can be distinguished: *a priori* and calibration (inverse modelling). *A priori* estimation seeks to assign parameter values based directly on observable physical quantities, e.g., soil properties, vegetation characteristics, and so forth (e.g., Koren et al. 2003). Calibration, in this work, refers to any procedure for estimating model parameters (and possibly their uncertainties) from available observations of quantities the model is supposed to predict (e.g., Tarantola 2005). *A priori* estimation tends to depend on model structure and physical basis (e.g., see the debates in Abbott et al. 2003; Pappenberger and Beven 2006). Calibration in this respect represents a more general mathematical operation. In principle any model can – or, as many have argued, should – be calibrated, yet in practice it is often a formidable challenge to calibrate a model suitable for extrapolation and reliably account for estimation uncertainties. An important subset of parameter estimation is estimation under data-scarce conditions, including in ungauged basins – these applications tend to use a combination of *a priori* estimation and calibration (e.g., see Hrachowitz et al. 2013).

In recognition of these challenges, parameter estimation in hydrology and environmental modelling is shifting from a reliance on manual expertise, especially in research applications – initially to the largely mathematical task of finding (optimizing) the “best” model parameters according to a given performance metric (e.g., Ibbitt and O’Donnell 1971; Gupta and Sorooshian 1985; Duan et al. 1992) – and ultimately to a more “holistic” treatment that seeks to reflect multiple competing objectives in model calibration (e.g., Gupta et al. 1998; Efstratiadis and Koutsoyiannis 2010), multiple sources of uncertainties (e.g., Beven and Binley 1992; Kavetski et al. 2002; Reichert and Mieleitner 2009; Renard et al. 2011), stringent diagnostics of model “realism” (e.g., Gupta et al. 2008; Clark et al. 2011), operational reliability (e.g., Krzysztofowicz 1999; Cloke and Pappenberger 2009; Wang et al. 2009; McInerney et al. 2017), and so forth.

The aims of this chapter are to review the main types of parameter estimation methods with a focus on calibration, to provide a rigorous but accessible summary of key ideas and methods, and to direct the interested reader to the rich scientific and operational literature. In the author’s experience, there is often a disconnect between

the intuitive objective function techniques used by practitioners and the statistically motivated likelihood functions used in the research literature. This chapter attempts to close this gap and provides a unified perspective that ties together seemingly distant techniques such as single- and multi-objective optimization, Bayesian inference, residual error diagnostics, and numerical model implementation aspects. Emphasis is placed on technical aspects and practical recommendations, including discussions of pros and cons of individual techniques. However, philosophical aspects are also relevant, and the practitioner should be aware of the types of assumptions being made and limitations arising thereof.

The chapter is structured as follows. Section 2 establishes the notation and background of parameter estimation, including a brief review of manual calibration and goodness-of-fit functions. Section 3 motivates automatic calibration – using digital computers rather than humans – and focuses on the optimization approach and its advantages, challenges, and limitations. Section 4 considers multiple competing objectives. Section 5 surveys the vast topic of probabilistic estimation and uncertainty quantification, with a focus on Bayesian techniques. Section 6 considers the critical topic of posterior diagnostics to check calibration assumptions. Section 7 picks up practicalities relevant to application, Sect. 8 describes ongoing research directions, and Sect. 9 wraps up with conclusions.

2 Basic Concepts of Parameter Estimation

2.1 Basic Setup of the Calibration Problem

A hydrological model $\mathcal{H}(\mathbf{x}; \boldsymbol{\theta})$ simulates catchment streamflow \mathbf{y} over a series of time steps t given forcing data \mathbf{x} and parameters $\boldsymbol{\theta}$:

$$\mathbf{y} = \mathcal{H}(\mathbf{x}; \boldsymbol{\theta}) \quad (1)$$

To keep notation simple for presentation purposes, it will be assumed that $\mathcal{H} = \{\mathcal{H}_t, t = 1, \dots, N_{\mathcal{H}}\}$ is a vector of length equal to the number of time steps in the forcing data $\mathbf{x} = \{\mathbf{x}_t, t = 1, \dots, N_{\mathbf{x}}\}$, i.e., $N_{\mathcal{H}} = N_{\mathbf{x}}$. A more general (but less transparent) notation can be deployed if the responses include streamflow at locations other than the catchment outlet, water depth levels, water quality, and so forth. Typical forcing required by hydrological models includes rainfall, potential evaporation, irrigation schedules, pumping schedules, and so forth. Most models contain internal states, typically storages across a collection of storage elements (conceptual models) or grid cells/finite elements (physical models discretized in space) (e.g., Singh and Woolhiser 2002; Fenicia et al. 2011; Clark et al. 2015, and many others).

In addition to inputs, outputs, and internal states, which are typically variables, models contain *parameters*, which are quantities intended to characterize the inherent properties of the modelled system (including the physical system *and* the observational system used to collect data). In Eq. (1), parameters are indicated as $\boldsymbol{\theta} = \{\theta_k, k = 1, \dots, N_{\boldsymbol{\theta}}\}$.

Parameters are typically defined subject to lower and upper bounds

$$\boldsymbol{\theta}^{(L)} \leq \boldsymbol{\theta} \leq \boldsymbol{\theta}^{(H)} \quad (2)$$

or more complex linear and nonlinear constraints (e.g., if parameters are mutually constrained).

The numbers of parameters in hydrological models vary widely. Lumped conceptual models, such as GR4J (Perrin et al. 2003), have just a handful of parameters intended to describe max storage values, routing characteristics, and groundwater exchange. Distributed physically based models such as SWAT (Arnold and Fohrer 2005) may have hundreds of parameters, describing soil hydraulic properties (such as conductivity), surface lags, crop growth rates, and so forth. The distinction between model parameters and states is not always clear-cut, and it frequently depends on the context in which the variables appear. For example, in some applications, parameters are defined as time- and/or state-dependent (e.g., Young 1998; Kuczera et al. 2006; Reichert and Mieleitner 2009; Young and Ratto 2009; Westra et al. 2012). Note that the term “model parameterization” is sometimes used to refer to the form of the model equations and their parametric dependencies and other times to the actual parameter values, which can cause confusion.

When parameter values are unknown, a model will generally be unable to reproduce even known data, let alone future unknown data. Hence, parameter estimation is among the first steps of deploying a model.

The following sections describe the two main parameter estimation strategies, namely, a priori estimation and calibration.

2.2 A Priori Estimation

A priori estimation refers to establishing parameter values from measured physical system properties. This strategy presupposes that model parameters have a sufficiently reliable physical interpretation (Abbott et al. 2003; Ivanov et al. 2004). For example, consider the specification of channel geometry in flood models – in the case of engineered channels, their width and length can be usually established from maps and other records. Parameter estimation in models of natural systems may require measurements and tests. For example, the hydraulic conductivity of soils, a parameter within physically based groundwater models such as MODFLOW (Harbaugh 2005), may be obtained from laboratory analysis of core samples, in situ tests, and/or geology maps (Fetter 1994). Other examples of a priori estimation might include the specification of channel roughness in hydraulic models using Manning’s equation (e.g., Streeter and Wylie 1983).

A priori estimation can be effective, especially when modelling well-instrumented locations using equations that embody our best current understanding of environmental physics (e.g., Ivanov et al. 2004; Clark et al. 2015). In contrast, when working with conceptual models, it has proven difficult to reliably relate their parameters to available information (Koren et al. 2003; Duan et al. 2006), though in

some cases, useful relationships appear possible (Samaniego et al. 2010). Parameter estimation from observable catchment characteristics is challenged by the tremendous spatial variability of soils and vegetation, both within and across basins (Miller and White 1999), as well as by the frequent problem of non-commensurability of modelled and observed quantities (Kuczera and Franks 2002). Another question relates to the estimation of parameter and predictive uncertainties; model structural errors are particularly difficult to estimate a priori without recourse to at least some observed data. Advances in physical process representation notwithstanding, it has been argued that models can only be described as “truly” physical if their parameters are specified independently from observed responses (Grayson et al. 1992). That said, even quantities currently seen to have a firm physical basis, such as Darcian hydraulic conductivity, were established empirically by fitting to experimental data (Brown 2002) – it can hence be argued that all practical environmental models begin their life as empirical quantities. This observation leads us to the general class of parameter estimation given by calibration (inverse modelling).

2.3 Calibration

The idea of model calibration is to find parameter values $\boldsymbol{\theta}^{(\text{cal})}$ such that, given a set of observed (“known”) inputs $\tilde{\mathbf{x}}$, the model \mathcal{H} reproduces a set of known outputs $\tilde{\mathbf{y}} = \{\tilde{y}_t, t = 1, \dots, N_{\tilde{y}}\}$, at least to a degree that is satisfactory for the application of interest. In algorithmic/mathematical notation:

$$\begin{aligned} \text{Calibration : Find } \boldsymbol{\theta}^{(\text{cal})} \text{ such that } \mathcal{H}\left(\tilde{\mathbf{x}}; \boldsymbol{\theta}^{(\text{cal})}\right) \approx \tilde{\mathbf{y}} \\ \boldsymbol{\theta}^{(\text{cal})} : \mathcal{H}\left(\tilde{\mathbf{x}}; \boldsymbol{\theta}^{(\text{cal})}\right) \approx \tilde{\mathbf{y}} \end{aligned} \quad (3)$$

Figure 1, in conjunction with Eq. (3), illustrates the basics of going from an uncalibrated to a calibrated model. While superficially simple, hydrological model calibration is a rather challenging task, especially once we recognize that a perfect model fit is unattainable and wish to characterize the attendant trade-offs and uncertainties. Different calibration approaches are then distinguished by aspects such as (i) how is the approximate equality in Eq. (3) expressed mathematically or even visually, (ii) how is the search for parameters conducted, (iii) how many sets of estimated parameter values are retained (e.g., to represent uncertainty), and so forth.

2.4 Manual Calibration

A hydrologist or engineer familiar with the model and catchment system of interest will often be able to find parameter values for which the model behaves in a reasonable way. For example, when working with a flood model, an engineer will generally try to match the flood peak magnitude, the total flow volume, and ideally

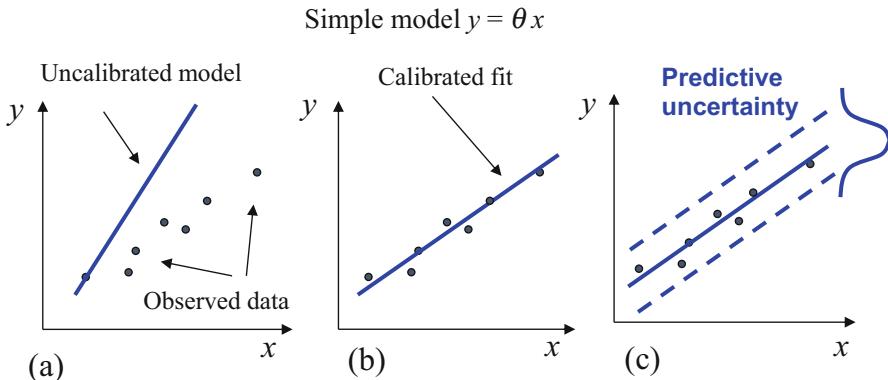


Fig. 1 Calibration concepts illustrated using a simple straight-line model. Panel *a* shows the uncalibrated model (poor value of model parameter θ), panel *b* shows the same model with a calibrated parameter θ , and panel *c* shows a more comprehensive application where model uncertainty is quantified

the timing (e.g., Maidment 1993). Characteristics such as the shape of recession would also be considered, though usually to a lesser extent as these aspects have less impact on flood damages. On the other hand, a water supply engineer working in an arid area may be far less interested in flood peaks but will try to get the model to reproduce low flows, which often contribute the most to cumulative flow volumes and hence to water availability. Finally, a hydrologist interested in understanding catchment dynamics may very well focus on the shape of recession, e.g., using master recession analysis (Tallaksen 1995).

Manual calibration allows the modellers to exploit their experience and hydrological understanding – which are formidable tools in the hands of an expert (Savenije 2009; Hrachowitz et al. 2014). Trade-offs in the ability of the model to reproduce different aspects of the data can be resolved based on the application objectives, once again exploiting expert judgment where available.

On the other hand, the subjectivity of manual calibration also creates inevitable weaknesses and limitations. Most notably, how do we establish if parameter set $\Theta^{(1)}$ is closer or further away from $\Theta^{(\text{cal})}$ than parameter set $\Theta^{(2)}$? The eye of an experienced modeller can provide superb expert judgment, but the resulting non-transparency and irreproducibility pose problems, both in scientific and operational contexts (Hill et al. 2015). Nor is it obvious how should a hydrologist quantify and report the uncertainty in manually calibrated parameters, especially if these parameters have been selected on the basis of a fit to visual hydrograph characteristics.

The laboriousness of manual calibration is another major practical downside – a human must select parameter values, run the model, inspect its output, suggest a new trial parameter set, rinse and repeat, and eventually decide when to stop. Clearly this is not only subjective but exhausting. Once again, these limitations become more pronounced in the case of large-scale national forecasting services, e.g., the US National Weather Service (NWS) (Demargne et al. 2014), the Australian Bureau of

Meteorology (Tuteja et al. 2017), and other agencies tasked with modelling and forecasting over hundreds and thousands of catchments spanning national- and continental-scale areas.

These limitations lead us to automatic calibration. But first we must solve the question of goodness-of-fit measures.

2.5 Goodness-Of-Fit Function as an Optimization Objective

The idea of a goodness-of-fit function Φ is to quantify how well the model reproduces the calibration data. Ideally we would like a perfect match of model to data, but we need to handle discrepancies in some reasonable way.

The most widely used goodness-of-fit function is the sum of squared errors (SSE), usually credited to Karl Gauss who developed it in the late eighteenth century (Merriman 1877):

$$\Phi_{\text{SSE}}(\boldsymbol{\theta}) = \Phi_{\text{SSE}}(\boldsymbol{\theta}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) = \sum_{t=1}^{N_y} (\tilde{y}_t - \mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}])^2 \quad (4)$$

where to avoid clutter, the dependence of Φ on $\tilde{\mathbf{x}}$ and $\tilde{\mathbf{y}}$ is omitted in the notation, and it is understood that, for time stepping models, $\mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}]$ only depends on inputs $\tilde{\mathbf{x}}_{1:t}$ up to and including time step t . Intuitively, the SSE function penalizes discrepancies between model and observations in a “reasonable” way (a larger discrepancy at any time step lowers the goodness of fit) and has the historical advantage that it is easy to manipulate analytically.

Given a goodness-of-fit function, model calibration problem can be articulated as an optimization problem:

$$\boldsymbol{\theta}^{(\text{opt})} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \Phi(\boldsymbol{\theta}) \quad (5)$$

subject to any external parameter constraints such as in Eq. (2).

In the optimization context given by Eq. (5), the goodness-of-fit function serves as the “objective” function – a naming convention that, perhaps unintentionally, hides that the choice of the error measure (e.g., the two-norm in Eq. (4)) is generally subjective. That said, Sect. 6 offers avenues to test these assumptions as part of the calibration process.

Figure 2 illustrates a typical least squares objective function of a hydrological model, shown as a cross section with respect to two parameters. Panel A provides an idealized schematic, with a well-defined optimum and smooth elliptic (quadratic) contours. Intuitively, the shape of the objective function indicates not just the optimal parameters but also parameter uncertainty: a peaky shape suggests well-defined parameters, whereas a flat shape indicates substantial uncertainty.

Figure 2 also illustrates parameter dependence – elongation of the objective function along certain parameter combinations. Parameter dependence is closely

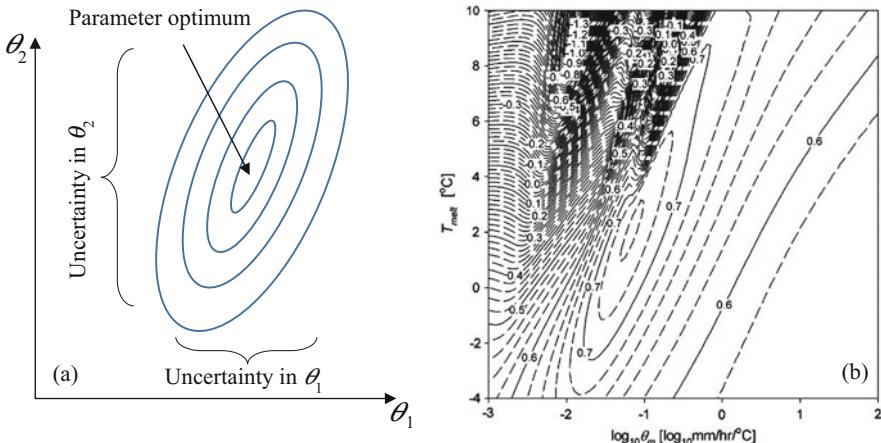


Fig. 2 Diagram of representative least squares objective functions. Panel *a* shows the idealized case of a two-parameter linear model, in which case the SSE objective function is exactly quadratic. Panel *b* (reproduced with permission from Elsevier) shows a real case example from the hydrological model case study of Kavetski et al. (2006c), where model nonlinearities lead to disturbances in the shape of the SSE

related to parameter identifiability: the objective function contours indicate parameter sets that produce predictions “indistinguishable” from each other according to the goodness-of-fit function. For example, consider the straight-line model from Fig. 1: if we simultaneously increase its slope and reduce its intercept, the changes could compensate for each other and maintain the same goodness-of-fit value. Characterizing these parameter interactions can yield insight into model deficiencies and is an important goal of parameter calibration and uncertainty quantification.

The shape of SSE functions depends on the model \mathcal{H} . Eq. (4) indicates that, to the extent that a model is linear with respect to its parameters, i.e., $\partial^2 \mathcal{H} / \partial \boldsymbol{\theta}^2 \approx \mathbf{0}$, its SSE objective function will be *quadratic* and have a single optimum irrespective of the data. In addition, a model that is smooth with respect to its parameters is guaranteed to have a smooth SSE objective function. In practice, most hydrological models exhibit nonlinearities, which induce irregularities in the objective function surface. Figure 2 panel *b* shows a well-behaved near-optimal region, as well as regions of irregular geometry and flat (insensitive) regions. In some instances, multiple optima can arise (e.g., Duan et al. 1992), which raises an even more immediate question of parameter identifiability than insensitive parameters.

The use of an objective function makes manual calibration more systematic, but its true power shines when used in conjunction with mathematical techniques such as optimization, which can find parameter optima analytically or numerically. For example, in the case of a straight-line model, $\mathcal{H}(x; a) = ax$, $\boldsymbol{\theta}^{(\text{opt})}$ is obtained analytically as

$$a^{(\text{opt})} = \sum_{t=1}^{N_{\tilde{\mathbf{y}}}} \tilde{x}_t \tilde{y}_t / \sum_{t=1}^{N_{\tilde{\mathbf{y}}}} \tilde{x}_t^2 \quad (6)$$

Similar expressions exist for more general linear models. Nonlinear models, for which the SSE function cannot be optimized analytically, can be handled using numerical optimization (Sect. 3). Before considering these techniques, it is insightful to consider alternative objective functions.

2.6 Other Objective Functions: How Different Are They?

The SSE objective function makes intuitive sense but is not without some limitations. For example, its values (and units) are not easy to interpret or compare across time series of unequal length. Two SSE-derived functions are common in hydrology, the root mean squared error (RMSE) and the Nash-Sutcliffe efficiency (NSE).

The RMSE metric, widely used in engineering and physics, is defined as

$$\Phi_{\text{RMSE}}(\boldsymbol{\theta}) = \sqrt{\frac{1}{N_{\tilde{\mathbf{y}}}} \sum_{t=1}^{N_{\tilde{\mathbf{y}}}} (\tilde{y}_t - \mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}])^2} = \sqrt{\frac{1}{N_{\tilde{\mathbf{y}}}} \Phi_{\text{SSE}}(\boldsymbol{\theta})} \quad (7)$$

It offers the advantage of having the same units as the quantity of interest (e.g., m³/s or mm/d in case of flowrates and catchment-average daily streamflow, respectively), as well as being scaled with respect to record length.

The NSE metric is a modification of the SSE function with a long tradition in hydrology (Nash and Sutcliffe 1970):

$$\Phi_{\text{NSE}}(\boldsymbol{\theta}) = 1 - \frac{\sum_{t=1}^{N_{\tilde{\mathbf{y}}}} (\tilde{y}_t - \mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}])^2}{\sum_{t=1}^{N_{\tilde{\mathbf{y}}}} (\tilde{y}_t - \text{ave}[\tilde{\mathbf{y}}])^2} = 1 - b \times \Phi_{\text{SSE}}(\boldsymbol{\theta}) \quad (8)$$

where $\text{ave}[\tilde{\mathbf{y}}]$ is the sample mean of observed data. The asymptotic identity $\Phi_{\text{NSE}}(\boldsymbol{\theta}) \xrightarrow[N_{\tilde{\mathbf{y}}} \rightarrow \infty]{} 1 - \left(\frac{\Phi_{\text{RMSE}}(\boldsymbol{\theta})}{\text{sdev}[\tilde{\mathbf{y}}]} \right)^2$, where $\text{sdev}[\cdot]$ denotes the standard deviation, elucidates that the NSE quantifies the fraction of streamflow variability captured by the hydrological model. Schaeafi and Gupta (2007) suggest generalizing the NSE by replacing $\text{ave}[\tilde{\mathbf{y}}]$ with a reference model, e.g., seasonal means, to provide a more informative and stringent benchmark.

The RMSE and NSE functions are related to the SSE kernel through monotonic transformations, and hence their optimal parameter sets (both local and global) are the same. Collectively, we shall refer to the optimization of these objective functions as least squares estimation. Connections to probabilistic estimation will be made in Sect. 5.2.

Genuinely different parameter estimates and objective function behavior are obtained with non-quadratic goodness-of-fit functions, e.g., the sum of absolute errors (SAE):

$$\Phi_{\text{SAE}}(\boldsymbol{\theta}) = \sum_{t=1}^{N_{\tilde{y}}} |\tilde{y}_t - \mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}]| \quad (9)$$

An attractive feature of the SAE formulation is that it is more robust with respect to outliers. SSE squares individual errors, which tends to exaggerate the influence of outliers – the calibration can distort parameter values just to get the model closer to the outlier. The idea of robust regression is to reduce the leverage of individual points, and weighting functions exist that discount outliers altogether, such as Tukey’s biweight and others (Press et al. 1992). That said, SAE functions are relatively uncommon in hydrology: their benefits are not always demonstrable, and the absolute value function is surely less smooth than SSE functions. Outlier detection and leverage analysis appear more attractive from the perspective of hydrological model setup and data analysis rather than just optimization (e.g., Wright et al. 2015; Hill et al. 2015).

The goodness-of-fit function framework allows the hydrologists to craft their own objective functions to reflect modelling objectives of interest – mimicking the hydrologist’s intuition mentioned in Sect. 2.4. For example, SSE and SAE functions can be computed separately for high and/or low flows; response weights and transformations such as logarithmic can be used to emphasize the fitting of particular data points and so forth (Chapman 1970; Chiew et al. 1993; Pushpalatha et al. 2012). These aspects are revisited in Sect. 5.4 from a statistical perspective. Further examples of goodness-of-fit functions are provided by Legates and McCabe Jr. (1999).

Several questions arise at this stage. Do we need to restrict attention to a single goodness-of-fit function? Given that some objective functions, notably SSE, RMSE, and NSE, have the same optimum but a different shape (curvature), how can we unambiguously quantify parameter uncertainty? And more generally, how do we navigate the vast range of potential performance measures? The following sections will describe how to overcome some of the challenges and present objective functions and calibration approaches from a more systematic perspective.

3 Automatic Calibration Via Optimization

The idea of automatic calibration is to reduce the need for human intervention and tackle the calibration problem in Eq. (5) using mathematical algorithms. As even simple goodness-of-fit function will be impossible to optimize analytically for most hydrological models, numerical optimization is employed. A plethora of numerical optimization algorithms have been employed in hydrological calibration, ranging from local methods such as classical Newton and quasi-Newton methods that

assume the objective function is smooth and near-quadratic (Gill et al. 1981) to global evolutionary methods such as the shuffled complex evolution (SCE) algorithm (Duan et al. 1992) and the dynamically dimensioned search (DDS) algorithm (Tolson and Shoemaker 2007) that make few if any such assumptions.

The selection of an optimization algorithm depends on the hydrological model and objective function. For example, Gauss-Newton-type algorithms are tailored to (possibly transformed) sum of squared errors (SSE) objective functions and are implemented in packages such as PEST (Doherty 2005), the Australian eWater Source platform (Welsh et al. 2013), and the BATEAU toolkit (Kavetski 2005) available to calibrate groundwater, water resources, and hydrological models.

In many cases, optimization works remarkably well. For example, a single-reservoir nonlinear model can be fitted to the Maimai catchment data using the Excel Solver tool, as shown in Fig. 3. For more complex modelling scenarios, calibration toolkits such as PEST (Doherty 2005) and BATEAU (Kavetski 2005) can be used, offering model coupling through ASCII input/output files and/or DLLs, visual interfaces, scripts, and other productivity features. Research is advancing into multi-start strategies and search randomization to increase the chance of finding the global optimum (e.g., Skahill and Doherty 2006; Kavetski et al. 2007; Tolson and Shoemaker 2007), as well as “multi-method” approaches that run multiple optimizers in parallel and pick the ones making the most progress (Vrugt and Robinson 2007).

That said, off-the-shelf optimization of hydrological models is not yet routinely attainable. Hydrological models with highly nonlinear dependence on their parameters have markedly non-quadratic objective functions, often exhibiting macroscale

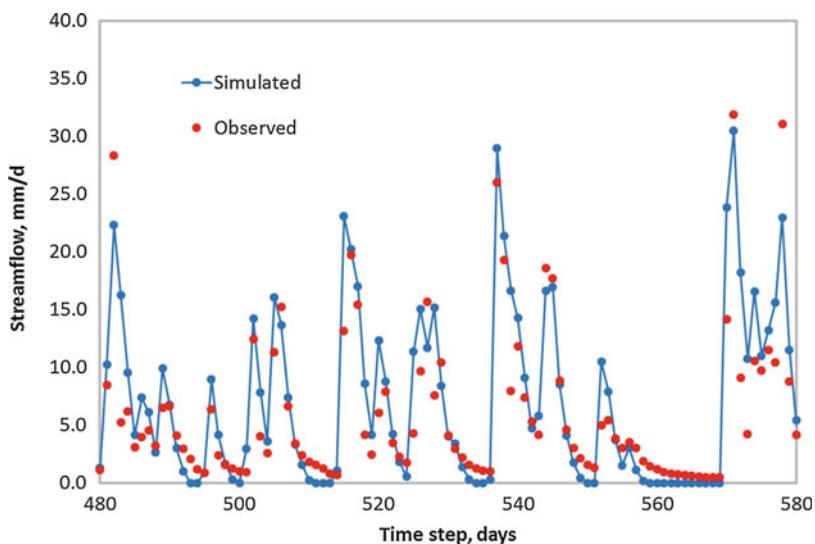


Fig. 3 Calibration of simple nonlinear reservoir bucket model, $dS/dt = P - kS^\alpha - E$, to the Maimai catchment, New Zealand

multi-optimality and microscale roughness (Duan et al. 1992). These problematic features are often exaggerated by fragile numerical implementations, such as the explicit Euler time stepping scheme, and by internal model thresholds (Kavetski and Kuczera 2007; Kavetski and Clark 2010). Under these conditions, gradient-based Newton-type algorithms typically converge only to the optimum nearest to the initial search point and generally behave erratically. Although current wisdom in hydrological modelling tends to favor evolutionary optimization methods, which tend to exhibit more robust global convergence and are less susceptible to microscale roughness, robust modifications of Newton-type methods offer the promise of comparable robustness at a much lower computational cost (Qin et al. 2018).

However, consider the following questions:

1. Identifiability problems. For example, Jakeman and Hornberger (1993) reported that typical rainfall-runoff data can support the identification of at most a “handful” of parameters in a lumped conceptual model. The optimization of distributed models solely to input-output data at the endpoints of their domain is clearly questionable – how would such data support the attribution of water flows through multiple internal pathways?
2. The very idea of looking solely for the global optimum at the expense of everything else can be questioned – nominally “slightly worse” optima may also be relevant and in some cases may provide more “realistic” model performance. In cases of pronounced multi-optimality, which could arise in case of grossly over-parameterized models, the optimum that becomes global may ultimately depend on data errors, subtle interplay of internal pathways, objective function idiosyncrasies, etc.
3. How do we estimate parameter uncertainty? For example, consider the endpoints of multiple optimization sequences – can these be treated as indicative of parameter uncertainty? Ultimately this approach gauges the effectiveness of the optimization algorithm and (potentially) the presence of multiple optima – somewhat counterintuitively, it would fail if the optimization algorithm is sufficiently robust to find the global optimum from most initial points. To estimate parameter uncertainty due to data and model errors, the shape of the objective function in the vicinity of the optimum should be investigated, e.g., using χ^2 methods (Press et al. 1992), which are related to the statistical ideas of Sect. 5.2.
4. A single objective is mathematically convenient but does not reflect the reality that multiple attributes might be of interest, e.g., low and high flows, timing of peaks, flow volumes, etc. (Sect. 2.4). In principle, a single-objective function can be constructed as a composite of multiple terms, e.g., separate SSE for low and high flows, water quality, etc., and melded together using weights. This approach goes some way toward recognizing the diverse nature of modelling objectives but is not quite “truly” multi-objective.

For these reasons, single-objective optimization on its own cannot be considered a complete solution to the calibration problem, even if it happens to be successful in terms of finding the global optimum.

4 Multi-Objective Optimization

Multi-objective optimization seeks to find the optimum of multiple objective functions simultaneously:

$$\boldsymbol{\theta}^{(\text{opt})} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} [\Phi_1(\boldsymbol{\theta}), \Phi_2(\boldsymbol{\theta}), \dots, \Phi_{N_\Phi}(\boldsymbol{\theta})] \quad (10)$$

It is well known that the optima of multiple general functions will not coincide except in very special cases (notably if the model is perfect or at least flexible enough to meet every objective – unlikely!). Instead, trade-offs arise between the degree to which individual objectives are optimized. Multi-objective optimization revolves around the concept of a “non-dominated” solution, which is a solution such that none of its corresponding objective function values can be improved without worsening at least one other objective. The Pareto front is defined as the set of non-dominated solutions.

Multi-objective optimization is a huge field of research in engineering, sciences, and mathematics; see Gupta et al. (1998) and the thorough review by Efstratiadis and Koutsoyiannis (2010) in the context of hydrological model calibration. Examples of multi-objective algorithms used in hydrology include MOSCEM (Vrugt et al. 2003), AMALGAM (Vrugt and Robinson 2007), and generalizations of the DDS algorithm (Asadzadeh and Tolson 2013). More generally, multi-objective optimization can be used to incorporate performance metrics other than those that quantify the model fit to observed data. For example, water resource model optimization may include economic objectives, pollution factors, and so forth. These setups may not be directly relevant to hydrological model calibration per se but are frequently used in the setup of management models where cost-benefit analysis is a major consideration (e.g., Marchi et al. 2014).

The ensemble of parameter sets comprising the Pareto front can be seen as representing parameter nonuniqueness associated with the (nonunique) choice of objective function. However, the interpretation of the Pareto front spread as parameter uncertainty is questionable. For example, the Pareto front does not, by itself, represent sources of uncertainty such as observation errors in the data, etc. Some advances along the direction of combining probabilistic and multi-objective techniques have been reported by Reichert and Schuwirth (2012) and warrant further investigation.

5 Probabilistic/Statistical Uncertainty Quantification

It is well known that hydrological modelling is affected by multiple sources of uncertainty, entering at every stage of the modelling process. For example, rainfall observations are subject to substantial sampling errors (e.g., McMillan et al. 2011), and streamflow observations are affected by rating curve errors (e.g., Westerbeg et al. 2010), not to mention the approximation of natural systems by mathematical models

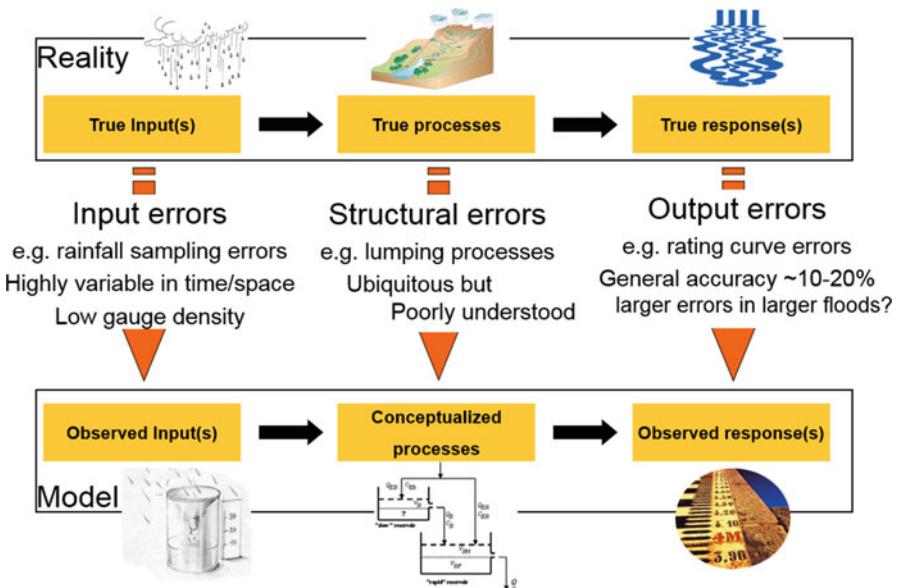


Fig. 4 Sources of uncertainty affecting parameter estimation in hydrological models

(e.g., Beven 2005; Renard et al. 2010). These sources of uncertainty are depicted schematically in Fig. 4.

Uncertainty is often classified as epistemic (i.e., due to model imperfections arising from incomplete knowledge of reality) versus aleatory (i.e., due to inherent randomness of the underlying phenomenon) – a distinction that is insightful yet not always clear-cut. As noted by Ang and Tang (2007), both types of uncertainty are tractable using probabilistic analysis, where uncertainty is described using probability theory. In this chapter, our primary focus is on the Bayesian paradigm, which provides a particularly appealing avenue for combining different sources of information.

5.1 Bayesian Inference: General Principles

Bayesian inference is a general class of probabilistic techniques, based on the premise that uncertainty in any quantity – including in model parameters – can be represented using random variables (probability distributions). Bayesian inference revolves around the posterior distribution of quantities of interest, $p(\boldsymbol{\theta} | \mathbf{D})$, which is given by Bayes equation:

$$p(\boldsymbol{\theta} | \mathbf{D}) = \frac{p(\mathbf{D} | \boldsymbol{\theta}) p(\boldsymbol{\theta})}{p(\mathbf{D})} \quad (11)$$

Bayesian inference requires two key conceptual ingredients: a prior $p(\boldsymbol{\theta})$ and a likelihood function $p(\mathbf{D}|\boldsymbol{\theta})$. The term $p(\mathbf{D})$ is independent from $\boldsymbol{\theta}$ and represents a normalizing constant, defined by the total probability integral $p(\mathbf{D}) = \int p(\mathbf{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$; while it is often a bear to compute, luckily this is not necessary in most modern Bayesian implementations (Sect. 5.3).

The prior distribution $p(\boldsymbol{\theta})$ is intuitively defined as what is known about parameters $\boldsymbol{\theta}$ *before* data \mathbf{D} has been observed. Prior information can come from multiple sources, including previous quantitative studies, qualitative expert judgment, and combinations of multiple such sources. For example, in flood frequency analysis, it is common to use regional information derived from previous studies in neighboring locations (e.g., Hailegeorgis and Alfredsen 2017). When developing rating curve models, priors may come from the analysis of hydraulic controls (Le Coz et al. 2014). In hydrologic models, it is common to use the admissible range of parameter values to specify flat (non-informative) priors. In some cases, when working with widely used models such as GR4J, it may be reasonable to specify the prior based on parameter values inferred in previous calibrations – either worldwide or in nearby or similar locations (Perrin et al. 2001).

The likelihood function $p(\mathbf{D}|\boldsymbol{\theta})$ represents, loosely speaking, the probability of observing the data \mathbf{D} given a particular set of model parameters $\boldsymbol{\theta}$. To obtain this, we need to specify a statistical model that may have “reasonably” generated the observed data. We are as free to specify this “reasonable” model just as hydrologists are free to specify their bucket model – guided by knowledge and intuition – and making practical judgments to simplify where appropriate. Section 5.2 will walk the interested reader through such a derivation. Section 6 will consider how to test calibration and model assumptions.

More formally, the likelihood function is the probability density function of the data-generating model, evaluated at the observed data, and viewed as a function of the model parameters. To emphasize its primary argument, the likelihood function is often written as $\mathcal{L}(\boldsymbol{\theta}; \mathbf{D})$. This notation is convenient for defining multiple likelihood functions depending on the specific assumptions made, e.g., as in Sects. 5.2 and 5.4.

The idea of Bayesian inference is that we start with a vague initial knowledge (the prior) and use the information contained in the data to refine this knowledge and obtain a (hopefully) sharper posterior. In this respect, posterior knowledge represents prior knowledge updated using the data. This principle is illustrated schematically in Fig. 5.

Note that in Bayesian methods, the outcome of the inference is the entire posterior distribution – in contrast to single-objective optimization, where the outcome of the inference is a single parameter set. That said, in practice, Bayesian posteriors are often summarized using properties such as the (posterior) mean, mode, or median, and their uncertainty (spread) is often summarized using the posterior covariance and so forth (Sect. 5.3).

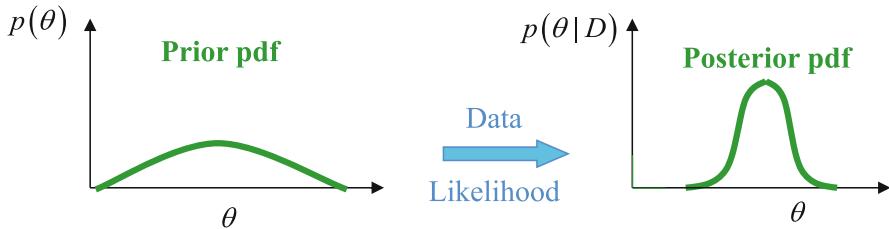


Fig. 5 Conceptual schematic of Bayesian inference. Parameter uncertainty is expressed using probability distributions. The combination of the prior and the likelihood yields the posterior, which represents the results of the inference

5.2 Least Squares Techniques as Gaussian Error Models

Bayesian inference often appears mysterious to modellers with a deterministic modelling background, especially when focusing solely on Eq. (11). Let us demystify Bayesian inference with a basic example.

Suppose we hypothesize that the original (deterministic) hydrological model provides a description of the observed data that is accurate on average but subject to random errors. In this case, the probabilistic model can be articulated as

$$\mathbf{Y} = \mathcal{H}(\tilde{\mathbf{x}}; \boldsymbol{\theta}) + \boldsymbol{\varepsilon} \quad (12)$$

where the term $\boldsymbol{\varepsilon}$ is the residual error, intended to represent the effect of all source of uncertainty contributing to differences between observed and modelled streamflow values.

Suppose these residual errors are independent and identically distributed (iid) Gaussian with zero mean and variance σ_e^2 :

$$\varepsilon_t \sim \mathcal{N}(0, \sigma_e^2) \quad (13)$$

Figure 6 illustrates this conceptualization. Equations (12 and 13) are referred to as the *error model* equations, as they provide a description of the errors affecting the model simulations and predictions.

The likelihood function $\mathcal{L}_{SLS}(\boldsymbol{\theta}, \sigma_e; \tilde{\mathbf{y}}, \tilde{\mathbf{x}})$ corresponding to this error model can be derived as

$$\begin{aligned} \mathcal{L}_{SLS}(\boldsymbol{\theta}, \sigma_e; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}) &= p(\tilde{\mathbf{y}} | \boldsymbol{\theta}, \tilde{\mathbf{x}}, \sigma_e) \underset{\text{indep}}{=} \prod_{t=1}^{N_{\tilde{\mathbf{y}}}} p(\tilde{y}_t | \boldsymbol{\theta}, \sigma_e, \tilde{\mathbf{x}}) \underset{\text{identic \& Gauss}}{=} \prod_{t=1}^{N_{\tilde{\mathbf{y}}}} f_{\mathcal{N}}(\tilde{y}_t; \mathcal{H}_t(\tilde{\mathbf{x}}; \boldsymbol{\theta}), \sigma_e^2) \end{aligned} \quad (14)$$

where $f_{\mathcal{N}}(y; \mu, \sigma^2)$ denotes the Gaussian probability density function (pdf). The annotation “indep” refers to simplifications arising from the independence assumption and “idetic & Gauss” to the identical Gaussian assumption.

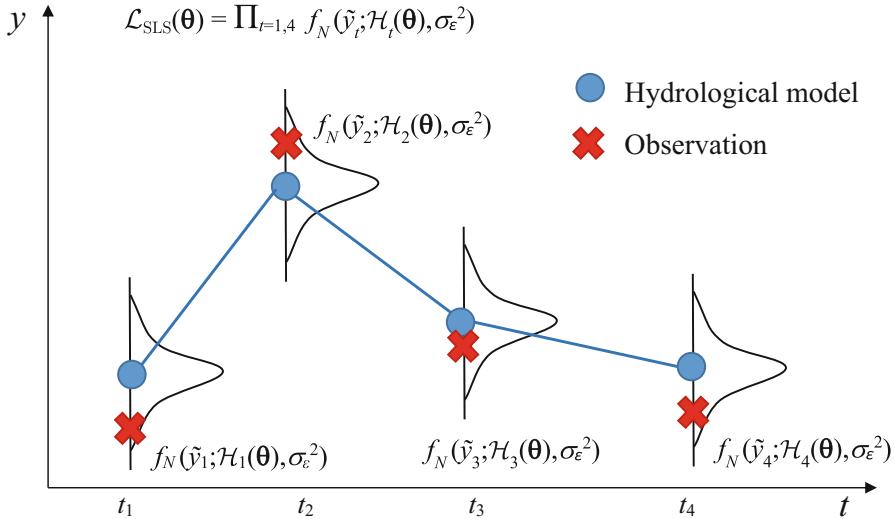


Fig. 6 Probabilistic model corresponding to Bayesian least squares inference. A Gaussian error model is assumed around each value simulated by the hydrological model, with the error variance being an error model parameter. The likelihood function is then defined as the (Gaussian) probability density of observation values of streamflow within this (joint) distribution, which in this case is the product of the (Gaussian) densities of individual observations

Equation (14) can be also articulated in terms of the residuals:

$$\mathcal{L}_{\text{SLS}}(\boldsymbol{\theta}, \sigma_\varepsilon; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}) = \prod_{t=1}^{N_{\tilde{\mathbf{y}}}} f_N(\tilde{y}_t - \mathcal{H}_t(\tilde{\mathbf{x}}; \boldsymbol{\theta}); 0, \sigma_\varepsilon^2) = \prod_{t=1}^{N_{\tilde{\mathbf{y}}}} f_N(\varepsilon_t[\tilde{\mathbf{y}}; \tilde{\mathbf{x}}; \boldsymbol{\theta}]; 0, \sigma_\varepsilon^2) \quad (15)$$

$$\varepsilon_t = \tilde{y}_t - \mathcal{H}_t(\tilde{\mathbf{x}}; \boldsymbol{\theta}) \quad (16)$$

Note that the change of variables from \mathbf{Y} to $\boldsymbol{\varepsilon}$ in the pdfs given by Eqs. (14) and (15) requires a Jacobian term $\partial \boldsymbol{\varepsilon} / \partial \mathbf{y}|_{\mathbf{y}=\tilde{\mathbf{y}}}$; however, this term is unity in view of Eq. (16) and is hence omitted.

If we substitute the Gaussian pdf expression into Eq. (15) and take logs, we get

$$\log \mathcal{L}_{\text{SLS}}(\boldsymbol{\theta}, \sigma_\varepsilon; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}) = -\frac{N_{\tilde{\mathbf{y}}}}{2} \log 2\pi - N_{\tilde{\mathbf{y}}} \log \sigma_\varepsilon - \frac{1}{2\sigma_\varepsilon^2} \Phi_{\text{SSE}}(\boldsymbol{\theta}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \quad (17)$$

which establishes the close correspondence of probabilistic inference under Gaussian assumptions with the SSE objective function in Eq. (4). For example, it can be readily shown that the parameter set $\boldsymbol{\theta}^{(\text{SLS})}$ that maximizes the (log-) likelihood function in Eq. (17) also minimizes the SSE objective function. This equivalence holds whether the error variance σ_ε^2 is inferred or assumed known. For this reason, we can refer to Eqs. (14, 15, 16, and 17) as Bayesian least squares inference. The

derivations needed to arrive at these equations put the choice of error norm in Eq. (4) on a more defined theoretical basis and in doing so highlight its implicit assumptions (here, iid Gaussian errors).

If the primary interest is in the model parameters $\boldsymbol{\theta}$, the error parameter σ_e can be integrated out (Kavetski et al. 2006a):

$$\log \mathcal{L}_{SLS2}(\boldsymbol{\theta}; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}) = \log \int \mathcal{L}_{SLS}(\boldsymbol{\theta}, \sigma_e; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}) d\sigma_e \propto \frac{N_{\tilde{\mathbf{y}}} + 2}{2} \log \Phi_{SSE}(\boldsymbol{\theta}; \tilde{\mathbf{x}}, \tilde{\mathbf{y}}) \quad (18)$$

which still retains the SSE kernel within the (now lower-dimensional) likelihood function. A similar expression holds for certain non-rectangular priors on σ_e .

5.3 Tools for Analyzing Bayesian Posteriors

The formulation of the posterior is only the first step of the inference. The equation defining the posterior distribution in effect plays the role of the “objective function” in Bayesian estimation. The same intuitive ideas established in Sect. 2.5 for interpreting the shape of the objective function apply to the posterior distribution – near-optimal regions suggest the most likely values of the parameters, the spread of the distribution is indicative of posterior parameter uncertainty, and elongated contours sloping with respect to the parameter axes indicate parameter correlations/dependencies. Once we have derived the functional form of the posterior, how do we use it to get a parameter value and its uncertainty?

In some simple cases, where the likelihood and prior are given by “conjugate” distributions – which for the most part are common textbook distributions such as Gaussian, Gamma, and so forth – the posterior will itself come out as a known distribution, with parameters derived from the parameters of the prior and likelihood (Box and Tiao 1992). However, this simplification is seldom possible with most hydrological models.

In practice, Bayesian posteriors are either summarized by their estimated mode (optimum) and covariance or explored wholesale using Markov Chain Monte Carlo (MCMC) algorithms.

The posterior mode can be found using optimization as $\hat{\boldsymbol{\theta}} = \text{mode}[\boldsymbol{\theta}] = \underset{\boldsymbol{\theta}}{\text{argmax}} p(\boldsymbol{\theta} | \tilde{\mathbf{y}})$, i.e., directly treating the Bayesian posterior as an objective function (cf Sect. 5.2). The posterior covariance can then be approximated as $\text{cov}[\boldsymbol{\theta}] \approx -\mathbf{H}_{\boldsymbol{\theta}}^{-1} [\log p(\boldsymbol{\theta} | \tilde{\mathbf{y}})]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}$ where $\mathbf{H}_{\boldsymbol{\theta}} = \partial^2 / \partial \boldsymbol{\theta}^2$ is the Hessian (second derivative) matrix operator (Gelman et al. 1998), which can itself be approximated using finite differences. The Hessian matrix reflects the curvature of the posterior distribution (objective function) – peaky posteriors have large negative curvature and hence represent little posterior uncertainty, whereas flat posteriors have near-zero curvature and hence represent large uncertainty.

MCMC sampling provides an alternative – and very powerful – numerical technique for uncertainty quantification. A challenge of Bayesian posteriors is that they seldom take the form of common distributions. MCMC is a numerical sampling technique that (asymptotically) produces samples from the (possibly un-normalized) probability density function it is applied to (Gelman et al. 1998). It is most relevant to (i) nonstandard distributions for which off-the-shelf samplers are not available, (ii) high-dimensional distributions, and (iii) distributions with pdf known only up to a constant of proportionality. These are precisely the features of most Bayesian posteriors, making MCMC a Bayesian’s best friend! MCMC algorithms used in hydrology include multistage implementations of the classic Metropolis algorithm (Thyer et al. 2009), methods based on differential evolution (Vrugt et al. 2009a), and many others.

It is worth noting the important distinction between the choice of likelihood function and prior versus the choice of tools used to compute and analyze the posterior distribution. Just as the use of an optimization algorithm such as SCE to find the posterior mode does not make SCE into a “Bayesian algorithm” neither does the routine use of MCMC to sample from posterior distributions make MCMC into a “Bayesian algorithm.” MCMC itself is *not* a Bayesian technique – it is a general numerical method for sampling from any probability distribution. Provided the MCMC algorithm is at all convergent, its results are determined by the function it is applied to, not the MCMC algorithm itself. For this reason, statements such as “we calibrated our model using MCMC” are about as informative (while still technically correct) as “we calibrated our model using MATLAB” – what should be reported first and foremost are the equations and assumptions defining the posterior distribution. A poorly chosen optimizer or MCMC sampler will surely degrade even the best posed inference, but algorithmic sophistication can hardly rescue a calibration from a poorly chosen objective function.

The next sections describe two distinct strategies for Bayesian inference, namely, aggregational and decompositional approaches, which are distinguished by the way they attempt to represent uncertainty.

5.4 Aggregational Methods

Aggregational approaches attempt to describe all sources of error using a single term. The simple least squares technique from Sect. 5.2 represents the prototypical implementation of this idea. However, its iid assumptions are questionable (e.g., Sorooshian and Dracup 1980). For example, errors of hydrological models typically exhibit heteroscedasticity, meaning larger errors in larger flows, which invalidates the assumption of identical distribution. In addition, errors typically exhibit persistence, meaning multiple consecutive errors of similar sign and magnitude, which invalidates the independence

assumption. These statistical features can and should be reflected in the likelihood function.

Heteroscedasticity can be dealt with using weighted least squares and transformed least squares, in which case the assumption of identical distribution is applied to “normalized” residuals η , rather than to raw residuals ϵ :

$$\eta_t \sim \mathcal{N}\left(0, \sigma_\eta^2\right) \quad (19)$$

The use of weights (weighted least squares) corresponds to defining the normalized residuals as

$$\eta_t(\tilde{\mathbf{y}}; \boldsymbol{\theta}, \tilde{\mathbf{x}}) = \frac{\tilde{y}_t - \mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}]}{\sigma_{\epsilon(t)}} \quad (20)$$

which effectively gives some data points more influence than others. Under the heteroscedastic assumption that large flows have larger errors, data points corresponding to peak flows should receive *reduced* weight, so it is more accurate to say the influence is being “balanced.”

If the residuals are assumed to represent all sources of error, their statistical properties are generally unknown, and additional assumptions are required, e.g.,

$$\sigma_{\epsilon(t)} = a + b \mathcal{H}_t \quad (21)$$

where a and b are unknown parameters inferred along with the hydrological model parameters (Evin et al. 2014). Alternatively, the weights could be specified a priori, e.g., as in the PEST package (Doherty 2005). In other words, probabilistic inference often introduces parameters *in addition* to those of the hydrological model itself.

The use of response transformations (possibly with their own parameters $\boldsymbol{\theta}_Z$) represents an alternative strategy, where

$$\eta_t(\tilde{\mathbf{y}}, \boldsymbol{\theta}, \tilde{\mathbf{x}}; \boldsymbol{\theta}_Z) = Z(\tilde{y}_t; \boldsymbol{\theta}_Z) - Z(\mathcal{H}_t[\tilde{\mathbf{x}}; \boldsymbol{\theta}]; \boldsymbol{\theta}_Z) \quad (22)$$

A common choice of transformation Z is the Box-Cox transformation:

$$Z(y; \lambda) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log y & \text{if } \lambda = 0 \end{cases} \quad (23)$$

which includes as special cases the logarithmic transformation ($\lambda = 0$), the inverse transformation ($\lambda = -1$), the square-root transformation ($\lambda = 0.5$), and, trivially, the null transformation ($\lambda = 1$).

Despite superficial differences, weighting and transformational strategies are closely related. Consider the likelihood function formulated in terms of normalized residuals η :

$$\begin{aligned} \log \mathcal{L}_H(\boldsymbol{\theta}, \sigma_\eta; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}, \boldsymbol{\theta}_Z) &= p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \tilde{\mathbf{x}}, \sigma_\eta, \boldsymbol{\theta}_Z) \\ &= \frac{\partial \mathbf{\eta}}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\tilde{\mathbf{y}}} \times f_N(\mathbf{\eta}(\tilde{\mathbf{y}}; \boldsymbol{\theta}, \boldsymbol{\theta}_Z, \tilde{\mathbf{x}}); \boldsymbol{\theta}, \sigma_\eta^2 \mathbf{I}) \end{aligned} \quad (24)$$

where the Jacobian term $\partial \mathbf{\eta} / \partial \mathbf{y}|_{\mathbf{y}=\tilde{\mathbf{y}}}$ accounts for the change of variables from \mathbf{y} to $\mathbf{\eta}$ and \mathbf{I} is the identity matrix. Taylor series can be used to establish a first-order equivalence of linear weights in Eq. (21) and the log transformation given by Eq. (23) with $\lambda = 0$ (McInerney et al. 2017). This equivalence holds in terms of variances, but there are important differences in terms of skew and kurtosis that can impact practical performance (e.g., Schoups and Vrugt 2010; McInerney et al. 2017).

Persistence can be dealt with by incorporating autoregressive terms, e.g., the simplest AR(1) assumption yields

$$\eta_t = \phi \eta_{t-1} + W_t \quad (25)$$

$$W_t \sim N(0, \sigma_w^2) \quad (26)$$

If the AR(1) assumption is applied to residuals after the Box-Cox transformation, the likelihood function is

$$\begin{aligned} \log \mathcal{L}_H(\boldsymbol{\theta}, \phi, \sigma_w; \tilde{\mathbf{y}}, \tilde{\mathbf{x}}, \lambda) &= p(\tilde{\mathbf{y}}|\boldsymbol{\theta}, \tilde{\mathbf{x}}, \phi, \sigma_w, \lambda) \\ &= \frac{\partial \mathbf{w}}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\tilde{\mathbf{y}}} \times f_N(\mathbf{w}(\tilde{\mathbf{y}}; \boldsymbol{\theta}, \tilde{\mathbf{x}}, \phi, \lambda); \boldsymbol{\theta}, \sigma_w^2 \mathbf{I}) \end{aligned} \quad (27)$$

where we allow for the autocorrelation parameter ϕ to be inferred along with the error variance σ_w^2 while keeping the BC transformation parameter λ fixed.

The practicalities of representing heteroscedasticity and persistence within the likelihood function are often subtle and have a major impact on parameter estimation. First, the order of treatment is important – it is best to start by stabilizing the residual variance using a transformation (or weighting) and then treating persistence (Evin et al. 2013). Second, the parameters of the error models, notably the error variance and autocorrelation, can be inferred either jointly with the hydrological parameters or in a separate post-processing step. Evin et al. (2014) considered a post-processing approach that first estimated the hydrological parameters under the assumption of no persistence and then separately estimated the error variance and autocorrelation. Although the joint approach is a more pure application of the Bayesian paradigm, it can suffer from multi-way interactions between the mass-balance parameters, the autocorrelation coefficient, and the error variance, which lead to poor quality predictions; the post-processing approach appears more robust, as seen in Fig. 7 adopted from Evin et al. (2014). Third, in terms of transformation parameter, values of the Box-Cox λ in the range 0–0.5 appear to provide the best empirical performance (McInerney et al. 2017), with trade-offs arising between the reliability, precision, and bias of the resulting predictions.

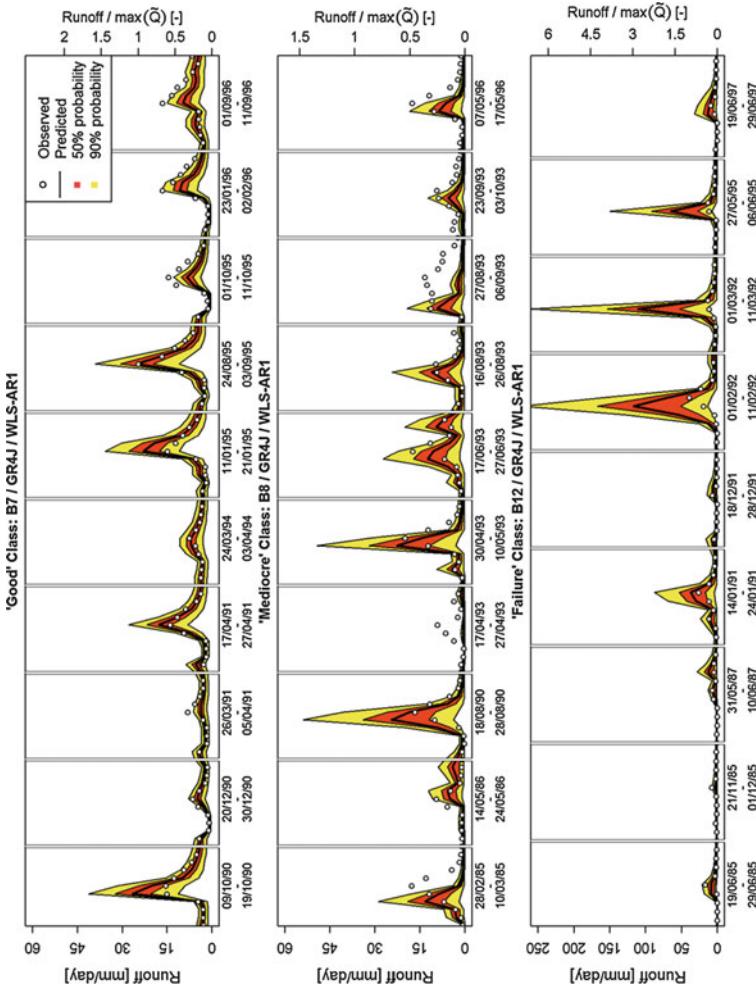


Fig. 7 Representative hydrological predictions obtained in the case studies of Evin et al. (2014), where an aggregational approach was employed with all sources of error represented using a single residual error term. The likelihood function was formulated accounting for heteroscedasticity and persistence. Three classes of results are presented, ranging from “good” to “mediocre” and “failure” categories, based on predictive performance of the parameter estimation approach. The notation B7, B8 and B12 refers to the French Broad, English River and San Marcos catchments in the USA. Reproduced with permission from John Wiley and Sons

Many aspects of residual error modelling are of interest beyond Eqs. (19, 20, 21, 22, 23, 24, 25, 26, and 27). For example, Gaussian assumptions can be replaced with more general skewed power exponential (SEP) distribution, which allows for skewness and kurtosis parameters (Schoups and Vrugt 2010), and AR(1) assumptions can be replaced with more general autoregressive models (e.g., Morawietz et al. 2011). The treatment of zero and near-zero flows has been investigated using approaches such as censoring (Wang and Robertson 2011) and “zero-flow inflation” (Smith et al. 2010). Residual error models based on mixtures (Schaeffli et al. 2007) and conditioned on covariates other than streamflow (e.g., Pianosi and Raso 2012) have also been investigated.

Aggregated strategies are well suited to operational applications, as the resulting inference can produce reliable and precise predictions at a low-moderate cost in terms of algorithm complexity and computational effort. In other words, aggregated strategies allow the modeller to focus on the pragmatic goal of overall predictive uncertainty quantification (e.g., Krzysztofowicz 1999; Lerat et al. 2015). The next section details decompositional approaches – which are more ambitious in their objectives yet are also harder and more expensive to implement.

5.5 Decompositional Methods

Decompositional approaches attempt to explicitly disentangle the contributions of individual sources of error, such as those seen in Fig. 4 (e.g., Krzysztofowicz 1999; Kavetski et al. 2002; Seo et al. 2006; Huard and Mailhot 2008; Vrugt et al. 2008; Reichert and Mieleitner 2009). An example of the decompositional approach in hydrological modelling is given by the Bayesian total error analysis (BATEA) (Kavetski et al. 2006a; Kuczera et al. 2006). Decompositional approaches require more data and are more complex than aggregated approaches. Notably, Renard et al. (2010) established that the decomposition is inherently ill-posed in the absence of (approximate) prior knowledge; as noted by Beven (2005), the problem can be conceptualized as inferring individual error terms $\boldsymbol{\varepsilon}_x$, $\boldsymbol{\varepsilon}_y$, and $\boldsymbol{\varepsilon}_{\mathcal{H}}$ associated with input data errors, output data errors, and model structural errors, respectively,

$$\boldsymbol{\varepsilon} = \boldsymbol{\varepsilon}_x + \boldsymbol{\varepsilon}_y + \boldsymbol{\varepsilon}_{\mathcal{H}} \quad (28)$$

which is clearly ill-posed without at least some information about at least two of the three error terms.

Renard et al. (2011) tackled this challenge in a case study of the Yzeron catchment (France), where a dense rain gauge network, R13H, was available over a 2-year period. Data from R13H was exploited using conditional simulation (Tompson et al. 1989) to develop a prior error model for a sparse rain gauge network, R3D, active over a much longer time period. The priors for parameters describing streamflow observation errors were obtained using rating curve error analysis (Thyer et al. 2009). The representation of model structural error is another major challenge in a decompositional approach. Unlike data errors, which in

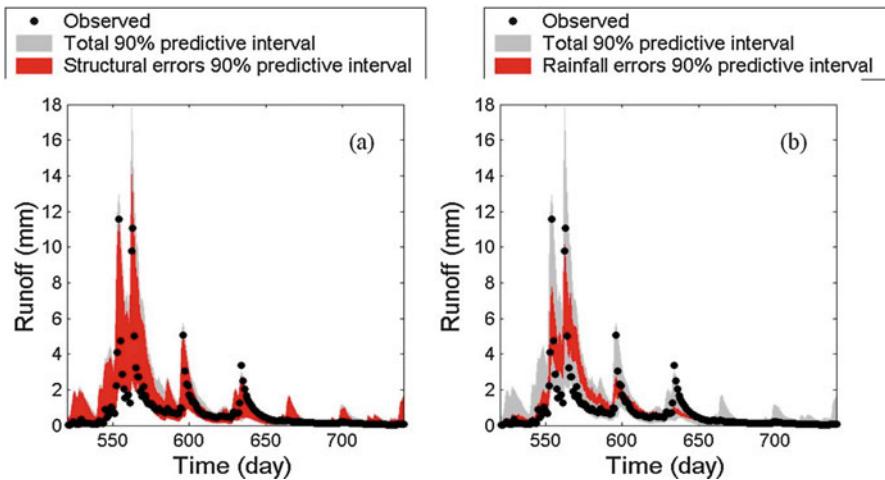


Fig. 8 Insights from the application of the BATEA decompositional parameter estimation approach in the Yzeron catchment, France (Renard et al. 2011). In this case study, structural errors of the GR4J model appear to dominate streamflow uncertainty due to the effects of rainfall errors. Figure reproduced from Renard et al. (2009)

principle can be estimated by comparison to a more accurate data set, there is no analogous concept for model structural errors – model comparison experiments do not provide much evidence of a particular model being consistently more accurate than others (e.g., Duan et al. 2006). Renard et al. (2011) treated model structural errors as what's left behind after the other errors have been characterized. At the cost of setting them up, decompositional approaches can offer fascinating insights into the relative contributions of different sources of error to total predictive uncertainty. For example, when using GR4J and the R3D data set to predict streamflow in the Yzeron catchment, structural errors dominated rainfall-induced errors, as shown in Fig. 8 adopted from Renard et al. (2011). These insights can guide efforts to improve the predictions, e.g., in this instance by prioritizing improvements to model structure over looking for more accurate data sets (a statement specific to this case study and not intended to detract from the general importance of data to hydrology!).

5.6 Methods Other than Bayesian and Other than Probabilistic

The presentation of probabilistic inference thus far has focused primarily on Bayesian methods, which tend to be commonly used for hydrological models. This section provides a brief summary of other related techniques.

Frequentist techniques are common in statistical estimation, particularly in flood frequency analysis. The method of moments (MoM) can be used whenever the

parameters of a distribution can be related to the moments of the calibration data (e.g., Salas 1993). Maximum likelihood (ML) methods (e.g., Martins and Stedinger 2000) are broadly similar to Bayesian methods (barring some philosophical differences in the interpretation of probability) but (generally) do not allow for a prior distribution. Procedurally, MoM and ML proceed by first estimating the optimal parameters and then estimating parametric uncertainty – as opposed to Bayesian inference where the modeller must first derive the posterior distribution and then summarize/use it. In many rainfall-runoff model applications, there is little difference between frequentist and Bayesian approaches, as the influence of the prior is close to negligible for typical lengths of data. However, it is less clear how to develop a decompositional approach, e.g., analogous to BATEA but without using priors to handle the disaggregation exemplified by Eq. (28). Priors can also be valuable in data-sparse contexts.

The generalized likelihood uncertainty estimation (GLUE) (Beven and Binley 1992) is an estimation technique common in conceptual hydrological modelling. GLUE is often described as an “informal” technique, in the sense that it does not seek to construct probabilistic descriptions of uncertainty such as the error models in Sects. 5.2 and 5.4. For example, many GLUE publications have used the Nash-Sutcliffe efficiency *as if* it were a likelihood function and have relied solely on parametric uncertainty (without a residual error model) when generating prediction limits (e.g., Freer et al. 1996). The motivation for GLUE has evolved since its original development and has more recently focused on the challenges of describing epistemic uncertainty using probability theory. The solutions suggested by GLUE have elicited much debate, from the role of parameters and parametric uncertainty in modelling to whether prediction limits should satisfy probabilistic criteria such as enveloping a prescribed proportion of observations (e.g., Mantovan and Todini 2006; Stedinger et al. 2008; Beven 2006; Beven et al. 2012; Clark et al. 2012). Despite many divergent perspectives, there is also important commonality. The concept of equifinality, central to the original motivation for GLUE (Beven 2006), corresponds broadly to non-identifiability and ill-posedness (Sects. 2.1 and 5.3). More recently, the GLUE community has drawn attention to “disinformative data” (Beven and Westerberg 2011), which in the context of probabilistic techniques represents data that violates the assumed error models. Some work has attempted to bridge the gap between GLUE and Bayesian methods (e.g., Vrugt et al. 2009b; Nott et al. 2012; Kavetski et al. 2018).

Ultimately, conceptual and algorithmic similarities will necessarily arise between all calibration approaches that work with (optimize and/or sample) functions of the form $\mathcal{L}(\boldsymbol{\theta}; \tilde{\mathbf{y}})p(\boldsymbol{\theta})$. From this perspective, genuine differences can only arise from the way $\mathcal{L}(\boldsymbol{\theta}; \tilde{\mathbf{y}})$ and $p(\boldsymbol{\theta})$ are constructed: the use of probability theory leads to Bayesian (and maximum likelihood) methods and probabilistic prediction, whereas the use of other principles, such as fuzzy set theory (Freer et al. 2004) and others, leads to correspondingly different interpretations (Smith et al. 2008). For this reason, a modeller that wishes to obtain prediction limits that have a probabilistic interpretation is best advised to use the tools of probability theory.

But the extent to which a probabilistic interpretation is a desirable attribute of an estimate or prediction is a much deeper question. This chapter takes the perspective that probability theory is a suitable – indeed advantageous – platform for describing the data and model uncertainties of relevance to scientific and engineering applications (Ang and Tang 2007); the interested reader is directed to discussions in the broader scientific community (e.g., de Finetti 1964; Oberkampf et al. 2004; O'Hagan and Oakley O'Hagan and Oakley 2004; Reichert et al. 2015, among others).

6 Model Diagnostics as Part of Parameter Estimation

The specific task of parameter estimation cannot be seen in isolation from the broader modelling process of hypothesis testing, refinement, and prediction. As vividly seen in Sects. 2, 3, 4, and 5, parameter estimation is necessarily based on assumptions, such as the applicability of data for a priori estimation, the selection of an objective function for optimization or the selection of an error model for probabilistic inference. In order to gain confidence that parameter estimation has been successful, these assumptions should be scrutinized and, if necessary, replaced. In the absence of such checks, there is little guarantee that, as eloquently noted by Kirchner (2006), the calibrated models are not merely dancing like mathematical marionettes to the tune of the calibration data. The topic of posterior diagnostics is extensive in its own right; this chapter highlights some of the key principles and practical implementations but does not attempt to be truly comprehensive.

We first consider what kind of model fit can be expected after calibration and how uncertainty limits behave. To this end, Fig. 9 shows the results of a GR4J model calibration. For didactic reasons, synthetic data with Gaussian errors was used, so that the inference assumptions are met by construction. Should parametric uncertainty encompass the observations? In Fig. 9b, they clearly do not – does this mean something is wrong? To answer this question, consider the error model underlying the inference – Eqs. (12 and 13) clearly assume that differences between the model simulations and the observed values are explained by an additive noise term. Therefore, if we want to construct probability limits that encompass the data, we need to account for the residual errors. Adding this term now produces the expected results, as seen in Fig. 9c. Another questionable aspect is the “white noisiness” of predictive limits seen in Fig. 9c – this is a consequence of the independence assumption in the residual error model. Figure 9d illustrates, for a different synthetic dataset, the much smoother and realistic (for a streamflow prediction) behavior of the autocorrelated error model in Eq. (25). A different perspective is taken in an approach such as GLUE, where no probabilistic error model is used (Sect. 5.6). In this case, parametric uncertainty is used to describe all sources of error, which in turn requires a much flatter (“lenient”) pseudo-likelihood function. Hence, depending on the assumptions made by the estimation framework, different behaviors may be expected, and diagnostics must be crafted accordingly.

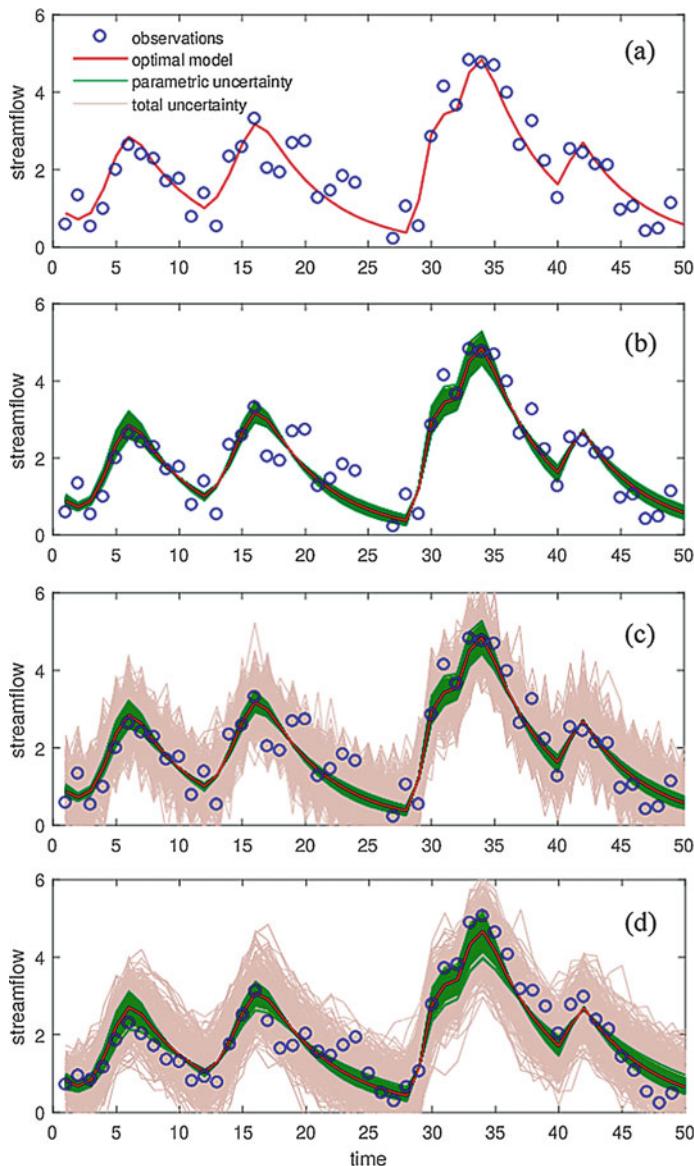


Fig. 9 Optimization, parametric uncertainty, and total predictive uncertainty in a simple Bayesian least squares framework. Panel *a* shows the optimal predictions; panel *b* shows the parametric uncertainty. Under Gaussian error model assumption, discrepancies between observed and simulated streamflows are described by the (Gaussian) residual error. Hence, residual error uncertainty, shown in panel *c*, must be added in order for the predictive limits to envelop the observed data. With reference to Fig. 6, panel *c* represents the use of the probability model to generate the predictive distribution and reconcile it against the actual observations. Panel *d* illustrates (using a different data set) the much “smoother” behavior of autocorrelated error models, which reduce the unrealistic jitter in hydrograph replicates generated using white noise error models

The individual assumptions underlying error models must also be tested. For example, if the Box-Cox Gaussian AR(1) error model is used, the modeller should test that (i) normalized residuals are approximately homoscedastic (e.g., no dependence on magnitude of simulated response), (ii) innovations are approximately independent (e.g., using PACF plots), and (iii) innovations are approximately Gaussian (e.g., using histograms and Gaussian QQ plots) (e.g., Thyer et al. 2009; Schoups and Vrugt 2010; Morawietz et al. 2011; McInerney et al. 2017). Model diagnostics represent a form of hypothesis testing and are most effective when applied in a structured and systematic way (Clark et al. 2011). For example, diagnostics are most inquisitive and informative when applied to stratified data, e.g., by flow magnitude, season, etc., as this can yield more insights into individual model deficiencies without being masked by averaging effects, etc.

What period should these diagnostics be applied to? Testing over the calibration period is generally weak. Arguably, one cannot claim a model to be “predicting” a quantity (e.g., data period) already explicitly used in its calibration – at best, the model is “simulating” or “reproducing” data already known to it. This perspective leads to the concept of a “validation” period, e.g., split-sample calibration where the available data is split into calibration and validation periods (Kuczera and Franks 2002), and more complex cross-validation setups (Tuteja et al. 2017). Validation mimics the way the model will be used in practice and arguably offers the best chance to detect deficiencies in the model and calibration. However, testing on a new period can create genuine complications in the case of non-stationarities (e.g., land-use change and/or climate variability) (Westra et al. 2012) – which highlights the formidable challenge of environmental prediction. Even the very semantics of the term “validation” have been questioned (e.g., Konikow and Bredehoeft 1992), on the grounds that it can provide a misleading impression of the model’s abilities to make predictions.

Testing on validation periods can detect instances of over-parameterization, where a model has been over-fitted to spurious features in the calibration data and performs poorly on new data. A simple example is the fitting of a high-degree polynomial to a few data points. While any model can in principle be over-fitted, complex highly parameterized models are more susceptible, notably models based on neural networks (Kingston et al. 2008), but also physically based distributed models calibrated solely to catchment-average rainfall and runoff (e.g., Grayson et al. 1992; Jakeman and Hornberger 1993). Parameter estimation – whether via calibration or a priori estimation – is hence inevitably an exercise in balancing model complexity with available data (e.g., Fenicia et al. 2008).

Finally, we note that diagnostics are generally based on comparing simulated and observed responses and relating any deficiencies to the parameter estimates. In this respect, response diagnostics work with “tangible” quantities (to an extent) and seek to draw conclusions about parameters, which ultimately are “intangible” quantities. Mantovan et al. (2007) go as far as to refer to parameters as “abstract devices,” which is of course far from ideal from the perspective of physical interpretability of models. Parameters, ultimately, do hold insights into systems, e.g., residence time, etc. (Fenicia et al. 2010). Nevertheless, it should be clear that

parameter estimation is not as dependable as the prediction of observable quantities – while it may be reasonable to rely on validated prediction of quantities such as streamflow, relying on the corresponding parameter values requires a longer leap of faith!

7 Practicalities

This section lists some empirical “tricks” to supplement the theory of parameter estimation in hydrology.

7.1 Parameter Transformations

Parameter transformations often improve numerical algorithm performance when working with highly non-quadratic objective functions. Figure 10 shows an example from hydrological modelling, where a “banana-shaped” objective function becomes much better behaved when expressed in log-transformed parameter space. In some estimation problems, transformations are simply essential (Thyer et al. 2002). Parameters of hydrological models often benefit from log transformations, especially when appearing in exponents and multiplicative factors. Note the fundamental distinction between transforming parameters and responses: the former is purely a numerical device to improve the shape of the objective function “as seen” by an analysis method, whereas the latter yields a genuinely different objective function.

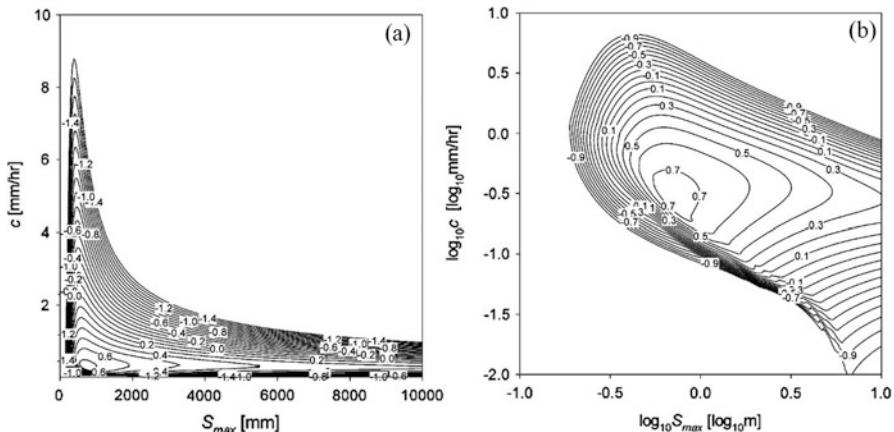


Fig. 10 Illustration of parameter transformation to improve the conditioning of the objective function (Kavetski et al. 2006c). Panel *a* shows a least squares objective function exhibiting a banana-type shape. Panel *b* shows the same objective function plotted in log-transformed parameter space, exhibiting a much better-behaved near-quadratic shape. Reproduced with permission from Elsevier

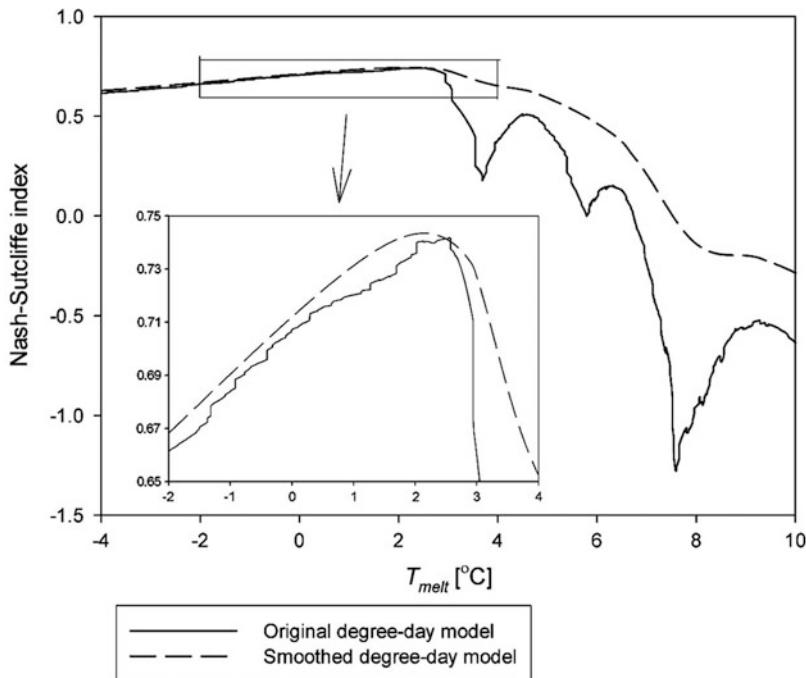


Fig. 11 Effect of non-smooth models on the objective function, illustrated using results from the case study of Kavetski et al. (2006b). Both macro- and microscale irregularities are visible. Model smoothing removes these artefacts and yields a remarkably well-behaved surface. Reproduced with permission from Elsevier

7.2 Impact of Model Non-smoothness/Discontinuities

Section 3 alluded to the difficulties posed by non-smooth models. The effect can be dramatic, especially on optimization algorithms that rely on gradients to establish the search direction. Figure 11 shows an example of a discontinuous objective function, due to internal model thresholds. When working with non-smooth models, hydrologists have two options: either use a more robust but slower algorithm or modify the model to remove the thresholds (Kavetski and Kuczera 2007). As seen in Fig. 11, model smoothing can successfully remove discontinuities from the objective function and substantially simplify the estimation process. Interestingly, previous work on smoothing often reported an improvement in model performance, which suggests that, at least on the large scale, environmental dynamics are not as threshold-driven as the models themselves (Kavetski and Clark 2010). For this reason, Hill et al. (2015) recommend a more concerted effort by environmental modellers to use robust numerics and smooth constitutive functions; this philosophy has been adopted in modelling frameworks such as FUSE (Clark et al. 2008), SUPERFLEX (Fenia et al. 2011), SUMMA (Clark et al. 2015), and RAVEN (Craig et al. 2017).

7.3 Initial Conditions: Estimate or Warm-Up

Most hydrological models are dynamic in time, and initial state values must be specified before the model can be deployed. In practice, initial model states are unknown. Three options can be considered:

- (i) Use a warm-up period after setting the initial conditions to some arbitrary values (e.g., 50% full). Warm-up periods are easy to implement but can chew up a lot of data in slow-responding catchments. The warm-up period is sufficiently long if the objective function (and hence calibrated parameters) exhibits little sensitivity to the (arbitrary) initial states.
- (ii) Estimate initial conditions along with the model parameters. This approach does not waste data but can distort parameter values by favoring the fitting of the initial data period, where the fitted initial conditions in effect provide another degree of freedom.
- (iii) Estimate initial conditions using other considerations, e.g., by solving the model equations for the steady-state storage values. This approach avoids the limitations of approaches (i) and (ii) but has the drawback that the theoretical steady state may not be representative of actual catchment conditions at the beginning of the calibration period. This approach is best used to inform the selection of initial values to shorten (but not avoid) the warm-up period.

7.4 Estimation of Expensive Models

Distributed models, such as MODFLOW (Harbaugh 2005), SWAT (Arnold and Fohrer 2005), and MHM (Samaniego et al. 2010), are increasingly used in environmental work to generate distributed predictions. These benefits accrue at substantial computational costs, with model runs taking as long as minutes or even hours per simulation. Computational costs inevitably impose restrictions on parameter estimation. For example, multi-start optimization and MCMC sampling may be severely limited or perhaps precluded altogether. When working with expensive models, it may be necessary to set a computational budget for analyses such as optimization, e.g., as implemented in the DDS optimizer (Tolson and Shoemaker 2007); parallel computing offers a pragmatic way to reduce wall-clock runtimes. Highly parameterized models can be handled using parameter regularization (Doherty 2003), model emulation (e.g., Albert 2012; Laloy et al. 2013), multi-scale parameter estimation techniques (Samaniego et al. 2010), and/or multistage estimation approaches where parameters are calibrated step-by-step rather than all at once (Fenia et al. 2016).

8 Research Directions

Parameter estimation is a huge field with many unresolved challenges. This section lists some (not all) research directions of interest from practical and scientific perspectives.

8.1 Operational Improvements

Environmental and water agencies are increasingly interested in tackling challenging streamflow forecasting problems, including temporally consistent (“seamless”) predictions over seasonal (3 months) lead times (e.g., Tuteja et al. 2011). Achieving these outcomes requires robust treatment of error persistence (e.g., Evin et al. 2014), balancing predictive reliability and precision (McInerney et al. 2017), and finding hydrological model parameters that perform well at multiple spatial and temporal scales (e.g., Samaniego et al. 2010). Similar requirements hold when seeking spatially coherent predictions over large catchment systems and river networks. In addition, since forecasting models are often calibrated using observed input-output data (e.g., rainfall-streamflow) but produce response forecasts (e.g., streamflow 3 months ahead) using forecasted forcings (e.g., rainfall from a numerical weather prediction model), better integration of multiple models representing individual sources of uncertainty is also of interest – this is one of the operational motivations for decompositional approaches (Sect. 5.5).

8.2 Sparse-Data Problems

Many locations around the globe are poorly gauged or ungauged. For example, in Australia, as much as 90–95% of the subcatchments of the Murray-Darling basin are ungauged (e.g., Chiew and Siriwardena 2005). Modelling these locations requires estimating model parameters from a combination of local properties (if available) and extrapolation from “similar” catchments. This was the theme of the Prediction in Ungauged Basins (PUB) decade (Sivapalan et al. 2003b). Indirect calibration approaches include non-concomitant calibration, where input and output data from different time periods are utilized. Spectral methods (e.g., Montanari and Toth 2007; De Vleeschouwer and Pauwels 2013; Schaefli and Kavetski 2017) and signature calibration (e.g., Yilmaz et al. 2008; Shafii and Tolson 2015; Westerberg and McMillan 2015; Fenicia et al. 2018) are of interest; e.g., signatures computed from simulated responses in Period A can be compared to the corresponding signatures of observed data in Period B. Computationally, signature calibration and uncertainty quantification can be approached using the fascinating class of methods known as approximate Bayesian computation (e.g., Nott et al. 2012; Vrugt and Sadegh 2013; Kavetski et al. 2018) – these methods avoid the need to derive the likelihood function in closed form and instead require sampling from the underlying probability

model (Toni et al. 2009). Estimation in ungauged basins, where no data is available for model calibration, is being investigated using various parameter regionalization approaches (e.g., Bulygina and Gupta 2009; Hrachowitz et al. 2013).

8.3 Recursive Estimation and Data Assimilation

In many cases, calibration data is not available all at once and/or arrives in real time. For example, in applications such as flood forecasting, exploiting real-time information such as local observations and/or satellite imagery is of interest (e.g., Neal et al. 2007; Reichle 2008; Hostache et al. 2010; Giustarini et al. 2016; Revilla-Romero et al. 2016). In the context of parameter estimation, one can then distinguish between batch estimation (where the entire data set is used at once) and recursive estimation (where data is ingested sequentially, e.g., one data point at a time).

One of the simplest recursive estimation algorithms is the classic Kalman filter (KF) (Kalman 1960), which treats the problem of a linear model under conditions of Gaussian errors. The KF equations comprise a propagation (forecast) step and a correction (assimilation) step; the latter can be derived as a Bayesian inference not similar to the least squares problem but with a prior given by the posterior from the previous (forecast) step. The beauty of the KF is that its equations have an elegant and computationally fast solution (Gelb 1974). Since the assumptions of model linearity and Gaussian errors are restrictive, various generalizations of the KF equations have been proposed, including extended and ensemble Kalman filters and particle filters (e.g., Arulampalam et al. 2002; Vrugt et al. 2005, 2013; Weerts and El Serafy 2006); development, application, and improvement of these real-time techniques is of practical interest.

9 Summary and Conclusion

Parameter estimation in hydrological modelling is a common scientific and operational task and has received a tremendous amount of research and industry attention. This chapter has reviewed the broad classes of parameter estimation techniques, namely, a priori estimation and calibration (inverse modelling), with an emphasis on parameter estimation through calibration. Strategies reviewed include manual calibration, optimization, multi-objective optimization, and probabilistic estimation, with Bayesian inference receiving most attention. Manual calibration offers the ability to exploit expert understanding of the model and fit features in an intuitive way. However, reliance on expert knowledge makes it nontransparent and difficult to reproduce independently. The formulation of a goodness-of-fit function allows a modeller to quantify model performance for a given parameter set and lends itself to automatic implementation using optimization algorithms. Multi-objective optimization avoids one of the main limitations of single-objective work and allows exploring trade-offs between different aspects of the model fit (e.g., low vs high flows, timing

of peaks, etc.). Probabilistic estimation, on the other hand, allows reflecting the uncertainty in the modelling process, which leads to parameter uncertainty. Bayesian inference supports probabilistic estimation from observed data while allowing for the use of additional information through prior distributions. Bayesian estimation can be implemented within aggregational approaches – where all sources of uncertainty are lumped into a single residual error term – or decompositional approaches, where there is an attempt to disentangle the effects of individual sources of errors (such as observational errors in rainfall forcings and streamflow responses and model structural errors). Irrespective of the calibration strategy, its assumptions must be scrutinized using posterior diagnostics, including tests for assumptions such as error heteroscedasticity, persistence, Gaussianity, and so forth. Practical implementations may also benefit from tricks such as parameter transformations and model smoothing. Directions of ongoing and future research include improvements in error model specification and the development of approaches for parameter estimation under sparse-data and ungauged conditions.

References

- M.B. Abbott, V.M. Babovic, J.A. Cunge, Reply to comment by Beven et al on “Towards the hydraulics of the hydroinformatics era” by Abbott et al. *J. Hydraul. Res.* **41**(3), 333–336 (2003)
- C. Albert, A mechanistic dynamic emulator. *Nonlinear Anal. Real World Appl.* **13**(6), 2747–2754 (2012)
- A.H.-S. Ang, W.H. Tang, *Probability Concepts in Engineering: Emphasis on Applications to Civil and Environmental Engineering* (Wiley, Hoboken, 2007)
- S.A. Archfield, M. Clark, B. Arheimer, L.E. Hay, H. McMillan, J.E. Kiang, J. Seibert, K. Hakala, A. Bock, T. Wagener, W.H. Farmer, V. Andréassian, S. Attinger, A. Viglione, R. Knight, S. Markstrom, T. Over, Accelerating advances in continental domain hydrologic modeling. *Water Resour. Res.* **51**(12), 10078–10091 (2015)
- J.G. Arnold, N. Fohrer, SWAT2000: Current capabilities and research opportunities in applied watershed modelling. *Hydrol. Process.* **19**(3), 563–572 (2005)
- M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50**(2), 174–188 (2002). Special Issue on Monte Carlo Methods for Statistical Signal Processing
- M. Asadzadeh, B.A. Tolson, Pareto archived dynamically dimensioned search with hypervolume-based selection for multiobjective optimization. *Eng. Optim.* **45**(12), 1489–1509 (2013)
- K. Beven, TOPMODEL: A critique. *Hydrol. Process.* **11**(9), 1069–1085 (1997)
- K. Beven, On the concept of model structural error. *Water Sci. Technol.* **52**, 167–175 (2005)
- K.J. Beven, A manifesto for the equifinality thesis. *J. Hydrol.* **320**, 18–36 (2006)
- K.J. Beven, A.M. Binley, The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* **6**, 279–298 (1992)
- K. Beven, I. Westerberg, On red herrings and real herrings: Disinformation and information in hydrological inference. *Hydrol. Process.* **25**, 1676–1680 (2011)
- K. Beven, P. Smith, I. Westerberg, J. Freer, Comment on “Pursuing the method of multiple working hypotheses for hydrological modeling” by P. Clark et al. *Water Resour. Res.* **48**, W11801 (2012)
- G.E.P. Box, G.C. Tiao, *Bayesian Inference in Statistical Analysis* (Wiley, New York, 1992)
- G.O. Brown, Henry Darcy and the making of a law. *Water Resour. Res.* **38**(7), 1–12 (2002)
- N. Bulygina, H. Gupta, Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation. *Water Resour. Res.* **45**, W00B13 (2009)

- T.G. Chapman, Optimization of a rainfall-runoff model for an arid zone catchment, in *I.A.S.H.-UNESCO Symposium on the Results of Research on Representative and Experimental Basins*, (IASH-AISH Publ, Wellington, 1970), pp. 126–144
- F.H. Chiew, L. Siriwardena, Estimation of SIMHYD parameter values for application in ungauged catchments, in *MODSIM 2005 International Congress on Modelling and Simulation*, ed. by A. Zerger, R.M. Argent (Modelling and Simulation Society of Australia and New Zealand, Melbourne, Australia, 2005), pp. 2883–2889
- F.H.S. Chiew, M.J. Stewardson, T.A. McMahon, Comparison of six rainfall-runoff modelling approaches. *J. Hydrol.* **147**, 1–36 (1993)
- M.P. Clark, A.G. Slater, D.E. Rupp, R.A. Woods, J.A. Vrugt, H.V. Gupta, T. Wagener, L.E. Hay, Framework for understanding structural errors (FUSE): A modular framework to diagnose differences between hydrological models. *Water Resour. Res.* **44**, W00B02 (2008). <https://doi.org/10.1029/2007WR006735>
- M.P. Clark, D. Kavetski, F. Fenicia, Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* **47**, W09301 (2011)
- M.P. Clark, D. Kavetski, F. Fenicia, Reply to comment by K. Beven et al. on “Pursuing the method of multiple working hypotheses for hydrological modeling”. *Water Resour. Res.* **48**, W11802 (2012)
- M.P. Clark, B. Nijssen, J.D. Lundquist, D. Kavetski, D.E. Rupp, R.A. Woods, J.E. Freer, E.D. Gutmann, A.W. Wood, L.D. Brekke, J.R. Arnold, D.J. Gochis, R.M. Rasmussen, A unified approach for process-based hydrologic modeling: 1. Modeling concept. *Water Resour. Res.* **51**(4), 2498–2514 (2015)
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: A review. *J. Hydrol.* **375**, 613–626 (2009)
- J. Craig, et al., Raven User’s and Developer’s manual v2.7, <http://www.civil.uwaterloo.ca/jrcraig/Raven/>. (University of Waterloo, 2017)
- B. de Finetti, Foresight: Its logical laws, its subjective sources, in *Studies in Subjective Probability*, ed. by H.E. Kyburg (Wiley, New York, 1964), pp. 93–158
- N. De Vleeschouwer, V.R.N. Pauwels, Assessment of the indirect calibration of a rainfall-runoff model for ungauged catchments in Flanders. *Hydrol. Earth Syst. Sci.* **17**, 2001–2016 (2013)
- J. Demargne, L. Wu, S.K. Regonda, J.D. Brown, H. Lee, M. He, D.J. Seo, R. Hartman, H.D. Herr, M. Fresch, J. Schaake, Y. Zhu, The science of NOAA’s operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* **95**(1), 79–98 (2014)
- J. Doherty, Ground water model calibration using pilot points and regularization. *Ground Water* **41**, 170–177 (2003)
- J. Doherty, *PEST: Model Independent Parameter Estimation*, 5th edn. (Watermark Numerical Computing, Brisbane, 2005)
- Q. Duan, S. Sorooshian, V. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **28**(4), 1015–1031 (1992)
- Q. Duan, J. Schaake, V. Andreassian, S.W. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, E.F. Wood, Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320**(1–2), 3–17 (2006)
- A. Efstratiadis, D. Koutsoyiannis, One decade of multi-objective calibration approaches in hydrological modelling: A review. *Hydrol. Sci. J.* **55**(1), 58–78 (2010)
- G. Evin, D. Kavetski, M. Thyre, G. Kuczera, Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. *Water Resour. Res.* **49**, 4518–4524 (2013)
- G. Evin, M. Thyre, D. Kavetski, D. McInerney, G. Kuczera, Comparison of joint versus post-processor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. *Water Resour. Res.* **50**, 2350–2375 (2014)
- F. Fenicia, H.H.G. Savenije, P. Matgen, L. Pfister, Understanding catchment behavior through stepwise model concept improvement. *Water Resour. Res.* **44**, W01402 (2008)

- F. Fenicia, S. Wrede, D. Kavetski, L. Pfister, L. Hoffmann, H. Savenije, J.J. McDonnell, Impact of mixing assumptions on mean residence time estimation. *Hydrol. Process.* **24**(12), 1730–1741 (2010). (Special Issue on Residence Times and Preferential Flows)
- F. Fenicia, D. Kavetski, H.H.G. Savenije, Elements of a flexible approach for conceptual hydrological modeling: Part 1. Motivation and theoretical development. *Water Resour. Res.* **47**, W11510 (2011)
- F. Fenicia, D. Kavetski, H.H.G. Savenije, P. L, From spatially variable streamflow to distributed hydrological models: Analysis of key modeling decisions. *Water Resour. Res.* **52**, 954–989 (2016)
- F. Fenicia, D. Kavetski, P. Reichert, C. Albert, Signature-domain calibration of hydrological models using approximate Bayesian computation: Empirical analysis of fundamental properties. *Water Resour. Res.* in press, <https://doi.org/10.1002/2017WR021616> (2018)
- C.W. Fetter, *Applied Hydrogeology*, 3rd edn. (Prentice-Hall, Upper Saddle River, 1994)
- J. Freer, K. Beven, B. Ambroise, Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resour. Res.* **32**(7), 2161–2173 (1996)
- J.E. Freer, H. McMillan, J.J. McDonnell, K.J. Beven, Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures, *J. Hydrol.* **291**(3–4), 254–277 (2004)
- R.A. Freeze, R.L. Harlan, Blueprint for a physically-based, digitally-simulated hydrologic response model. *J. Hydrol.* **9**, 237–258 (1969)
- A. Gelb (ed.), *Applied Optimal Estimation* (MIT Press, Cambridge, MA, 1974)
- A. Gelman, J.B. Carlin, H.S. Stern, D.B. Rubin, *Bayesian Data Analysis* (Chapman and Hall, London, 1998)
- P.E. Gill, W. Murray, M.H. Wright, *Practical Optimization* (Academic, London, 1981)
- L. Giustarini, R. Hostache, D. Kavetski, M. Chini, G. Corato, S. Schlaffer, P. Matgen, Probabilistic flood mapping using synthetic aperture radar data. *IEEE Trans. Geosci. Remote Sens.* **54**(12), 6958–6969 (2016)
- R.S. Govindaraju, Artificial neural networks in hydrology. I: Preliminary concepts. *J. Hydrol. Eng.* **5**(2), 115–123 (2000)
- R.B. Grayson, I.D. Moore, T.A. McMahon, Physically based hydrologic modeling: 2. Is the concept realistic? *Water Resour. Res.* **28**(10), 2659–2666 (1992)
- V.K. Gupta, S. Sorooshian, The automatic calibration of conceptual catchment models using derivative-based optimization algorithms. *Water Resour. Res.* **21**(4), 473–485 (1985)
- H.V. Gupta, S. Sorooshian, P.O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.* **34**(4), 751–763 (1998)
- H.V. Gupta, T. Wagener, Y. Liu, Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Process.* **22**, 3802–3813 (2008)
- T.T. Hailegeorgis, K. Alfredsen, Regional flood frequency analysis and prediction in ungauged basins including estimation of major uncertainties for mid-Norway. *J. Hydrol.* **9**, 104–126 (2017)
- A.W. Harbaugh, MODFLOW-2005, the U.S. Geological Survey modular ground-water model – the Ground-Water Flow Process, U.S. Geological Survey Techniques and Methods 6-A16 (2005)
- M.C. Hill, D. Kavetski, M.P. Clark, M. Ye, M. Arabi, D. Lu, L. Foglia, S. Mehl, Practical use of computationally frugal model analysis methods. *Groundwater* **54**(2), 159 (2015)
- R. Hostache, X. Lai, J. Monnier, C. Puech, Assimilation of spatially distributed water levels into a shallow-water model. Part II: Use of a remote sensing image of Mosel River. *J. Hydrol.* **390**(3–4), 257–268 (2010)
- M. Hrachowitz, H.H.G. Savenije, G. Blöschl, J.J. McDonnell, M. Sivapalan, J.W. Pomeroy, B. Arheimer, T. Blume, M.P. Clark, U. Ehret, F. Fenicia, J.E. Freer, A. Gelfan, H.V. Gupta, D.A. Hughes, R.W. Hut, A. Montanari, S. Pande, D. Tetzlaff, P.A. Troch, S. Uhlenbrook, T. Wagener, H.C. Winsemius, R.A. Woods, E. Zehe, C. Cudennec, A decade of predictions in ungauged basins (PUB) – A review. *Hydrol. Sci. J.* **58**(6), 198–1255 (2013)

- M. Hrachowitz, O. Fovet, L. Ruiz, T. Euser, S. Gharari, R. Nijzink, J. Freer, H.H.G. Savenije, C. Gascuel-Odoux, Process consistency in models: The importance of system signatures, expert knowledge, and process complexity. *Water Resour. Res.* **50**(9), 7445–7469 (2014)
- D. Huard, A. Mailhot, Calibration of hydrological model GR2M using Bayesian uncertainty analysis. *Water Resour. Res.* **44**, W02424 (2008)
- R.P. Ibbitt, T. O'Donnell, Designing conceptual catchment models for automatic fitting methods, in *Mathematical Models in Hydrology Symposium*, IAHS-AISH Publication No. 101(2) (1971), pp. 461–475
- V.Y. Ivanov, E.R. Vivoni, R.L. Bras, D. Entekhabi, Catchment hydrologic response with a fully distributed triangulated irregular network model. *Water Resour. Res.* **40**(11), W11102 (2004). <https://doi.org/10.1029/2004WR003218>
- A.J. Jakeman, G.M. Hornberger, How much complexity is warranted in a rainfall-runoff model? *Water Resour. Res.* **29**(8), 2637–2649 (1993)
- R.E. Kalman, A new approach to linear filtering and prediction problems. *J. Basic Eng.* **82**(1), 35–45 (1960)
- D. Kavetski, Analysis of input data uncertainty and numerical robustness in conceptual rainfall-runoff modelling, PhD Thesis, Faculty of Engineering and Built Environment, University of Newcastle (2005)
- D. Kavetski, M.P. Clark, Ancient numerical daemons of conceptual hydrological modeling. Part 2: Impact of time stepping schemes on model analysis and prediction. *Water Resour. Res.* **46**, W10511 (2010). <https://doi.org/10.1029/2009WR008896>
- D. Kavetski, G. Kuczera, Model smoothing strategies to remove microscale discontinuities and spurious secondary optima in objective functions in hydrological calibration. *Water Resour. Res.* **43**, W03411 (2007). <https://doi.org/10.1029/2006WR005195>
- D. Kavetski, S. Franks, G. Kuczera, Confronting input uncertainty in environmental modelling, in *Calibration of Watershed Models*. Water Science and Application Series 6, ed. by Q.Y. Duan, H.V. Gupta, S. Sorooshian, A. Rousseau, R. Tourcotte. (American Geophysical Union, Washington, DC, 2002), pp. 49–68
- D. Kavetski, G. Kuczera, S.W. Franks, Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resour. Res.* **42**(3), W03407 (2006a)
- D. Kavetski, G. Kuczera, S.W. Franks, Calibration of conceptual hydrological models revisited: 1. Overcoming numerical artefacts. *J. Hydrol.* **320**(1–2), 173–186 (2006b)
- D. Kavetski, G. Kuczera, S.W. Franks, Calibration of conceptual hydrological models revisited: 2. Improving optimisation and analysis. *J. Hydrol.* **320**(1–2), 187–201 (2006c)
- D. Kavetski, G. Kuczera, M. Thyre, B. Renard, Multistart Newton-type optimisation methods for the calibration of conceptual hydrological models, In Proceedings of Oxley, L. and Kulasinghe, D. (eds) MODSIM 2007 International Congress on Modelling and Simulation, Christchurch, New Zealand. (Modelling and Simulation Society of Australia and New Zealand, 2007)
- D. Kavetski, F. Fenicia, P. Reichert, C. Albert, Signature-domain calibration of hydrological models using approximate Bayes computation: Theory and comparison to existing applications. *Water Resour. Res.* in press, <https://doi.org/10.1002/2017WR020528> (2018)
- G.B. Kingston, H.R. Maier, M.F. Lambert, Bayesian model selection applied to artificial neural networks used for water resources modeling. *Water Resour. Res.* **44**, W04419 (2008)
- J.W. Kirchner, Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resour. Res.* **42**(3), W03S04 (2006). <https://doi.org/10.1029/2005WR004362>
- L.F. Konikow, J.D. Bredehoeft, Ground-water models cannot be validated. *Adv. Water Resour.* **15**, 75–83 (1992)
- V. Koren, M. Smith, Q. Duan, Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall-runoff models, in *Calibration of Watershed Models*, ed. by Q. Duan, H.V. Gupta, S. Sorooshian, A.N. Rousseau, R. Turcotte (AGU Press, Washington, DC, 2003)
- R. Krzysztofowicz, Bayesian theory of probabilistic forecasting via a deterministic hydrologic model. *Water Resour. Res.* **35**(9), 2739–2750 (1999)

- G. Kuczera, S. Franks, Testing hydrologic models: Fortification or falsification? in *Mathematical Modelling of Large Watershed Hydrology*, ed. by V.P. Singh, D.K. Frevert (Water Resources Publications, Littleton, 2002)
- G. Kuczera, D. Kavetski, S.W. Franks, M. Thyer, Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *J. Hydrol.* **331**(1–2), 161–177 (2006)
- E. Laloy, B. Rogiers, J.A. Vrugt, D. Mallants, D. Jacques, Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion. *Water Resour. Res.* **49**(5), 2664–2682 (2013)
- J. Le Coz, B. Renard, L. Bonnifait, F. Branger, R. Le Boursicaud, Combining hydraulic knowledge and uncertain gaugings in the estimation of hydrometric rating curves: A Bayesian approach. *J. Hydrol.* **509**, 573–587 (2014)
- D.R. Legates, G.J. McCabe Jr., Evaluating the use of “goodness-of-fit” measures in hydrologic and hydroclimatic model validation. *Water Resour. Res.* **35**(1), 233–241 (1999)
- J. Lerat, C. Pickett-Heaps, D. Shin, S. Zhou, P. Feikema, U. Khan, R. Laugesen, N. Tuteja, G. Kuczera, M. Thyer, D. Kavetski, Dynamic streamflow forecasts within an uncertainty framework for 100 catchments in Australia, in *Hydrology and Water Resources Symposium: The Art and Science of Water*, (Engineers Australia, Barton, ACT, Australia, 2015), pp. 1396–1403
- G. Lindstrom, B. Johansson, M. Persson, M. Gardelin, S. Bergstrom, Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **201**, 272–288 (1997)
- D.P. Loucks, J.R. Stedinger, D.A. Haith, *Water Resource Systems Planning and Analysis* (Prentice-Hall, Englewood Cliffs, 1981)
- D.R. Maidment, *Handbook of Hydrology* (McGraw-Hill, New York, 1993)
- P. Mantovan, E. Todini, Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. *J. Hydrol.* **330**(1–2), 368–381 (2006)
- P. Mantovan, E. Todini, M.L.V. Martina, Reply to comment by Keith Beven, Paul Smith and Jim Freer on “Hydrological forecasting uncertainty assessment: Inconherence of the GLUE methodology”. *J. Hydrol.* **338**, 319–324 (2007)
- A. Marchi, E. Salomons, A. Simpson, A. Zecchin, H. Maier, Z. Wu, C. Stokes, W. Wu, G.C. Dandy, The battle of the water networks II (BWN-II). *J. Water Resour. Plann. Manage.* **140**, 04014009:04014001–04014009:04014014 (2014)
- E.S. Martins, J.R. Stedinger, Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.* **36**(3), 737–744 (2000)
- D. McInerney, M. Thyer, D. Kavetski, J. Lerat, G. Kuczera, Improving probabilistic prediction of daily streamflow by identifying Pareto optimal approaches for modeling heteroscedastic residual errors. *Water Resour. Res.* **53**, 2199–2239 (2017)
- H. McMillan, B. Jackson, M. Clark, D. Kavetski, R. Woods, Rainfall uncertainty in hydrologic modelling: An evaluation of multiplicative error models. *J. Hydrol.* **400**, 83–94 (2011)
- M. Merriman, On the history of the method of least squares. *Analyst* **4**(2), 33–36 (1877)
- D.A. Miller, R.A. White, A conterminous United States multi-layer soil characteristics data set for regional climate and hydrology modeling. *Earth Interact.* **2**, 2 (1999)
- A. Montanari, E. Toth, Calibration of hydrological models in the spectral domain: An opportunity for scarcely gauged basins? *Water Resour. Res.* **43**, W05434 (2007)
- M. Morawietz, C.-Y. Xu, L. Gottschalk, L.M. Tallaksen, Systematic evaluation of autoregressive error models as post-processors for a probabilistic streamflow forecast system. *J. Hydrol.* **407**(1–4), 58–72 (2011)
- J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models. Part 1 – A discussion of principles. *J. Hydrol.* **10**, 282–290 (1970)
- J.C. Neal, P.M. Atkinson, H.C. W, Flood inundation model updating using an ensemble Kalman filter and spatially distributed measurements. *J. Hydrol.* **336**, 401–415 (2007)
- J. Neal, G. Schumann, P. Bates, A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. *Water Resour. Res.* **48**, W11506 (2012)

- D.J. Nott, L. Marshall, J. Brown, Generalized likelihood uncertainty estimation (GLUE) and approximate Bayesian computation: What's the connection? *Water Resour. Res.* **48**, W12602 (2012)
- W.L. Oberkampf, J.C. Helton, C.A. Joslyn, S.F. Wojtkiewicz, S. Ferson, Challenge problems: Uncertainty in system response given uncertain parameters. *Reliab. Eng. Syst. Saf.* **85**(1–3), 11–19 (2004)
- A. O'Hagan, J. Oakley, Probability is perfect, but we can't elicit it perfectly. *Reliab. Eng. Syst. Saf.* **85**(1–3), 239–248 (2004)
- F. Pappenberger, K.J. Beven, Ignorance is bliss: Or seven reasons not to use uncertainty analysis. *Water Resour. Res.* **42**, W05302 (2006). <https://doi.org/10.1029/2005WR004820>
- C. Perrin, C. Michel, V. Andreassian, Does a large number of parameters enhance model performance? Comparative assessment of common catchment model structures on 429 catchments. *J. Hydrol.* **242**(3–4), 275–301 (2001)
- C. Perrin, C. Michel, V. Andreassian, Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **279**(1–4), 275–289 (2003)
- F. Pianosi, L. Raso, Dynamic modeling of predictive uncertainty by regression on absolute errors. *Water Resour. Res.* **48**, W03516 (2012)
- W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes in Fortran-77: The Art of Scientific Computing* (Cambridge University Press, Cambridge, 1992)
- R. Pushpalatha, C. Perrin, N.L. Moine, V. Andréassian, A review of efficiency criteria suitable for evaluating low-flow simulations. *J. Hydrol.* **420**, 171–182 (2012)
- Y. Qin, D. Kavetski, G. Kuczera, A robust Gauss-Newton algorithm for the optimization of hydrological models: 2. Benchmarking against industry-standard algorithms. *Water Resour. Res.* in review, <https://doi.org/10.1029/2017WR022489> (2018)
- P. Reichert, J. Mieleitner, Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resour. Res.* **45**, W10402 (2009)
- P. Reichert, N. Schuwirth, Linking statistical bias description to multiobjective model calibration. *Water Resour. Res.* **48**, W09543 (2012)
- P. Reichert, S.D. Langhans, J. Lienert, N. Schuwirth, The conceptual foundation of environmental decision support. *J. Environ. Manag.* **154**, 316–332 (2015)
- R.H. Reichle, Data assimilation methods in the Earth sciences. *Adv. Water Resour.* **31**(11), 1411–1418 (2008)
- B. Renard, E. Leblois, G. Kuczera, D. Kavetski, M. Thyre, S. Franks, Characterizing errors in areal rainfall estimates: Application to uncertainty quantification and decomposition in hydrologic modelling. H2009: 32nd Hydrology and Water Resources Symposium, Newcastle (Engineers Australia, Barton ACT, 2009), pp. 505–516
- B. Renard, D. Kavetski, M. Thyre, G. Kuczera, S.W. Franks, Understanding predictive uncertainty in hydrologic modeling: Le challenge of identifying input and structural errors. *Water Resour. Res.* **46**, W05521 (2010). <https://doi.org/10.1029/2009WR008328>
- B. Renard, D. Kavetski, E.T. Leblois, M. Thyre, G. Kuczera, S.W. Franks, Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resour. Res.* **47**(11), W11516 (2011)
- B. Revilla-Romero, N. Wanders, P. Burek, P. Salamon, A. de Roo, Integrating remotely sensed surface water extent into continental scale hydrology. *J. Hydrol.* **543**(Pt B), 659–670 (2016)
- J.D. Salas, Analysis and modeling of hydrologic time series, in *Handbook of Hydrology*, ed. by D.R. Maidment (McGraw-Hill, New York, 1993), pp. 19.11–19.72
- L. Samaniego, R. Kumar, S. Attinger, Multiscale parameter regionalization of a grid-based hydrologic model at the mesoscale. *Water Resour. Res.* **46**(5), W05523 (2010)
- H.H.G. Savenije, The art of hydrology. *Hydrol. Earth Syst. Sci.* **13**, 157–161 (2009)
- B. Schaefli, H.V. Gupta, Do Nash values have value? *Hydrol. Process.* **21**(15), 2075–2080 (2007)
- B. Schaefli, D. Kavetski, Bayesian spectral likelihood for hydrological parameter inference. *Water Resour. Res.* **53**, 6857–6884 (2017)
- B. Schaefli, D.B. Talmaba, A. Musy, Quantifying hydrological modeling errors through a mixture of normal distributions. *J. Hydrol.* **332**, 303–315 (2007)

- G. Schoups, J.A. Vrugt, A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors. *Water Resour. Res.* **46**, W10531 (2010)
- D.-J. Seo, H.D. Herr, J.C. Schaake, A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci.* **3**, 1987–2035 (2006)
- M. Shafii, B.A. Tolson, Optimizing hydrological consistency by incorporating hydrological signatures into model calibration objectives. *Water Resour. Res.* **51**(5), 3796–3814 (2015)
- V.P. Singh, D.A. Woolhiser, Mathematical modeling of watershed hydrology. *J. Hydrol. Eng.* **7**(4), 270–292 (2002)
- M. Sivapalan, G. Bloschl, L. Zhang, R. Vertessy, Downward approach to hydrological prediction. *Hydrol. Process.* **17**(11), 2101–2111 (2003a)
- M. Sivapalan, K. Takeuchi, S.W. Franks, V.K. Gupta, H. Karambiri, V. Lakshmi, X. Liang, J.J. McDonnell, E.M. Mendiondo, P.E. O'Connell, T. Oki, J.W. Pomeroy, D. Schertzer, S. Uhlenbrook, E. Zehe, IAHS decade on predictions in ungauged basins (PUB). *Hydrol. Sci. J.* **48**(6), 857–880 (2003b)
- B.E. Skahill, J. Doherty, Efficient accommodation of local minima in watershed model calibration. *J. Hydrol.* **329**, 122 (2006). in press
- P. Smith, K.J. Beven, J.A. Tawn, Informal likelihood measures in model assessment: Theoretic development and investigation. *Adv. Water Resour.* **31**(8), 1087–1100 (2008)
- T. Smith, A. Sharma, L. Marshall, R. Mehrotra, S. Sisson, Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. *Water Resour. Res.* **46**(12), W12551 (2010). <https://doi.org/10.1029/2010WR009514>
- S. Sorooshian, J.A. Dracup, Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resour. Res.* **16**(2), 430–442 (1980)
- J.R. Stedinger, R.M. Vogel, S.U. Lee, R. Batchelder, Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resour. Res.* **44**, W00B06 (2008)
- V.L. Streeter, E.B. Wylie, *Fluid Mechanics*, First SI Metric Edition. (McGraw-Hill, Singapore, 1983)
- L.M. Tallaksen, A review of baseflow recession analysis. *J. Hydrol.* **165**, 349–370 (1995)
- A. Tarantola, *Inverse Problem Theory and Methods for Model Parameter Estimation* (Society for Industrial and Applied Mathematics, Philadelphia, 2005)
- M. Thyer, G. Kuczera, Q.J. Wang, Quantifying parameter uncertainty in stochastic models using the Box-Cox transformation. *J. Hydrol.* **265**(1–4), 246–257 (2002)
- M. Thyer, B. Renard, D. Kavetski, G. Kuczera, S. Franks, S. Srikanthan, Critical evaluation of parameter consistency and predictive uncertainty in hydrological modelling: A case study using Bayesian total error analysis. *Water Resour. Res.* **45**, W00B14 (2009)
- B.A. Tolson, C.A. Shoemaker, Dynamically dimensioned search algorithm for computationally efficient watershed model calibration. *Water Resour. Res.* **43**, W01413 (2007)
- A.F.B. Tompson, R. Ababou, L.W. Gelhar, Implementation of the 3-dimensional turning bands random field generator. *Water Resour. Res.* **25**(10), 2227–2243 (1989)
- T. Toni, D. Welch, N. Strelkowa, A. Ipsen, M.P.H. Stumpf, Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *J. R. Soc. Interface* **6**(31), 187–202 (2009)
- N.K. Tuteja, D. Shin, R. Laugesen, U. Khan, Q. Shao, E. Wang, M. Li, H. Zheng, G. Kuczera, D. Kavetski, G. Evin, M. Thyer, A. MacDonald, T. Chia, B. Le, *Experimental Evaluation of the Dynamic Seasonal Streamflow Forecasting Approach* (Australian Bureau of Meteorology, Melbourne, 2011)
- N.K. Tuteja, S. Zhou, J. Lerat, Q.J. Wang, D. Shin, D.E. Robertson, Overview of communication strategies for uncertainty in hydrological forecasting in Australia, in *Handbook of Hydrometeorological Ensemble Forecasting*, ed. by Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H.L. Cloke, J.C. Schaake (Springer, Berlin/Heidelberg, 2017), pp. 1–19

- R.M. Vogel, Stochastic watershed models for hydrologic risk management. *Water Secur.* **1**, 28–35 (2017)
- J.A. Vrugt, B.A. Robinson, Improved evolutionary optimization from genetically adaptive multi-method search. *Proc. Natl. Acad. Sci. U. S. A.* **104**(3), 708–711 (2007)
- J.A. Vrugt, M. Sadegh, Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resour. Res.* **49**(7), 4335–4345 (2013)
- J.A. Vrugt, H.V. Gupta, L.A. Bastidas, W. Bouten, S. Sorooshian, Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.* **39**(8), 1214 (2003)
- J.A. Vrugt, C.G.H. Diks, H.V. Gupta, W. Bouten, J.M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resour. Res.* **41**(1), W01017 (2005)
- J.A. Vrugt, C.J.F. ter Braak, M.P. Clark, J.M. Hyman, B.A. Robinson, Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.* **44**, W00B09 (2008)
- J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, B.A. Robinson, J.M. Hyman, D. Higdon, Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlinear Sci. Numer. Simul.* **10**(3), 273–290 (2009a)
- J.A. Vrugt, C.J.F. ter Braak, H.V. Gupta, B.A. Robinson, Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stoch. Env. Res. Risk A.* **23**(7), 1011–1026 (2009b)
- J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, G. Schoups, Hydrologic data assimilation using particle Markov chain Monte Carlo simulation: Theory, concepts and applications. *Adv. Water Resour.* **51**, 457–478 (2013)
- Q.J. Wang, D.E. Robertson, Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.* **47**, W02546 (2011)
- Q.J. Wang, D.E. Robertson, F.H.S. Chiew, A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.* **45**(5), W05407 (2009)
- A.H. Weerts, G.Y.H. El Serafy, Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resour. Res.* **42**, W09403 (2006)
- W.D. Welsh, J. Vaze, D. Dutta, D. Rassam, J.M. Rahman, I.D. Jolly, P. Wallbrink, G.M. Podger, M. Bethune, M.J. Hardy, J. Teng, J. Lerat, An integrated modelling framework for regulated river systems. *Environ. Model Softw.* **39**, 81–102 (2013)
- I. Westerberg, J.-L. Guerrero, J. Seibert, K.J. Beven, S. Halldin, Stage-discharge uncertainty derived with a non-stationary rating curve in the Choluteca River, Honduras. *Hydrol. Process.* **25**(4), 603–613 (2010)
- I.K. Westerberg, H.K. McMillan, Uncertainty in hydrological signatures. *Hydrol. Earth Syst. Sci.* **19**(9), 3951–3968 (2015)
- S. Westra, M. Thyre, M. Leonard, D. Kavetski, M. Lambert, A strategy for diagnosing and interpreting hydrological model nonstationarity, *Water Resources Research*, **50**(6), 5090–5113 (2014)
- D.P. Wright, M. Thyre, S. Westra, Influential point detection diagnostics in the context of hydrological model calibration. *J. Hydrol.* **527**, 1161–1172 (2015)
- K.K. Yilmaz, H.V. Gupta, T. Wagener, A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. *Water Resour. Res.* **44**, W09417 (2008)
- P. Young, Data-based mechanistic modelling of environmental, ecological, economic and engineering systems. *Environ. Model Softw.* **13**(2), 105–122 (1998)
- P.C. Young, M. Ratto, A unified approach to environmental systems modeling. *Stoch. Env. Res. Risk A.* **23**(7), 1037–1057 (2009)



Methods to Estimate Optimal Parameters

Tiantian Yang, Kuolin Hsu, Qingyun Duan, Soroosh Sorooshian, and Chen Wang

Contents

1	Introduction	524
2	Hydrologic Model Parameter Calibration	526
3	Overview of Optimal Parameter Estimation Approaches	528
3.1	Local Search Methods	529
3.2	Global Search Methods	532
3.3	Surrogate Modeling-Based Methods	541
3.4	Deterministic Multiobjective Search Methods	544
4	Examples of Hydrological Applications	547
5	Summary and Conclusion	554
	References	554

Abstract

Model, data, and parameter estimation are three fundamental elements in hydrologic process modeling and forecasting. Recent progresses in hydrologic modeling have been made toward more efficient and effective estimation of model

T. Yang · S. Sorooshian
University of California, Irvine, CA, USA
e-mail: tiantiay@uci.edu

K. Hsu (✉)
Civil and Environmental Engineering, The Henry Samueli School of Engineering, University of California, Irvine, CA, USA
e-mail: kuolin@uci.edu

Q. Duan
Faculty of Geographical Science, Beijing Normal University, Beijing, China
e-mail: qyduan@bnu.edu.cn

C. Wang
South China Botanical Garden, Chinese Academy of Sciences, Richland, WA, USA
e-mail: chen.wang@scbg.ac.cn

parameters. In this chapter, classical and recently developed parameter optimization methods and their applications in hydrological model calibration are reviewed. Those methods include gradient-based optimization methods, direct search methods, and recently developed stochastic global optimization methods. A recently developed surrogate model approach, with the purpose to reduce computational burden of model which runs through replacing the hydrologic process model with a cheaper-to-run surrogate model, is also discussed. Extending from a single objective function parameter optimization, multiobjective optimization methods and their core concept in deriving trade-offs are also summarized. Examples are provided to demonstrate the strengths and limitations of optimization algorithms summarized in this chapter.

Keywords

Optimization · Hydrologic Model · Evolutionary Algorithm · Automatic Parameter Estimation · Surrogate Model

1 Introduction

Hydrologic models are extensively used in academia, industry, and operating agencies for flood forecasting, streamflow simulation, and water resources management. The successful use of hydrologic models to simulate natural processes depends on many factors, including (1) the mathematical formulation of hydrologic model, i.e., the mathematical representation of natural rainfall-runoff processes in a certain level of sophistication and its corresponding assumptions; (2) sufficiency and accuracy of observation data at proper temporal and spatial resolutions, such as in situ streamflow observations and precipitation measurements from rain gauge, radar network, or remotely sensed information; (3) the properly calibrated model parameters (i.e., the global optimal parameters in the feasible domain), which significantly affect the accuracy and uncertainty of hydrological prediction.

This chapter presents model parameter calibration methods in three parts. The first part reviews recent development of the methods to estimate optimal parameters of hydrologic models, especially those heuristic methods used in automatic parameter estimation. The second part focuses on the search mechanism and procedures employed in different methods. And the third part provides examples to illustrate the strengths and limitations of different methods. An overall review of this chapter is summarized below.

This chapter starts with an introduction of hydrologic models and a general mathematical formulation of parameter estimation from the maximum likelihood perspective. Two classical parameter estimation methods are introduced, namely the *Steepest decent method* and *Newton method*, known as the gradient-based local

search methods. Those classical *gradient-based methods* are efficient, but require the objective function to be differentiable, i.e., gradient of the objective function exists for the entire parameter space. Differing from the gradient-based methods, another type of parameter estimation methods is called *direct search methods*, which does not require information or the existence of gradient of the objective function. A *Simplex Downhill method* is presented in detail as one of the efficient and robust *direct search methods*. Both the methods introduced as *gradient-based methods* and the *Simplex Downhill method* belong to the same category of *local search method*, because the search always starts at certain location in the parameter space, evolves toward a gradient decreasing direction in the objective function space, and finally stops when method identifies the gradient of objective function equals or approximately equals zero. As the complexity of hydrologic model and number of parameters to calibrate increase, the response surface of an objective function sometime become multimodal, i.e., there are multiple local optimums instead of a single local optimum for concave or convex problems. The roughness and multimodality place great challenges for those local search methods, and global optimization methods turn out to be powerful in addressing those issues. In the global optimization subsection, a number of nature phenomenon inspired optimization algorithms are summarized in detail, including the *genetic algorithms* (GAs), the *simulated annealing* method (SA), the *particle swarm optimization* (PSO), the *ant colony optimization* (ACO), and the *shuffled complex evolution UA* global optimization scheme (SCE-UA). Though the global optimization methods are effective in finding the global optimal parameter set, they generally require up to tens of thousands of model runs to find the global optimal solution. Given the severe computational constraint on solving such an optimization problem and the further increases of model complexity in operation (e.g., the models require a large amount of CPU time to run), many efforts are made by researchers to reduce the computational burden and replace the expensive simulation model with a cheaper-to-run surrogate model. Some fields also refer to the surrogate modeling as function approximation, meta-modeling, response surface method, or model emulation. Once the surrogate model is constructed, a global optimization algorithm can be used to identify the optimal parameter set. These kinds of algorithms are called surrogate modeling-based optimization methods. Last, we summarize several fundamental differences between single and multiple objective optimization due to the fact that many real-world problems are intrinsically multiobjective optimization problems. Examples of using some introduced methods in real-world study are provided at the end of this chapter, through which authors want to deliver the message that the practical use and selection parameter estimation method should be originated from the ultimate goal of application. There will be no single parameter estimation method that is superior than another considering all aspects of performances, i.e., evaluation metrics. One algorithm is inevitably better than another in a certain way, and vice versa. Detailed introduction of each type of method, discussion of the pros and cons, and examples are included in the rest of chapter.

2 Hydrologic Model Parameter Calibration

The rainfall-runoff conversion is a highly nonlinear process, which can be simulated by a hydrologic model (Chiu and Huang 1970; Kulandaiswamy and Subramanian 1967; Pilgrim 1976; Singh 1964). There exists a plethora of rainfall-runoff models of various complexities, from simple black-box models which are derived from statistical relationships between rainfall and runoff observations, to conceptual models based on the physical principles or empirical relationships among hydrological variables, and to the physically-based distributed models which are based on physical laws of mass, energy, and momentum conservation. Many of those models are shown to be effective in forecasting certain important features of the hydrograph, such as the rising limb, the peaking time and the peak flow rate, and/or the flow volume (Kitanidis and Bras 1980; Sorooshian 1983). All of those models contain model parameters which appear in model equations as constants or exponents and are generally nonobservable at the scale of applications. The ability of the rainfall-runoff models to capture the real-world hydrological processes is dependent on how those parameters are specified (Duan et al. 2006). In practice, model parameters are often tuned to improve the fitting between model simulations and observations. This process is also known as model calibration. Because of the highly nonlinear nature of the hydrological processes, calibration of rainfall-runoff models is faced with enormous challenges that require sophisticated mathematical tools, significant amounts of calibration data, and some degree of model knowledge (Duan et al. 1992).

A rainfall-runoff model can be represented as a mathematical function of numerous variables, including the forcing inputs (e.g., precipitation and temperature), the outputs (e.g., streamflow discharge, evapotranspiration), the transfer functions (i.e., the nonlinear equations governing the relationships among variables), the model states (e.g., river stage, soil moisture storage, snow cover, and snow water equivalent), and the model parameters. The transfer functions (g) consist of either a set of physically based or conceptual hydrologic functions or a list of experimental functions:

$$y_t = g(x_0, x_t, I, \theta) + \epsilon_t \quad (1)$$

where x_t represents the model state variables; x_0 is the initial model states; I is the input variables; θ is the model parameter vector; y_t is the model outputs and the last term; and ϵ_t is the model estimation error. Model calibration can be formulated as an inverse problem as illustrated in Fig. 1.

As in any inverse problems, a proper objective function must be specified in order to evaluate the goodness-of-fit between the model simulations and the actual observations. The error term $\epsilon_t = y_t - g(x_0, x_t, I, \theta)$ is a function of the parameter vector θ . θ can be treated as a random variable. The objective function can therefore be represented by a likelihood function expressed as below:

$$L(\theta | \text{data}) = f(y_1, y_2, \dots, y_n | \theta) \quad (2)$$

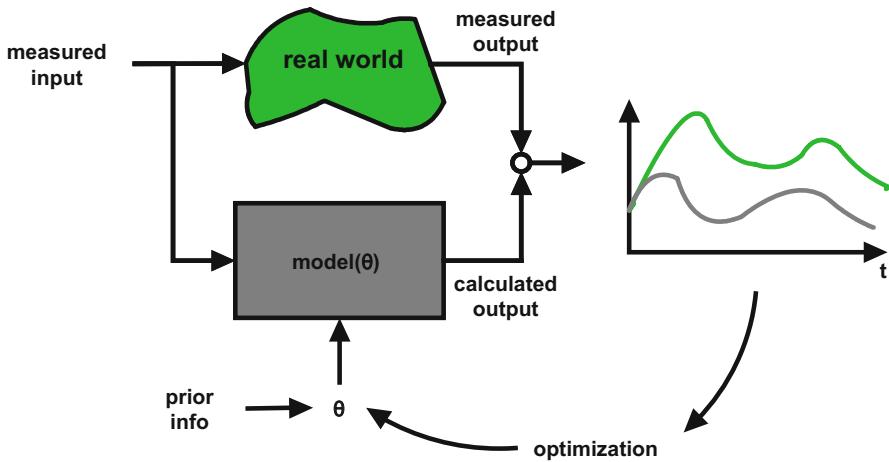


Fig. 1 Strategy for model calibration

where $L(\theta | \text{data})$ is the likelihood of θ given the data , which is equal to the joint probability density of all observations y_t , $t = 1, \dots, n$, given θ , $f(y_1, y_2, \dots, y_n | \theta)$. If one assumes statistical independence among y_t , then Eq. (2) can be written as:

$$L(\theta | \text{data}) = \prod_{t=1}^N f(y_t | \theta) \quad (3)$$

In practice, it is more convenient to compute the log likelihood function $\ln L(\theta | \text{data})$. For a likelihood function with i.i.d. Gaussian errors, it leads to the equation below:

$$\begin{aligned} \ln L((\theta | \text{data})) &= \ln \left\{ (2\pi\sigma_e^2)^{-N/2} \cdot \exp \left[-\frac{1}{2\sigma_e^2} \left(\sum_{t=1}^N \epsilon_t^2 \right) \right] \right\} \\ &= -N/2 \cdot \ln(2\pi\sigma_e^2) + \left[-\frac{1}{2\sigma_e^2} \left(\sum_{t=1}^N \epsilon_t^2 \right) \right] \end{aligned} \quad (4)$$

The maximum likelihood estimate of θ , θ^* can be obtained by solving the following optimization problem:

$$\theta^* = \text{Maximize}_{w.r.t. \theta} \ln L(\theta | \text{data}) \cong \text{Minimize}_{w.r.t. \theta} \left(\sum_{t=1}^N \epsilon_t^2 \right) \quad (5)$$

There are two types of approaches to solving Eq. (5). One approach is known as the deterministic approach which assumes that there exists a unique set of extrema of the objective function (i.e., the maximum or the minimum value of the objective function). Another approach is the stochastic approach or the Bayesian approach which assumes that the optimal solution to Eq. (5) is not a unique set of parameters,

but a posterior distribution of θ inferred based on all observations. In this chapter, we focus on estimation of optimal parameters using deterministic approaches. The next chapter discusses the approaches for estimating parameter distributions using stochastic approaches.

3 Overview of Optimal Parameter Estimation Approaches

There are different deterministic approaches to identify the optimal parameter estimates. There is a broad class of local search methods which are presented in classical nonlinear programming textbooks. Those approaches can only guarantee to find a local optimum in the presence of multiple local optima. Local search methods can be further divided into gradient methods which require the calculation of the first and/or the second derivatives of the objective function, and direct methods which do not need the derivative information of the objective function. Obviously, gradient methods require the objective function to be continuous and smooth, which can be a problem for hydrological models as many of them contain threshold parameters (Duan et al. 1992).

Another category of approaches is designed to overcome the limitations of the local search methods. Among the methods in this category, some stochastic and global search mechanisms are often used. Some popular search algorithms include genetic algorithm (GA), the shuffled complex evolution methods developed at the University of Arizona (SCE-UA), particle swarming (PS), ant colony optimization (ACO), and simulated annealing (SA), among others. Many of those algorithms are also called evolutionary algorithms, or EAs because the search strategies follow the evolutionary principles. Another group of global search methods follows the global strategies such as branch and bound methods, cutting plane methods, interval methods, filling function methods, among others. The global search methods usually require smooth and continuous objective functions and are rarely effective in solving high-dimensional hydrological model calibration. Those methods are not reviewed in this chapter. For those interested in those methods, readers may refer to Pintér (1996) and Duan et al. (2006).

More recently, there is a new category of approaches that aim to deal with large complex system models, known as the surrogate modeling-based optimization methods. Those methods are designed to use only a limited number of objective function evaluations to identify the approximate optimal solutions. The idea behind this category of methods is to construct a response surface using a small number of parameter sample sets to approximate the objective function surface. Once this response surface is found to be a reasonable approximation of the objective function, then the search would be conducted on this response surface, which is known as the surrogate model. The solution of this surrogate model would approximate the solution of the dynamic model.

Below, we provide the review of some of the most popular methods that have been used in practice, starting with several local search methods, then global search methods, and finally the surrogate modeling-based optimization methods. We also

include a section discussing deterministic multiobjective optimization search algorithms. At the end of this chapter, we provide some examples of the different search algorithms.

3.1 Local Search Methods

3.1.1 Gradient-Based Methods

Parameter estimation methods are used to find the minimum of unimodal functions for which the search is in the direction to improve function value continuously until reaching the local minimum (Singh 1995). If the derivatives of the objective function are used, the local search methods are classified as gradient search methods.

Steepest Descent

Steepest descent method is a gradient-based search method for unconstraint optimization. The search iteration can be written as follows:

$$\theta_{k+1} = \theta_k - \eta_k g_k \quad (6)$$

where η_k is the step size or learning rate; g_k is gradient direction of the likelihood function.

Newton Method

Newton method is also a second-order gradient-based method by taking the curvature of the space into consideration. The interactive form of the search algorithm is as follows:

$$\theta_{k+1} = \theta_k - \eta_k H_k^{-1} g_k \quad (7)$$

This method requires the estimation of H (Hessian matrix), which is a positive defined second-order derivative.

3.1.2 Direct Search Methods

Direct search methods are very popular because of their simplicity and do not require information about the gradient of the objective function.

Downhill Simplex

A direct search method is widely used for the objective function which is not directly differentiable. Downhill Simplex Method (Nelder and Mead 1965) is one popular direct search algorithm. The method uses the concept of a simplex of n dimensional parameters to set $n + 1$ test points to form a simplex. The objective function is evaluated at each test point and all of the simplex points are ordered (sorted) according to the function values from low to high values; meanwhile, the centroid of the simplex is estimated. A new test point is evaluated based on the reflection, expansion, and contraction of the centroid toward the worst point and

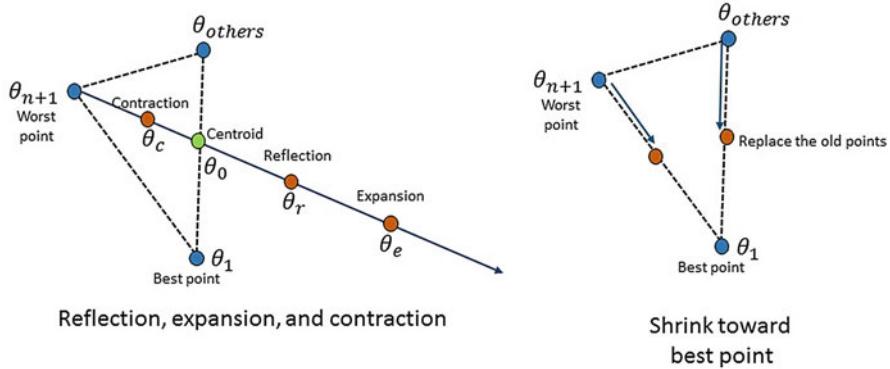


Fig. 2 Reflection, expansion, contraction, and shrinking stage of Simplex search

the worst point is replaced if the new test point is better than the worst point. At the same time, a new simplex is formed and the same process is repeated to move the simplex forward. On the other hand, if this new test point is not able to have better fitness value than the previous value, the simplex is then shrunk toward its best point. This strategy allows the simplex to continue to evolve till the convergence criteria are reached.

The key simplex evolving steps are shown in Fig. 2 and listed below:

- For an n -dimension parameter space, the test points $\theta_1, \dots, \theta_{n+1}$ are sorted based on their function values, i.e., $g(\theta_1) \leq g(\theta_2) \dots g(\theta_{n+1})$, and the location of the centroid of the simplex is calculated and represented with θ_0 .
- Test of reflection, expansion, contraction points:

$$\text{Reflection : } \theta_r = \theta_0 + \alpha(\theta_0 - \theta_{n+1})$$

$$\text{Expansion : } \theta_e = \theta_0 + \beta(\theta_r - \theta_0)$$

$$\text{Contraction : } \theta_c = \theta_0 + \gamma(\theta_{n+1} - \theta_0)$$

$\alpha, \beta, \gamma > 0$; in general, α, β, γ are set to 1.0, 2.0, and 0.5, respectively.

- Shrink the simplex if the function values of above test points $\{\theta_r, \theta_e, \theta_c\}$ are not better than the function value of the worst point $g(\theta_{n+1})$:
Shrink: $\theta_i = \theta_1 + \delta(\theta_i - \theta_1)$ for all $i = 2..n+1$ & $\delta = 0.5$.

Gradient-based and downhill simplex methods are local search methods, which are capable of finding the local minimum, but have no guarantee to find the global optimal parameter solution. Duan et al. (1992) conducted an analysis of the properties of the response surface associated with a rainfall-runoff model and found that the surface (1) contains more than one main region of attraction, (2) has many local

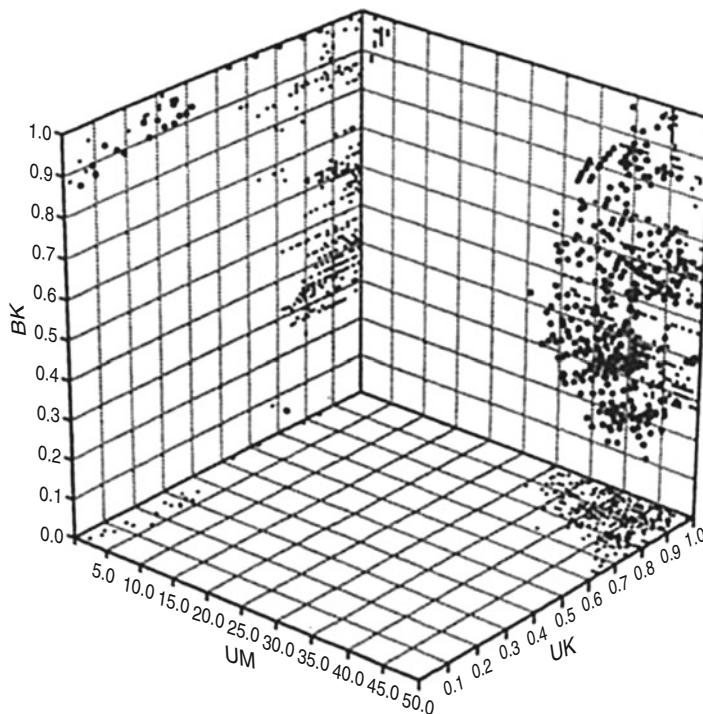


Fig. 3 Three parameters subspace of a simple conceptual catchment model (SIXPAR, Duan et al. 1992), showing locations of multiple local optima

optima within each region of attraction, (3) is rough with discontinuous derivatives, (4) is flat near the optimum with significantly different parameter sensitivities, and (5) includes long and curved ridges. Figure 3 shows many local optima within each region of attraction for a conceptual catchment model of three parameters (Duan et al. 1992). The response surface with above characteristics makes it very difficult to find the global minimum solution.

Finding global optimal solution for a complex optimization problem is of great importance to many real-world applications. Because of a large number of local minimum, it is a challenge for the local search algorithms to find the global optimal solution. Duan et al. (1992) developed a new optimization scheme, which combines the strength of the simplex method with the concept of information sharing from the evolution-based algorithms, and termed it as the shuffled complex evolution global optimization scheme – University of Arizona (SCE-UA). The concept of the SCE-UA algorithm is similar to those used in evolutional algorithms. For the past two to three decades, SCE-UA has been widely used in the parameter estimation of hydrologic models. In the next chapter, we will specifically introduce the mechanisms and concept of the SCE-UA algorithm along with other commonly used global search methods.

3.2 Global Search Methods

3.2.1 Genetic Algorithms

The genetic algorithms (GAs) belong to one of the most popular evolutionary algorithms that mimic the processes of natural selection (Goldberg 1989; Holland 1975). The natural selection is defined as the processes that organisms correspondingly survive and produce offspring with the tendency to adapt their environment. There are different types of natural selection processes, including chromosome heredity, mutation, crossover, and selection. The following Fig. 4 illustrates a conceptual GA algorithm, in which a population consists of a group of individuals (dashed-box 1). Each individual is denoted as a chromosome, which has a set of properties as genes. The chromosomes are able to mutate, carryover, and crossover with other chromosomes to produce a new generation of offspring. The production of next offspring generation is repeated following the natural selection criteria. After producing a number of new generations, the entire population will gradually carry the “good” genes from the chromosomes that have better fitness with respects to their environment, and eliminate the “bad” genes which are not suitable for surviving. In other words, the natural selection processes tend to generate offspring with better suitability to survive under the pressures from their living environment.

According to Simpson et al. (1994), the optimization of a particular problem using GA is achieved through the following concepts. First, the initial population is created by randomly selecting a number of individuals in the searching space, and each individual is called a chromosome. Second, the objective function value for each feasible solution, or individual, is defined as the fitness of chromosome to its environment. The individual (chromosome) with better objective function value (or fitness) is assumed to possess better genes (parameters), and therefore, has a higher chance to be selected to produce next generation. Last, when producing offspring from the selected parent individuals (chromosomes), the percentages of genes in each parent chromosome to crossover, carryover, and mutate are defined as algorithm parameters, i.e., crossover rate, elite rate, and mutation rate, respectively.

A generalized procedure of implementing GA is summarized as follows:

1. *Define objective function:*

Assign objective function (see Eq. (5)): $f(\theta_1, \theta_2, \dots, \theta_n)$, where n is the number of dimension.

2. *Initialization:*

Randomly sample k individuals in the parameter space to form the population

$P = \{p_1, p_2, \dots, p_k\}$. Each $p_i, i \in 1, 2, \dots, n$ is defined as an individual chromosome as shown in the dashed-box 1 of Fig. 4.

3. *Selection:*

Evaluate the fitness, i.e., objective function values for all individuals in population, and recursively select two individuals as parents for producing offspring. In the example shown in Fig. 4, chromosome 1 and 2 are selected as parents. Elite members, i.e., the individuals with high fitness values, are also selected and directly copied to next generation without any changes.

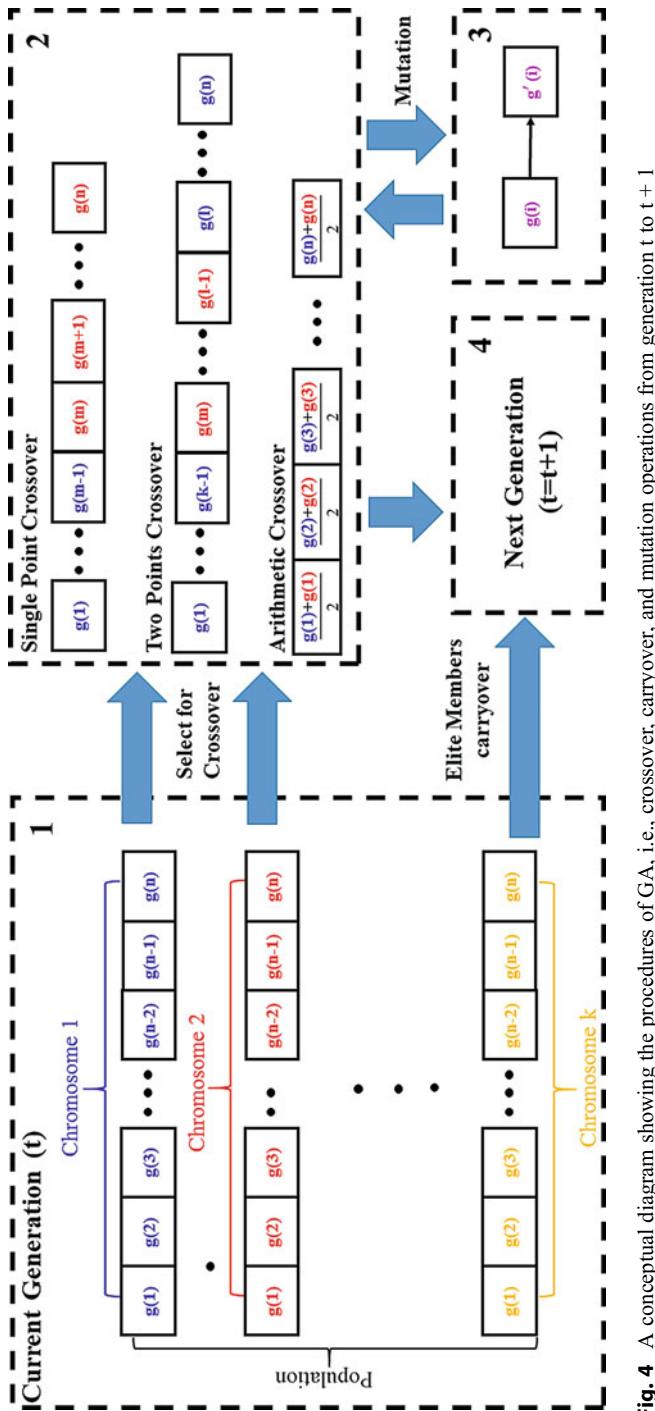


Fig. 4 A conceptual diagram showing the procedures of GA, i.e., crossover, carryover, and mutation operations from generation t to $t+1$

4. Crossover:

The production of offspring is firstly conducted by crossover operation on the selected parental individuals. There are many types of crossover operations as shown in the dashed-box 2 of Fig. 4: (a) Single-point crossover, in which a gene location m ($m \in 2, 3 \dots n - 1$) is defined and the genes of offspring before this location are from the first parent and the rest genes come from the second parent. (b) Two-point crossover, in which another gene location l ($l \in 2, 3 \dots n - 1$, and $l > m$) in the offspring is further defined and the genes after location l are copied from the first parent again instead of using the genes from the second parent. (3) Arithmetic crossover, in which the offspring genes are obtained from both parents with a average of the values of genes.

5. Mutation:

As shown in the dashed-box 3 of Fig. 4, after crossover operation, a portion of offspring genes is able to mutate from $g(i)$ to $g'(i)$, where $i \in 1, 2 \dots, n$, i.e., for real-value encoded GA, the mutation is conducted by adding a small random number to the values of offspring genes who are selected to mutate during an iteration. This mutation strategy prevents the population from trapping in local minima, and premature converging. The new genes $g'(i)$ are copied back to each offspring and replace the original genes $g(i)$.

6. Next Generation:

The new population for generation $(t + 1)$ consists of the offspring from the crossover and mutation operations, as well as the elite members directly copied from previous generation (t) .

7. Termination:

Repeat steps (3)–(6) until stopping criteria are met, i.e., total number of function evaluation reaches user-defined maximum, total number of generation reaches user-defined maximum, or the average relative changes in the objective function values over a number of generation is less than the function tolerance, etc.

The applications of GA in hydrological model calibration are extensive. Some early attempts to using GA in the field of automatic calibration of hydrological model parameters include the works by Wang (1991), Franchini (1996), Franchini and Galeati (1997), Wang (1997), Balascio et al. (1998), Savic et al. (1999), and Whigham and Crapper (1999). The usefulness of GA has also been demonstrated in many different hydrological models, such as the HYMOD model, SWAT model, the Xinanjiang Model, etc., as demonstrated by numerous studies (Babovic and Keijzer 2002; Liang et al. 2002; Srivastava et al. 2002; Cheng et al. 2006; Francés et al. 2007; Lin and Wang 2007; Zhang et al. 2009b; Wu et al. 2012). Some recent comparisons of GA against other stochastic optimization schemes are available from Wang et al. (2010) and Arsenault et al. (2013) for interested readers.

3.2.2 Simulated Annealing

The simulated annealing (SA) algorithm was originally introduced by Kirkpatrick (1984) as a robust global optimizer for addressing the issue of trapping in local minimums of classical gradient descent method. The concept of SA was inspired by

the process of annealing in metal work, in which a metal material was repeatedly heated and cooled down to improve the stiffness of metals. The heating process allows the metal molecules to vibrate in their neighborhood, and partially breaks the molecular bonds. The cooling process re-forms the molecular structure, and re-combines a stronger molecular bond, so that the whole physical system reaches an entropy maximum state. This metal work annealing concept can be creatively used for finding global optimums on multimodality response surfaces (Eglese 1990), and many real-world problems, i.e., the travel salesman (Černý 1985).

In SA implementation, it follows two conditions: (1) when the temperature is high, the status of the system is free to move to other energy states through random work and (2) when the temperature is lower, the system states are becoming restricted and therefore, the solutions can only move toward regions where energy states are lower. In each iteration, a nearby region near the current solution is tested. If the objective function of the new test point is better than the old one, the new point is used to replace the old point. Otherwise, a probability being a function of the annealing temperature is assigned to the new test point to decide whether this test point is acceptable. When the annealing temperature continues to move lower, the acceptable probability for a worse solution becomes lower. To accept worse solutions can be referred as a “hill-climbing” procedure, whereas this search strategy allows the algorithm to have the capability of escaping from local minimums. As the annealing temperature decreases, the chance of accepting worse solution will decrease. It is expected that only better solution is acceptable when annealing temperature reaches minimum. The annealing temperature can be controlled by a cooling scheme specifying how it should be progressively reduced over iteration. Theoretical study has proved that the algorithm can converge toward the optimal solution in a asymptotic manner (Granville et al. 1994).

The SA algorithm in general follows six steps as shown below:

1. *Define objective/energy function:*

Assign objective function or an energy function (see Eq. (5)): $f(\theta)$.

2. *Initialization:*

Assign initial parameters ($\theta_t = 0$), terminal temperature (T_E), cooling rate (\propto), for $t = 0$.

Find the solution ($f(\theta_t = 0)$) of the initial parameters $\theta_t = 0$.

3. *Selection of a new point (iteration):*

Set $t = t + 1$; generate a new parameters θ_t near θ_{t-1} .

4. *Selection/rejection of the selected point:*

Use Metropolis acceptance rule to accept or reject θ_t :

i.e., estimate $\Delta E = f(\theta_t) - f(\theta_{t-1})$;

For a minimization problem, θ_t is accepted to replace θ_{t-1} if $\Delta E < 0$.

Otherwise, θ_t can be accepted to replace θ_{t-1} with a probability of $p = e^{-\Delta E/T}$.

5. *Adjustment of anneal temperature:*

One way to adjust the anneal temperature is to reduce the temperature over time, such as $T = \infty \times T$ and $\infty \in [0, 1]$.

6. *Terminate*:

If $T > T_E$, repeat steps (3)–(5).

Otherwise, terminate the process. The final solution is assigned to θ_r . In addition, once other user-defined stopping criteria are met, i.e., the maximum of number of function evaluation is reached, or as the stagnation time of annealing temperature becomes zero, etc., the process is terminated.

With the extensive uses of SA in various fields, particularly, in the field of automatic calibration of hydrological model parameters, many studies have shown the usefulness of SA with different case studies. Bates (1994) pioneered in the use of SA for calibrating an SFB conceptual rainfall-runoff model. Sumner et al. (1997) applied a modified SFB model and SA optimization scheme for a large-scale case study in Australia. Thyer et al. (1999) and Madsen et al. (2002) compared SA strategy with many other population-based optimization schemes with regard to its performances in calibrating conceptual rainfall-runoff models. Bárdossy and Das (2006) applied the SA to a semidistributed HBV model and shown a good capability of SA in tuning model parameters. A number of other stochastic optimization algorithms, such as the Shuffled Complex Evolution Metropolis algorithm – University of Arizona (SCEM-UA) (Vrugt et al. 2003a), were also developed based on the combination uses of annealing concept of SA, and the Metropolis-Hastings algorithm (Hastings 1970).

3.2.3 Particle Swarm Optimization

Similar to the GAs, the particle swarm optimization (PSO) is another extensively used, population-based global optimizer, which simulates the social-individual behaviors of bird flocking and fish schooling (Kennedy 2011; Kennedy et al. 2001). The particle swarm optimization belongs to one type of swarm intelligence, in which the *particles* that mimic the behaviors of social animals are swarming in a manner of strategic movements, instead of randomly moving in the searching domain. Different from the adopted natural selection criteria in GA, i.e., mutation or crossover, in PSO the offspring production is based on the fitness of *particles* and their movement velocities toward the locations of current best, as well as the historical best location so far. This is a simplified social behavior of bird foraging. According to Eberhart and Kennedy (1995) and Shi (2001), there are three assumptions when interpreting the birds foraging behavior into PSO algorithms. First, all birds (*particles*) are assumed to be blind with regards to (i.e., do not know) the location of best food source (global optimum). Therefore, one of the effective foraging strategies for all birds (*particles*) is to fly toward the bird which is nearest to the food (the *particle* that has the best fitness). Secondly, each bird (*particle*) is assumed to be intelligent enough to remember the distance of the historical locations to food source (i.e., the fitness values during all movements during the entire search). Last, each bird (*particle*) is able to collectively adjust its next movement direction and position (i.e., the next moving direction and distance). Therefore, in PSO the population is updated by recursively approaching two best positions: (1) the best location that gives the best fitness value within current population, and (2) the historical best location that gives the best fitness value through

the entire evolution that the algorithm has achieved so far. In summary, the information sharing mechanism sorely relies on the best individual instead of chromosomes exchanges and mutations. In PSO, the movement of population is always toward the best two members, while in GA, the individuals move as a group approaching the global optimum (Panduro et al. 2009). One similarity between PSO and GA is that the population is randomly sampled from the feasible solution space (Arsenault et al. 2014).

A generalized procedure of implementing PSO is summarized as follows:

1. *Define objective function:*

Assign objective function (see Eq. (5)): $f(\theta_1, \theta_2, \dots, \theta_n)$, where n is the number of dimension.

2. *Initialization:*

Randomly sample k individuals (*particles*) in the parameter space to form the population $P = \{p_1, p_2, \dots, p_k\}$, and define each particle's neighborhood as $\mathcal{N}_i \in P$.

3. *Evaluation:*

Evaluate the fitness, i.e., objective function values for all *particles*.

4. *Swarming:*

Swarm each *particle* (p_i) in its neighborhood (\mathcal{N}_i), and store the neighborhood best location l_i , called *local best* location, for each *particle*. The *local best* location also includes the particle's current location before swarming. Then, evaluate the historical best location each particle reached so far, and store as g_i , called *global best location*. Note that *global best* for each particle is not worse than the *local best*, i.e., $f(l_i) \leq f(g_i), \forall p_i \in \mathcal{N}_i$.

5. *Update velocity*

The movement velocity for *particle* (p_i) at step t is defined as:

$$v_i^{t+1} = w v_i^t + c_1 U_1^t (g_i^t - x_i^t) + c_2 U_2^t (l_i^t - x_i^t) \quad (8)$$

where w is an user-defined algorithm parameter called *inertia weight*; c_1 and c_2 are also user-defined algorithm parameters called *acceleration coefficients*; U_1^t and U_2^t are n by n diagonal matrixes with diagonal components randomly drawn from a uniform distribution in the interval of $[0, 1]$.

6. *Update particle location*

The next movement location for each particle (p_i) is denoted as x_i^{t+1} , and it is updated based on previous location (x_i^t) and the movement velocity v_i^{t+1} that is obtained from step (5). The equation for updating *particle* location is expressed as:

$$x_i^{t+1} = x_i^t + v_i^{t+1} \quad (9)$$

7. *Termination:*

Repeat steps (3)–(6) until stopping criteria are met, i.e., total number of function evaluation reaches user-defined maximum, total number of generation reaches

user-defined maximum, or the average relative changes in the objective function value over a number of generation is less than the function tolerance, etc.

The use of PSO in the field of automatic calibration of hydrologic model parameter is extensive. Gill et al. (2006) tested both single-objective PSO and multiobjective PSO in calibrating the sacramental soil moisture accounting model, which has 13 parameters. Chau (2008) applied PSO to train a data-driven rainfall-runoff model. Zhang et al. (2009c) compared both GA and PSO with regard to their performances of calibrating Soil and Water Assessment Tool (SWAT) model. Zhang and Chiew (2009) applied PSO scheme in both the Xinanjiang and SIMHYD hydrological models. Further investigation of parameter sensitivity of the Xinanjiang model using PSO is also conducted by Kuok and Chan (2012) and Lü et al. (2013). Kamali et al. (2013) applied both single- and multiple objective PSO algorithms to the HEC-HMS model. A recent comparison of many stochastic optimization schemes in calibrating hydrological models is made available by Arsenault et al. (2014). In addition, PSO is a very useful tool in various fields, such as water stage forecasting (Chau 2007), power system design (Abido 2002), ground water management (Zambrano-Bigiarini and Rojas 2013), etc.

3.2.4 Ant Colony Optimization

The ant colony optimization (ACO) belongs to another type of swarm intelligence, which was first introduced by Dorigo (1992) in his Ph.D. dissertation, and further developed by Marco Dorigo and his colleagues (Dorigo et al. 2006; Dorigo and Blum 2005; Dorigo and Stützle 2009). The concept of ACO followed the social foraging behavior of social insects, in particular, the strategy that ants find food sources and the development of optimal paths from food sources to their nest. According to Marco Dorigo's description and the biology study by Deneubourg et al. (1983), ants initially are able to explore the area near their nest in a random manner. When ants are randomly moving on the ground, a chemical pheromone trail is left by each individual ant, which is detectable by other ants. An individual ant tends to follow the path, in probability, has the strongest pheromone concentrations that marked by other ants. Once a food source is located by an individual ant, this ant will evaluate the quantity and quality of the food source, and carry a small portion of food back to its nest. On the way back to the nest, the pheromone left by this ant will correspondingly change based on the quantity and quality of food source, so that other ants can be guided to this discovered food source. By using the pheromone trails, ants are able to indirectly exchange information of the location, quantity, and quality of food source. This communication strategy via pheromone trails is proven to be effective allowing ants to find the shortest paths between the food sources and nest (Deneubourg et al. 1990).

In ACO, an instantiated decision variable $X_i = v_i^j$ (i.e., a variable X_i with a value v_i^j assigned from its parameter domain θ_i) is termed a *solution component* and denoted by c_{ij} , where i and j are the locations connecting a searching domain. τ_{ij} is

the pheromone value or intensity associated with each solution component c_{ij} , and it is continuously updated based on time and the behavior of all ants. All possible *solution components* and feasible solutions consist a complete set of $\mathcal{N}(s^k)$, where s^k is a partial feasible solution that constructed from an empty set s by adding the 1st, 2nd . . . and k th *solution component* from the complete feasible solution set \mathcal{N} . A generalized procedure for ACO is as follows:

1. Define the objective function:

Assign objective function for the optimization problem (see Eq. (5)): $f(\theta)$.

2. Initialization:

Assign the number of ants (M) in ASO, locations of the ants in searching space, and randomly assign pheromone values τ_{ij} for each *solution components* c_{ij} that connect location i and j .

3. Define the pheromone model:

A commonly used pheromone model for ant system (Dorigo et al. 1996) is

$$p(c_i|s) = \frac{[\tau_{ij}]^\alpha \cdot [\eta(c)_{ij}]^\beta}{\sum_{c_{il} \in \mathcal{N}(s^k)} [\tau_{il}]^\alpha \cdot [\eta(c)_{il}]^\beta} \quad (10)$$

where α and β are algorithm parameters that control the significance of pheromone value τ_{ij} , and the visibility of pheromone trials $\eta(c)_{ij}$; the visibility $\eta(c)_{ij}$ is defined as the inverse Euclidean distance between location i and j .

4. Movements of ants

When individual ant (m) is in the position i and so far constructed the partial solution s^k , the probability of moving from position i to position j is given by the pheromone model in step (3). Each ant will move to its next position until all the M ants finish their movements.

5. Update pheromone values:

The pheromone values τ_{ij} are updated for all the M ants according to the following equation:

$$\tau_{ij,t} = (1 - \rho) \cdot \tau_{ij,t-1} + \sum_{m=1}^M \Delta \tau_{ij,t-1}^m \quad (11)$$

where ρ is an evaporation rate of pheromone, which uniformly decreases all the pheromone values in order to prevent the algorithm from a rapid convergence toward suboptimal; $\Delta \tau_{ij,t-1}^m$ is the quantity of pheromone left on a path connecting position i and j by m th ant during the previous movement.

6. Termination:

Repeat steps (3)–(5) until stopping criterion are met, i.e., the number of function evaluation reaches user-defined maximum, the number of total number of ants movement cycle reaches user-defined maximum, or the fitness (objective function values) for all M ants are limited within function tolerance (i.e., convergence is reached).

The original ACO was initially invented for solving optimization problems with discrete decision variable domain, and for finding optimal combinations of components, such as the travel-sale-man problem. Socha and Dorigo (2008) further developed the original ACO to solve optimization problems with a continuous domain. According to the literature, there are many successful applications of the original ACO and its developed versions to solve different problems in various fields, such as soil hydraulic parameters calibration (Abbaspour et al. 2001), water quality (Bowden et al. 2002), optimal open channel design (Nourani et al. 2009), hydrological model calibration (Olarte and Obregon 2004), water distribution system design and planning (Maier et al. 2003; Wang and Guo 2010; Zecchin et al. 2003), and optimal reservoir operation (Kumar and Reddy 2006; Madadgar and Afshar 2009; Zecchin et al. 2012). Two most recent reviews of ACO and its applications can be found in Afshar et al. (2015) and Ostfeld (2011) for interested readers.

3.2.5 Shuffled Complex Evolution-UA (SCE-UA)

The SCE-UA algorithm is a global search algorithm, which combines a number of different strategies, including the Downhill Simplex, the Controlled Random Search, the Competitive Evolution, and the Complex Shuffling scheme (Duan et al. 1992, 1994). Extensive testing of the SCE-UA algorithm by numerous researchers has proven its effectiveness and efficiency in reliably finding the global solution, when a unique solution exists. The SCE-UA algorithm includes the following steps:

1. Initialization:

Generate parameter samples θ_i from the feasible parameter space. Calculate the objective function value of each sample $f(\theta_i)$. Set initial sample size $s = pm$, where p is the number of complexes and m is the number of points in each complex.

2. Rank samples:

Sort the s samples based on $f(\theta_i)$ from small to large values, i.e., $f(\theta_i) \leq f(\theta_{i+1})$.

3. Partitioning into complexes:

The s samples are partitioned into p complexes, such that the complex k includes samples of $\{\theta_k, \theta_{p+k}, \dots, \theta_{(m-1)*p+k}\}$ and function value of samples: $\{f(\theta_k), f(\theta_{p+k}), \dots, f(\theta_{(m-1)*p+k})\}$; $k = 1..m$ complexes.

4. Evolution of complexes:

Based on a trapezoidal probability distribution, in which higher probability is assigned to lower function values, select a subcomplex of q samples from each complex. Use downhill simplex algorithm to evolve the samples in each subcomplex.

5. Complex shuffling:

Include all samples from the subcomplex in the sample pool.

6. Termination:

Repeat steps (2)–(5) until the stopping criteria are reached.

SCE-UA has been extensively used in hydrologic modeling and has shown to be robust and efficient for hydrologic model calibration. Gan and Biftu (1996) applied

SCE-UA to multiple operational Conceptual Rainfall-Runoff (CRR) models and proved the effectiveness of SCE-UA scheme in calibrating different CRR models. Eckhardt and Arnold (2001) used SCE-UA algorithm to calibrate the parameters of the SWAT model over the Dietzhölze catchment in central Germany and reached high accuracy. Skahill and Doherty (2006) demonstrated the strengths of SCE-UA and other automatic parameter calibration scheme on an operational hydrological model for the Wildcat Creek watershed located in Kitsap County, Washington. Yang et al. (2008) demonstrated the effectiveness and efficiency of SCE-UA over other automatic parameter calibration schemes using the SWAT model over the Chaohe Basin in China. Ludwig et al. (2009) compared a fully distributed, a semidistributed, and a lumped hydrological model using SCE-UA as automatic calibration scheme over the Ammer basin in the Southern Bavaria, Germany. Khakbaz et al. (2012) employed SCE-UA for calibrating the SAC-SMA models over the Illinois River Basin at Siloam Springs, Arkansas, and produced satisfactory streamflow simulation. Liu et al. (2017) applied SCE-UA to calibrate an operational hydrological model used by Tibet Government for simulating the streamflows for the Upper Yellow and Upper Yangtze River basins of China. There are numerous other successful applications of SCE-UA algorithm in the field of hydrological model parameter calibration. Authors are only able to provide limited references. The original publication of SCE-UA ranked as the top three most cited articles in Water Resources Research (data collected in Sep 2016). However, when the dimension of a given problem increases extensively, the global convergence of SCE-UA algorithm might not be guaranteed. The population could collapse to a subspace of the full span of the parameter space, which impedes SCE-UA algorithm to exploit the parameter space (Chu et al. 2010). Chu et al. (2010, 2011) further improved the SCE-UA algorithm with a principal component analysis for a remedy of this issue. The enhanced version of SCE-UA is termed as the Shuffle Complex Evolution global optimization with Principal Component Analysis – University of California, Irvine (SP-UCI). In SP-UCI, principal component analysis (PCA) is used to detect the occurrence of population degeneration. The PC coordinate system is determined by the data samples. By adding new particles along the PC with zero (or relatively small) variance, the search is ensured to maintain the diversity of the entire population, especially along the collapsed dimension. The SP-UCI algorithm is also proven to be effective and efficient for many high dimensional real-world applications (Chu et al. 2014; Yang et al. 2015, 2017a).

3.3 Surrogate Modeling-Based Methods

The global optimization methods generally require up to tens of thousands of model runs to find the global optimal solution. This may place significant computational burden on solving such an optimization problem, if the underlying model requires a large amount of CPU time to run. One approach to reduce the computational burden is to approximate and replace the expensive simulation model with a cheaper-to-run surrogate model. Some fields also refer to the

surrogate modeling as function approximation, meta-modeling, response surface method, or model emulation (Blanning 1975; O'Hagan 2006). Once the surrogate model is constructed, a global optimization algorithm can be used to identify the optimal parameter set. These kinds of algorithms are called surrogate modeling based optimization methods (Simpson et al. 2001; Jin et al 2001; Queipo et al. 2005; Razavi et al. 2012).

A surrogate model can be understood as a “model of model.” It is a statistical model of the response surface of a simulation model. A surrogate model describes the relationship between inputs (i.e., model’s adjustable parameters) and outputs (i.e., the performance measure of the simulation model). Training an accurate surrogate model needs adequate input–output data, which are obtained by running the simulation model with different sets of parameters selected in the feasible parameter space. Previous studies use the “one-shot” approach (i.e., using a set of samples at once) to obtain input–output data to construct the surrogate model. This method directly establishes a surrogate model on the utilized data. Then it runs global optimization algorithm on the surrogate model. A high number of model runs may be required to ensure that the surrogate model represents the response surface of the original simulation model well. One way to economically construct a surrogate model for optimization is to use adaptive sampling. Adaptive sampling means that a certain number of points are sampled in the initial stage, and then a number of additional points, which can most effectively increase the accuracy of the surrogate model, are adaptively sampled and added to the initial sampling. For the purpose of finding an optimum, it is not necessary to map out the whole surface in a surrogate model exploration. An adaptive sampling strategy can quickly move the experiment to a region containing the optimum of the input variables. Only within this region is a thorough exploration of the surrogate model warranted to find the optimum (Wu and Hamada 2009). Below is the procedure of adaptive surrogate modeling based optimization algorithm:

1. Generate an initial experimental design spread over the entire input space and do the costly function evaluations at the points.
2. Use function evaluations to fit a statistical surrogate model for the objective function.
3. Use the surrogate model to predict the objective function values at unsampled points in the variable domain to decide at which point to do the next expensive function evaluation.
4. Do the expensive function evaluation at the point selected in Step 3. Then, use the new data point to update the surrogate model.
5. Iterate through steps (3)–(4) until the stopping criterion has been met. Take the final optimal value on the newly updated surrogate model as the global optimization result.

There are three main components of adaptive surrogate modeling-based optimization algorithms, named initial sampling, constructing the surrogate model, and adaptive sampling. Below we described these three steps more specifically.

1. Initial sampling

The initial sampling is the sampling plan in the design variable space, also called experimental design. It is a body of techniques that enable an investigator to conduct better experiments, analyze data efficiently, and make the connections between the conclusions from the analysis and the original objectives of the investigation (Wu and Hamada 2009). Generally, at this stage, the location of the points is only required to satisfy some space-filling criteria. Because of the absence of prior knowledge of the problem under consideration, uniformity of the design points throughout the domain is favorable. The optimum size of initial sample points is an open question. Some people think it should be ten times the number of dimensions (Jones et al. 1998). Others think it should keep the initial sample size to a minimum (for example: two times the number of dimensions) (Sóbester et al. 2005; Regis and Shoemaker 2007). However, the initial sample size is highly correlated with the initial sample methods. For Latin Hypercube method, too small sample size leads to a slow convergence for the optimization problem. For a more uniform low-discrepancy quasi-Monte Carlo method, a small sample size makes the better results (Wang et al. 2014).

2. Constructing the surrogate model

Generally, surrogate model construction methods are statistical regression methods that estimate response surface of a simulation model. A variety of approximation techniques have been developed and applied as the surrogates of an original simulation model: polynomial regression (Fen et al. 2009), regression tree method (Breiman et al. 1984; Yang et al. 2016), Random Forest method (Breiman 2001; Yang et al. 2017b), Multivariate Adaptive Regression Splines (Friedman 1991), Support Vector Machines (Zhang et al. 2009a), Artificial Neural Networks (Behzadian et al. 2009), and Gaussian Process (Rasmussen and Williams 2006; Snelson 2007). At the highest level, response surfaces can be differentiated based on whether they are noninterpolating (i.e., it minimizes the sum of squared errors from some predetermined functional form) or interpolating (i.e., it passes through all points). It has been suggested that noninterpolating surfaces, such as fitted quadratic surfaces, are unreliable for surrogate-based optimization because the surface may not sufficiently capture the shape of the function (Jones 2001). On the other hand, interpolating methods can get more and more accurate as new points are added, eventually converging to the true function.

3. Adaptive sampling

Adaptive sampling methods (also called sequential design methods) are iterative algorithms that use data acquired from previous iterations to guide future sample selections. The points we selected are also called infill samples (Sóbester et al. 2005). All of the aforementioned approaches select a sample point by optimizing an auxiliary function (minimize the bumpiness measure, maximize the expected

improvement, minimize the response surface), which is in general itself a global optimization problem. Adaptive sampling methods allow significant reduction in the number of simulations of the original simulation model because they only search the area that may contain the optimum of the input variables.

3.4 Deterministic Multiobjective Search Methods

The multiobjective optimization scheme can be referred as an application of single-objective optimization for handling multiple objectives (Deb 2001). A classical approach of solving a multiobjective optimization problem is to create a new composite function by giving individual weight to each single-objective function and adding them together. Then, the new weighted objective function is optimized using classical gradient-based methods or direct search schemes. This is a very straightforward approach. However, the fundamental differences between multiobjective optimization and single-objective optimization are ignored when using the transformation of multiple single-objective functions into a weighted composite function.

One of the fundamental differences between multiobjective optimization and single-objective is the existence of trade-offs among different competitive objective functions. In other words, for any multiobjective optimization problem, any gain with respect to the fitness of one objective function requires a sacrifice of the fitness in another objective function. This is due to the nature of constraints associated with any given multiobjective optimization problem. For example, conceptually a shopper can only get either item A or item B from a supermarket, because the total costs of both items A and B is beyond his/her fixed total budget. Under another circumstance, there are five different brands of the identical item ($A, A', A'', A''',$ and A'''') available with different qualities and costs (subject to the fact that a higher quality item is associated with a higher cost). The item quality that the shopper gets from item A and his/her remaining budget become two competitive objective functions. These two shopper's situations will be further explained with illustration in the following sections and Fig. 5. Many real-word problems are essentially balancing different objective functions and benefit gains. For instance, in California, USA, the water in the Northern California is diverted by the California State Water Project to multiple water demand sectors in the Central and Southern parts of California, including ecosystem, industry, resident, agriculture, etc. The amounts of water allocation to different sectors are conflicting objectives with the constraint of total available surface water. In another reservoir operation example, assuming water can be released through either spill gates or hydropower turbines, the turbine flows and spills become two competing objectives. This is because that once the water is spilled, it is not able to be retrieved for hydropower generation. The constraints include the risks of dam seepage, flooding of downstream areas, and other facility engineering requirements.

Given an optimization problem with two or more competitive objective functions, the trade-offs among those objective functions essentially mean any gain in the

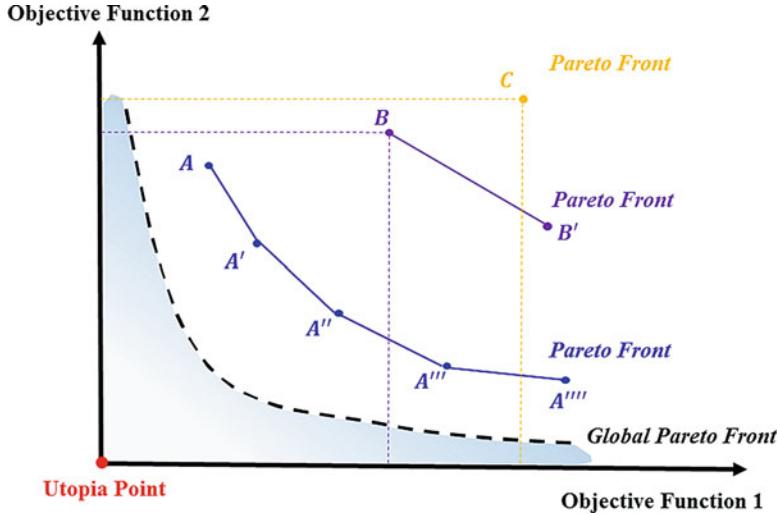


Fig. 5 An example for a two-objective minimization problem

fitness of one of the objective functions will call for the loss of fitness in at least one other objective functions of the optimization problem. Based on Eqs. (2) and (5), the mathematical expression of a multiobjective minimization problem with k objective functions is as follows:

$$\begin{cases} L_1(\theta_1 | \text{data}) = f_1(y_1, y_2, \dots, y_n | \theta_1) \\ L_2(\theta_2 | \text{data}) = f_2(y_1, y_2, \dots, y_n | \theta_2) \\ \vdots \\ L_k(\theta_k | \text{data}) = f_k(y_1, y_2, \dots, y_n | \theta_k) \end{cases} \quad (12)$$

$$\left\{ \begin{array}{l} \theta_1^* = \underset{\text{w.r.t } \theta_1}{\text{Maximize}} \ln L_1(\theta_1 | \text{data}) \cong \underset{\text{w.r.t } \theta_1}{\text{Minimize}} \left(\sum_{t=1}^N \epsilon_t^2 \right) \\ \theta_2^* = \underset{\text{w.r.t } \theta_2}{\text{Maximize}} \ln L_2(\theta_2 | \text{data}) \cong \underset{\text{w.r.t } \theta_2}{\text{Minimize}} \left(\sum_{t=1}^N \epsilon_t^2 \right) \\ \vdots \\ \theta_k^* = \underset{\text{w.r.t } \theta_k}{\text{Maximize}} \ln L_k(\theta_k | \text{data}) \cong \underset{\text{w.r.t } \theta_k}{\text{Minimize}} \left(\sum_{t=1}^N \epsilon_t^2 \right) \end{array} \right. \quad (13)$$

where $\theta(\theta_1, \theta_2, \dots, \theta_k)$ are subject to constraints $\Omega(\Omega_1, \Omega_2, \dots, \Omega_k)$, respectively.

Another difference between multi- and single-objective optimization is the notion of global optimality. In single-objective optimization, there is only one global optimum, while in the multiobjective context there exists multiple solutions that form a global optimal solution set, called *Global Pareto Optima* or *Global Pareto Front*. For example, in a two-objective minimization problem shown in Fig. 5, the x- and y-axis represent the first and second objective function, respectively. The origin represents

the minimization for all objectives without any constraint. However, due to imposed constraints, this point, called Utopia Point, as well as the light blue area in Fig. 5 are not feasible in reality. In the example of Fig. 5, the dashed line indicates the global optimal solutions of the conceptual two objective minimization problem, in which multiple global optimal solutions form a solution set, and this set of solutions is termed Global Pareto Optima or Global Pareto Front. For problems with higher number of dimensions, the *Global Pareto Optima* or *Global Pareto Front* can be a 3-D surface or a high-dimensional subspace. The actual shape and location of *Global Pareto Front* are dependent on the dimensionality, the conflicting characteristics of selected objective functions, and the parameter bounds.

In multiobjective optimization, the fitness of candidate solutions is defined based on the concepts of *Pareto Optimality* and *Pareto Front*, instead of simply using objective function values as solution fitness. For any multiobjective minimization problem, according to Deb (2001), a solution x is said to dominate the other solution x' , if (1) solution x is no worse than x' in all objectives, i.e., $\forall j \in \{1, 2, \dots, k\} : f_j(x) \leq f_j(x')$ and (2) solution x is strictly better than x' for at least one objective, i.e., $\exists j \in \{1, 2, \dots, k\} : f_j(x) < f_j(x')$. Among the eight solutions shown in Fig. 5, for instance, solution B dominates solution C , while solution B' does not dominate solution C . Similarly, both solutions A' and A''' dominate solution B and solution C . If there is no single solution dominating any others among a set of solutions, then this set of solutions forms a *Pareto Front* and is defined as a *nondominated solution set*. In Fig. 5, three different *Pareto Fronts* are defined with different colors, namely, the solution set $(A, A', A'', A''', A''''')$, (B, B') , and (C) . In each of the *nondominated solution set*, each individual solution is treated as equally important when comparing to others, i.e., solutions that belong to the same *Pareto Front* have equal fitness values even the associated objective function values can be different. Back to our shopper's examples, our shopper prefers item A than B because the fitness of item A is better than that of item B . Note that item A is associated with consistently smaller objective function values as compared to that with item B , i.e., item A is located on a *Pareto Front* that is closer to the *Global Pareto Optima* than item B . Furthermore, all five items $A, A', A'', A''',$ and A'''' are referred to the nondominated solutions, and are located on the same *Pareto Front* as shown in Fig. 5. The quality of our shopper gets from a single selection from items $A, A', A'', A''',$ and A'''' , and his/her remaining budget are the two objective functions to be optimized.

Theoretically, any single-objective optimization algorithm can be extended to solving multiobjective optimization if the fitness of population and updating rules are properly defined. For example, the GAs for multiobjective optimization (Deb 2001; Deb et al. 2000, 2002), the multiobjective PSO (Coello et al. 2004; Coello and Lechuga 2002), the multiobjective ACO (Alaya et al. 2007; Angus and Woodward 2009; Doerner et al. 2004; Gao et al. 2013), the multiobjective SA (Bandyopadhyay et al. 2008; Czyżak and Jaszkiewicz 1998; Serafini 1994; Suman 2004; Suppapitnarm et al. 2000), and different versions of SCE-UA algorithm for multi-objective optimization (Yang et al. 2015; Yapo et al. 1998) are all well developed and successful transformations from single-objective optimization algorithm to multi-objective searching schemes.

All kinds of multiobjective optimization schemes are very useful in real-world applications, by which the trade-offs among completing objectives can be analyzed and investigated. One of the commonly used multiobjective optimization algorithms is called the multiobjective complex evolution – University of Arizona (MOCOM-UA) global optimization method (Yapo et al. 1998). The MOCOM-UA method is one of the successors of the SCE-UA algorithm with a general purpose for solving global multiobjective optimization problems. The uses of MOCOM and other multiobjective optimization algorithms in the field of automatic hydrological model parameter calibration, and trade-off analysis are also extensive. Numbers of different variations of multiobjective algorithms and hydrological models have been tested, such as the studies conducted by Gupta et al. (1998), (2003), Madsen (2000, 2003), Tang et al. (2005), Bekele and Nicklow (2007), Hejazi et al. (2008), Moussa and Chahinian (2009), Shafii and Smedt (2009), Zhu et al. (2017), Zhang et al. (2010), Kollat et al. (2012), Sun et al. (2014), Reed et al. (2013), Asadzadeh et al. (2014), Yapo et al. (1998), and Vrugt et al. (2003b).

4 Examples of Hydrological Applications

Different optimization algorithms have their own strengths and limitations. In practical uses, it is suggested that users choose the most proper algorithm that meets the calibration requirements. However, the selection of algorithm could be tedious for some cases. In this section, we briefly introduce the practical uses of three different optimization algorithms and demonstrate the strengths of (1) the SCE-UA (Duan et al. 1992) algorithm for its global convergence; (2) a surrogate modeling scheme (ASMO) (Wang et al. 2014) and its improvements of computational efficiency; and (3) a multiobjective optimization scheme, termed MOSPD (Yang et al. 2015, 2017c), for its effectiveness of producing *Pareto Optimality*.

The case study is carried on a real-world application, the Sacramento Soil Moisture Accounting Model (SAC-SMA) model. The SAC-SMA model is a conceptual rainfall-runoff model that represents the soil column with upper and lower zones of multiple storages (Burnash 1995). It has been used extensively in both research and operational applications for river forecasting by the National Weather Service River Forecast Centers across the United States. According to literature, the SAC-SMA model is also one of the benchmark models in testing the performances of different automatic parameter calibration methods (Duan et al. 1992, 1994; Yapo et al. 1998; Chu et al. 2014). Figure 6 shows the structure of the SAC-SMA model. There are 16 parameters in the SAC-SMA model. We consider only 13 of them as adjustable parameters, whose feasible ranges and descriptions are listed in Table 1. Three parameters RSERV, RIVA, and SIDE are fixed at prespecified values according to Brazil (1988).

The study area is the South Branch Potomac River basin near Springfield, West Virginia, USA. It is one of the 12 experimental watersheds of the Model Parameter Estimation Experiment (MOPEX) (Duan et al. 2006). The total drainage area of the basin (U.S. Geological Survey Station No. 01608500) is about 3810 km². Historical precipitation, potential evapotranspiration, and streamflow observations from

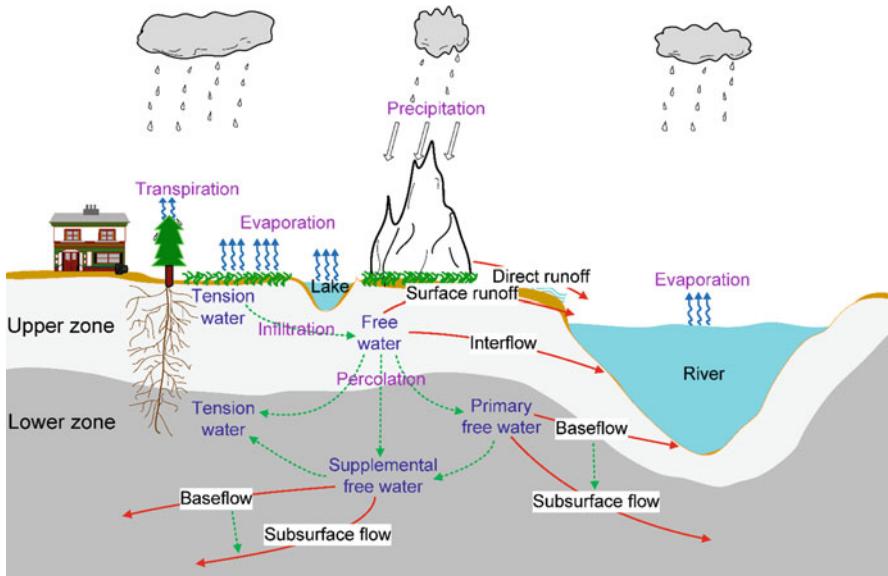


Fig. 6 A schematic of the SAC-SMA model (Source: Gan et al. 2014)

January 1, 1960 to December 31, 1979 were obtained from the MOPEX database for this study. The annual average precipitation over this period is 1021 mm, annual average potential evapotranspiration is 762 mm, and annual average streamflow discharge is $39.5 \text{ m}^3/\text{s}$. The hydrological simulations are run with a 6-h time step for each combination of model tunable parameters. Additional physical characteristics of the study area were presented by Duan et al. (2006). The purpose of a model calibration (i.e., parameter optimization) is to find the optimal parameter set for the SAC-SMA model such that the simulated streamflow would have the best overall match with the observed streamflow through minimizing the objective function value. In this study, we used the root mean square error (RMSE) as the objective function, which is defined as:

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{t=1}^n (\mathcal{Q}_{s,t} - \mathcal{Q}_{o,t})^2}$$

where t is the number of time steps, $\mathcal{Q}_{s,t}$ is the simulated flow for time step t , and $\mathcal{Q}_{o,t}$ is the observed flow for time step t . In the model simulation process, we used a period of 3 months to remove any impact of uncertain initial conditions, i.e., in computing RMSE, the streamflow values for the first 3 months of 1960 were intentionally removed from consideration in order to warm up the model.

A comparison of calibration performance of SCE-UA algorithm against the multistart downhill simplex (MSDS) algorithms (Duan et al. 1994) is shown in Figs. 7 and 8 for 4 of the 13 parameters in SAC-SMA model.

Table 1 The 13 parameters of the SAC-SMA model and their feasible ranges

No.	Parameter	Description	Range
1	UZTWM	Upper zone tension water maximum storage (mm)	[10, 300]
2	UZFWM	Upper zone free water maximum storage (mm)	[5, 150]
3	UZK	Upper zone free water lateral drainage rate (day^{-1})	[0.1, 0.75]
4	PCTIM	Impervious fraction of the watershed area (decimal fraction)	[0, 0.1]
5	ADIMP	Additional impervious area (decimal fraction)	[0, 0.2]
6	ZPERC	Maximum percolation rate (dimensionless)	[5, 350]
7	REXP	Exponent of the percolation equation (dimensionless)	[1, 5]
8	LZTWM	Lower zone tension water maximum storage (mm)	[10, 500]
9	LZFSM	Lower zone supplemental free water maximum storage (mm)	[5, 400]
10	LZFPM	Lower zone primary free water maximum storage (mm)	[10, 1000]
11	LZSK	Lower zone supplemental free water lateral drainage rate (day^{-1})	[0.01, 0.35]
12	LZPK	Lower zone primary free water lateral drainage rate (day^{-1})	[0.001, 0.05]
13	PFREE	Fraction of water percolating from upper zone directly to lower zone free water (decimal fraction)	[0.0, 0.9]
14	RIVA	Riverside vegetation area (decimal fraction)	0.3
15	SIDE	Ration of deep recharge to channel base flow (dimensionless)	0
16	RSERV	Fraction of lower zone free water not transferrable to lower zone tension water (decimal fraction)	0

As shown in Fig. 7, for 100 independent model runs, the parameters tuned by the MSDS algorithm all fail to converge to a single value at the end of evolution, which indicate the optimal solutions derived by the MSDS algorithm are local optima. On the contrary, all parameters with the SCE-UA algorithm are able to converge to consistent values (Fig. 8), suggesting the detection of global optimum for only ten independent model runs.

The SCE-UA algorithm is famous for its global convergence when the number of evaluation is not limited. However, it generally requires up to tens of thousands of model runs to find the global optimal solution. This may place severe computational constraint on solving such an optimization problem, if the underlying model requires a large amount of CPU time to run. For this situation, the surrogate modeling-based optimization algorithm (e.g., ASMO) is a good choice. Figure 9 shows the comparison between SCE-UA and ASMO applied to the 13 parameters SAC-SMA model. In this study, we set the number of complexes = 4 for SCE-UA and set maximum model evaluations = 200 for ASMO. From Fig. 9, we note that, for SCE-UA, after 2606 model evaluations, the optimization search converges to its optimal solution with an objective function value of 0.92722. For ASMO, the optimized objective function value is 0.92895424449 after 200 model runs. From this case, we conclude that ASMO has an obvious advantage in convergence speed over SCE-UA, with the former needing about 200 total sample points and the latter needing close to 1900

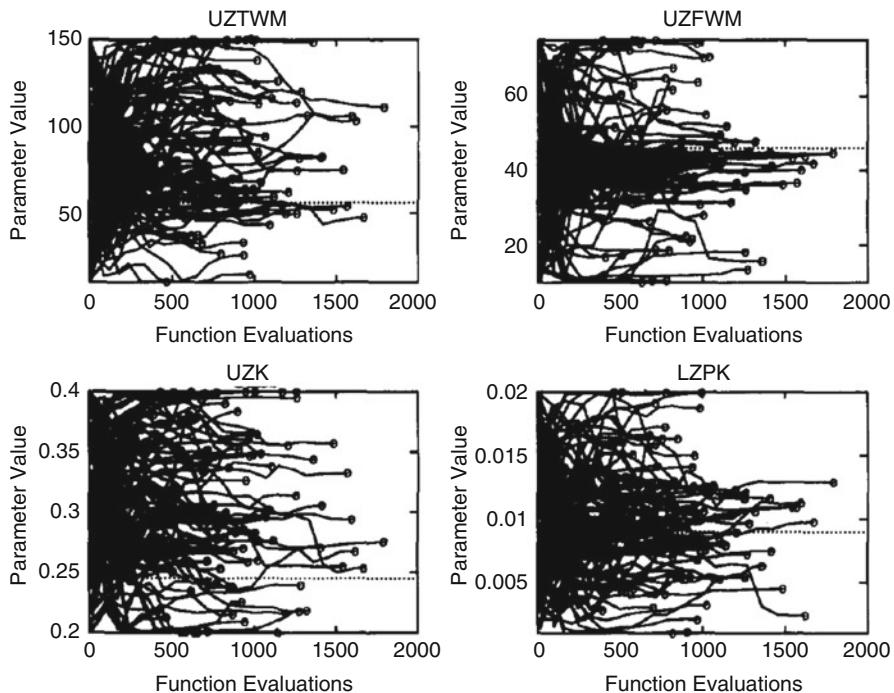


Fig. 7 MSDS search and convergence behavior for four SAC-SMA parameters (100 independent trails). (Source: Duan et al. 1994)

sample points to reach similar objective function value. The SCE-UA method possesses an edge over the surrogate-based optimization in converging to the “true” optimal parameter set if there is no limit on the number of sample points. In other words, the SCE-UA algorithm is capable of finding the exact “true” parameter values, while the ASMO can provide approximate optimization results with relatively small costs. There are also different types of ASMO algorithm available, for instance, the Multiobjective ASMO (Gong et al. 2016), and distribution-based parameter estimation with surrogate model ASMO-PODE (Gong and Duan 2017) for interested readers.

The parameter calibration in a multiobjective context is different from that in the single-objective framework as demonstrated in the previous examples of SCE-UA and AMSO algorithm. In the multiobjective optimization framework, the ultimate goal is to produce *nondominated solutions* that match *Global Pareto Optima* or *Global Pareto Front* in the objective function space. The less differences between the *nondominated solutions* and the solutions located on the *Global Pareto Optima* or *Global Pareto Front*, the better fitness the *nondominated solutions* have. In reality, to define many objective functions, which are completely competing to each other, is a tedious task for a given problem, and the *Global Pareto Front* for most of the cases cannot be explicitly formulated or even does not exist.

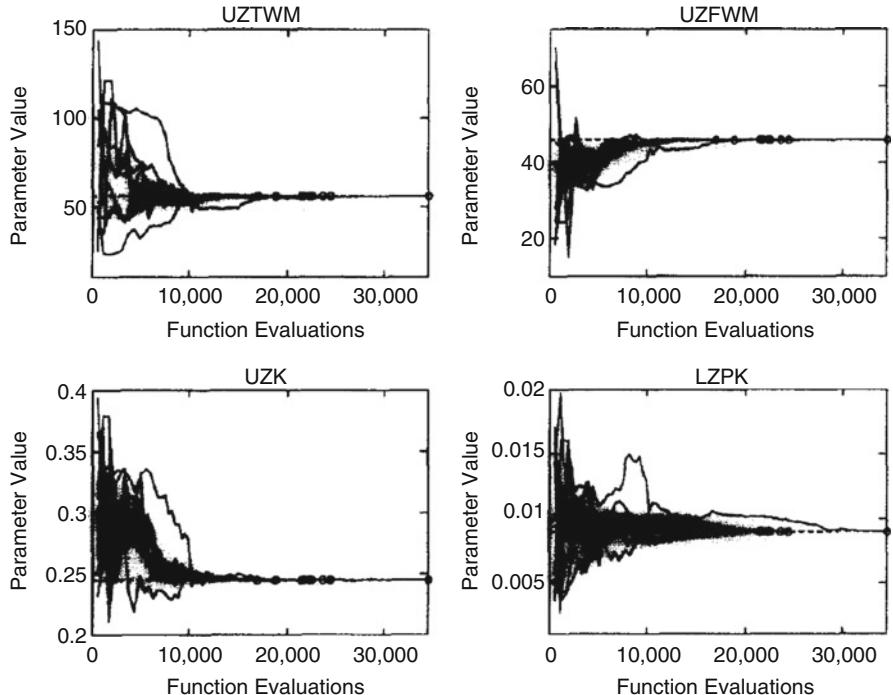


Fig. 8 SCE-UA search and convergence behavior for four SAC-SMA parameters (10 independent trials). (Source: Duan et al. 1994)

Therefore, many human-built, benchmark, conceptual test functions are used in both literature and real-world application to evaluate the performance of multiobjective optimization algorithms. In the following sections, we demonstrate the strength of a Multiobjective Shuffle Complex Evolution with Principal Component Analysis and Crowding Distance (MOSPD) (Yang et al. 2015) on two conceptual, benchmark, composite test functions with known *Global Pareto Front*. The two test functions are KUR function (Kursawe 1991) and ZDT1 function (Zitzler et al. 2000; Zitzler and Thiele 1999), which are both commonly used test functions in the literature. The detailed objective functions, dimensionality, parameter bounds, and characteristics are listed in Table 2. The total number of individuals in population is set to 124 for all simulation. The simulation results (red dots), along with the known *Global Pareto Front* (black line), are shown in Fig. 10a, b. The population in the objective function space during the entire evolution is shown in Fig. 10c, d with different color legends represents the locations of population during the evolution.

In the multiobjective context, a higher number of parameters will not only increase the dimensionality for any single-objective function, but also result in a more complex shape of *Global Pareto Front* in the objective functions' space. As shown in Fig. 10a, b, the final nondominated solutions derived by MOSPD generally match well with the known *Global Pareto Front* for both KUR and ZDT1 functions.

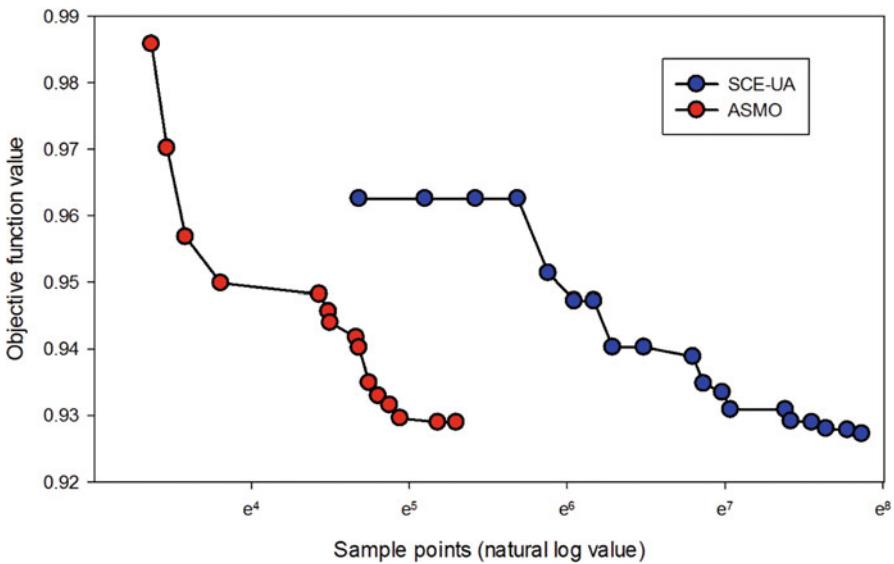


Fig. 9 Comparison of optimization results of SCE-UA and ASMO

Table 2 Details on test functions, including name, number of parameters (N), range, objective functions, optimal solutions, and shapes

Name	N	Variable range	Objective functions	Optimal solutions	Shape
KUR	3	[-5,5]	$f_1(x) = \sum_{i=1}^{n-1} (-10\exp(\sqrt{x_i^2 + x_{i+1}^2}))$	$x_1 \in [0, 1]$	Convex, disconnected
			$f_2(x) = \sum_{i=1}^n (x_i ^{0.8} + 5 \sin(x_i^3))$	$x_i = 0,$ $i = 2, \dots, n$	
ZDT1	30	[0,1]	$f_1(x) = x_1$	$x_1 \in [0, 1]$	Convex
			$f_2(x) = g(x)[1 - \sqrt{x_1/g(x)}]$	$x_i = 0,$ $i = 2, \dots, n$	
			$g(x) = 1 + 9(\sum_{i=2}^n x_i)/(n - 1))$	$i = 2, \dots, n$	

Different from the previous examples with single-objective function, the population evolution during the entire search (Fig. 10c, d) is gradually toward the *Global Pareto Front*, instead of optimizing any single-objective function value in a consistently decreasing pattern. According to Deb (2001), the evaluation criteria of non-dominated solutions for any algorithm have to take two aspects into consideration: (1) the closeness of *nondominated solutions* toward the *Global Pareto Front*, i.e., the convergence of solutions; and (2) the spread of *nondominated solutions* along the *Global Pareto Front*, i.e., the diversity of solutions that represents the coverage of the *nondominated solutions* on extreme values of objective functions. In the demonstrated examples (Fig. 10a, b), the produced *nondominated solutions* have satisfactory performances with regard to both convergence and diversity on the KUR and ZDT1 functions. More examples of different heuristic multiobjective optimization

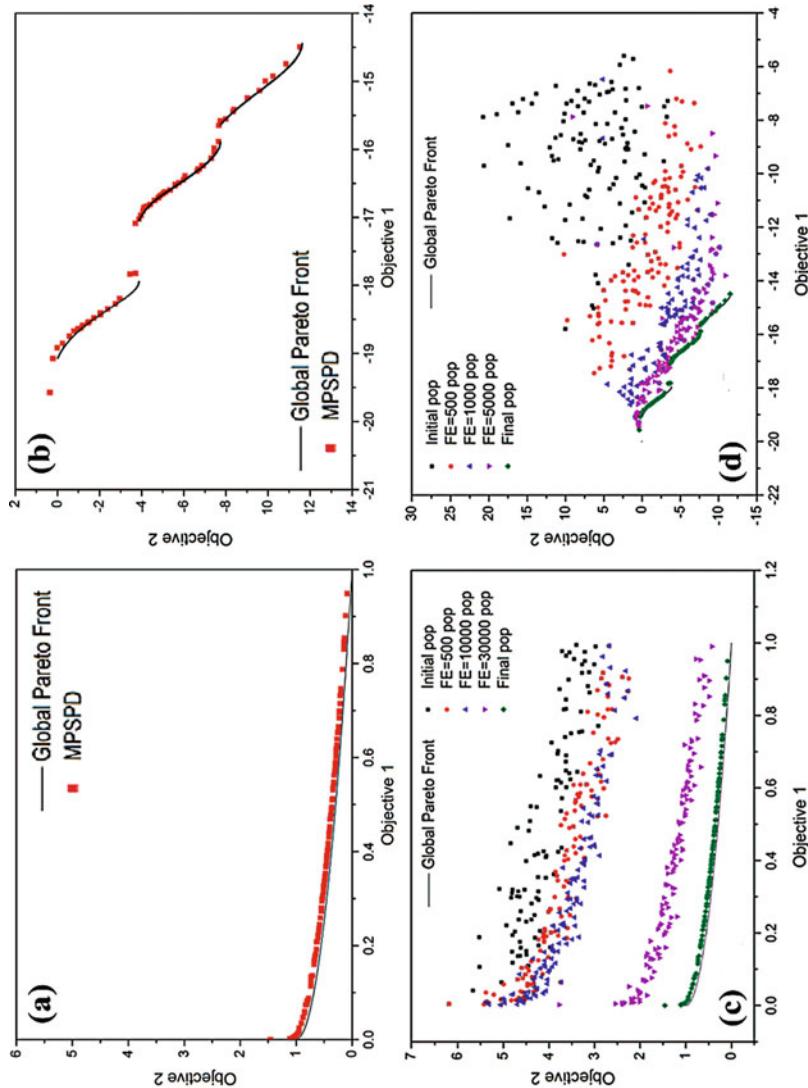


Fig. 10 The final nondominated solutions and Global Pareto Front for (a) KUR function and (b) ZDT1 function, and the evolution process of MOSPD on (c) KUR function and (d) ZDT1 function. Note: the acronym “pop” in the figure legend refers to the number of individuals in a population

algorithms on a large-scale tests are provided in Zitzler et al. (2000), Zitzler and Thiele (1999), Yang et al. (2015), Gong et al. (2016), and Gong and Duan (2017) for interested readers.

5 Summary and Conclusion

As a summary of this chapter, many types of deterministic optimization algorithms have been developed over the past decades, which have been useful for different types of optimization problems in general. With a limited amount of content of this chapter, authors only review few of the most popular methods in the field of hydrological model calibration, and there exists many other effective algorithms, which may fit better to various real-world optimization problems. It is also a fact that the development of optimization algorithm in the research community is rapid, and the number of new optimization algorithms has been increasing as time goes. However, it is worth mentioning that even the development of optimization algorithm has been prosperous over the years; all algorithms obey the No-Free-Lunch algorithm (Wolpert and Macready 1997). In other words, different algorithms have their own strengths and limitations regarding the efficiency, effectiveness, suitability, etc. to a particular problem. To reduce the computational burden of calibrating a complex model, one of the approaches is to approximate and replace the expensive simulation model with a cheaper-to-run surrogate model. On the other hand, if the computational resource is not a concern by users, many recently developed paralleled computing techniques and hybrid optimization approaches are also promising. Some challenges and future directions with respects to the development and applications of evolutionary algorithms can be found in a recent review paper by Maier et al. (2014).

References

- K. Abbaspour, R. Schulin, M.T. Van Genuchten, Estimating unsaturated soil hydraulic parameters using ant colony optimization. *Adv. Water Resour.* **24**(8), 827–841 (2001)
- M.A. Abido, Optimal design of power-system stabilizers using particle swarm optimization. *IEEE Trans. Energy Convers.* **17**(3), 406–413 (2002)
- A. Afshar, F. Massoumi, A. Afshar, M.A. Mariño, State of the art review of ant colony optimization applications in water resource management. *Water Resour. Manag.* **29**(11), 3891–3904 (2015)
- I. Alaya, C. Solnon, K. Ghedira, *Ant Colony Optimization for Multi-objective Optimization Problems* (Citeseer, Patras, 2007), pp. 450–457. <https://doi.org/10.1109/ICTAI.2007.108>
- D. Angus, C. Woodward, Multiple objective ant colony optimisation. *Swarm Intell.* **3**(1), 69–85 (2009)
- R. Arsenault, A. Poulin, P. Côté, F. Brissette, Comparison of stochastic optimization algorithms in hydrological model calibration. *J. Hydrol. Eng.* **19**(7), 1374–1384 (2013)
- R. Arsenault, A. Poulin, P. Côté, F. Brissette, Comparison of stochastic optimization algorithms in hydrological model calibration. *J. Hydrol. Eng.* **19**(7), 1374–1384 (2014)

- M. Asadzadeh, B.A. Tolson, D.H. Burn, A new selection metric for multiobjective hydrologic model calibration. *Water Resour. Res.* **50**(9), 7082–7099 (2014)
- V. Babovic, M. Keijzer, Rainfall runoff modelling based on genetic programming. *Hydrol. Res.* **33**(5), 331–346 (2002)
- C. Balascio, D. Palmeri, H. Gao, Use of a genetic algorithm and multi-objective programming for calibration of a hydrologic model. *Trans. ASAE* **41**(3), 615 (1998)
- S. Bandyopadhyay, S. Saha, U. Maulik, K. Deb, A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE Trans. Evol. Comput.* **12**(3), 269–283 (2008)
- A. Bárdossy, T. Das, Influence of rainfall observation network on model calibration and application. *Hydrol. Earth Syst. Sci. Discuss.* **3**(6), 3691–3726 (2006)
- B. Bates, Calibration of the SFB model using a simulated annealing approach. *Water Down Under 94: Surface Hydrology and Water Resources Papers; Preprints of Papers*, 1 (1994)
- K. Behzadian, Z. Kapelan, D. Savic, A. Ardeshir, Stochastic sampling design using a multi-objective genetic algorithm and adaptive neural networks. *Environ. Model. Softw.* **24**(4), 530–541 (2009)
- E.G. Bekele, J.W. Nicklow, Multi-objective automatic calibration of SWAT using NSGA-II. *J. Hydrol.* **341**(3), 165–176 (2007)
- R.W. Blanning, Construction and implementation of metamodels. *Simulation* **24**(6), 177–184 (1975)
- G. Bowden, G. Dandy, H. Maier, Ant colony optimisation of a general regression neural network for forecasting water quality, in *Hydroinformatics 2002: Proceedings of the FIFTH INTERNATIONAL Conference on Hydroinformatics*, ed. by R.A. Falconer et al., Cardiff (IWA Publishing, 2002), pp. 692–698
- L.E. Brazil, *Multilevel Calibration Strategy for Complex Hydrologic Simulation Models* (Colorado State University, Fort Collins, 1988)
- L. Breiman, Random forests. *Mach. Learn.* **45**(1), 5–32 (2001). <https://doi.org/10.1023/A.1010933404324>
- L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees* (Wadsworth, Belmone, 1984)
- R.J.C. Burnash, The NWS river forecast system: Catchment modeling, in *Computer Models of Watershed Hydrology*, ed. by V.P. Singh (Water Resources Publications, Highlands Ranch, 1995), pp. 311–366
- V. Černý, Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *J. Optim. Theory Appl.* **45**(1), 41–51 (1985)
- K. Chau, A split-step particle swarm optimization algorithm in river stage forecasting. *J. Hydrol.* **346**(3), 131–135 (2007)
- K. Chau, Application of a particle swarm optimization algorithm to hydrological problems, in *Water Resources Research Progress*, (Nova Science Publishers, New York, 2008), pp. 3–12
- C.-T. Cheng, M.-Y. Zhao, K. Chau, X.-Y. Wu, Using genetic algorithm and TOPSIS for Xinanjiang model calibration with a single procedure. *J. Hydrol.* **316**(1), 129–140 (2006)
- C.L. Chiu, J. Huang, Nonlinear time varying model of rainfall-runoff relation. *Water Resour. Res.* **6**(5), 1277–1286 (1970)
- W. Chu, X. Gao, S. Sorooshian, Improving the shuffled complex evolution scheme for optimization of complex nonlinear hydrological systems: Application to the calibration of the Sacramento soil-moisture accounting model. *Water Resour. Res.* **46**(9), W09530 (2010)
- W. Chu, X. Gao, S. Sorooshian, A new evolutionary search strategy for global optimization of high-dimensional problems. *Inf. Sci.* **181**(22), 4909–4927 (2011)
- W. Chu, T. Yang, X. Gao, Comment on “High-dimensional posterior exploration of hydrologic models using multiple-try DREAM (ZS) and high-performance computing” by Eric Laloy and Jasper A. Vrugt. *Water Resour. Res.* **50**(3), 2775–2780 (2014)
- C.C. Coello, M.S. Lechuga, *MOPSO: A Proposal for Multiple Objective Particle Swarm Optimization* (IEEE, Honolulu, 2002), pp. 1051–1056

- C.A.C. Coello, G.T. Pulido, M.S. Lechuga, Handling multiple objectives with particle swarm optimization. *IEEE Trans. Evol. Comput.* **8**(3), 256–279 (2004)
- P. Czyżak, A. Jaszkiewicz, Pareto simulated annealing – A metaheuristic technique for multiple-objective combinatorial optimization. *J. Multi-Criteria Decis. Anal.* **7**(1), 34–47 (1998)
- K. Deb, *Multi-objective Optimization Using Evolutionary Algorithms* (Wiley, Chichester, 2001)
- K. Deb, S. Agrawal, A. Pratap, T. Meyarivan, *A Fast Elitist Non-dominated Sorting Genetic Algorithm for Multi-objective Optimization: NSGA-II* (Springer, Berlin, 2000), pp. 849–858
- K. Deb, A. Pratap, S. Agarwal, T. Meyarivan, A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* **6**(2), 182–197 (2002)
- J.-L. Deneubourg, J.M. Pasteels, J.-C. Verhaeghe, Probabilistic behaviour in ants: A strategy of errors? *J. Theor. Biol.* **105**(2), 259–271 (1983)
- J.-L. Deneubourg, S. Aron, S. Goss, J.M. Pasteels, The self-organizing exploratory pattern of the argentine ant. *J. Insect Behav.* **3**(2), 159–168 (1990)
- K. Doerner, W.J. Gutjahr, R.F. Hartl, C. Strauss, C. Stummer, Pareto ant colony optimization: A metaheuristic approach to multiobjective portfolio selection. *Ann. Oper. Res.* **131**(1–4), 79–99 (2004)
- M. Dorigo, Optimization, learning and natural algorithms. Ph.D. Thesis, Politecnico di Milano (in Italian) 1992
- M. Dorigo, C. Blum, Ant colony optimization theory: A survey. *Theor. Comput. Sci.* **344**(2), 243–278 (2005)
- M. Dorigo, T. Stützle, *Ant Colony Optimization: Overview and Recent Advances*. Techreport, IRIDIA, Université Libre de Bruxelles (2009)
- M. Dorigo, V. Maniezzo, A. Colorni, Ant system: Optimization by a colony of cooperating agents. *IEEE Trans. Syst. Man Cybern. B Cybern.* **26**(1), 29–41 (1996)
- M. Dorigo, M. Birattari, T. Stützle, Ant colony optimization. *IEEE Comput. Intell. Mag.* **1**(4), 28–39 (2006)
- Q. Duan, S. Sorooshian, H.V. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **28**, 1015 (1992)
- Q. Duan, S. Sorooshian, V.K. Gupta, Optimal use of the SCE-UA global optimization method for calibrating watershed models. *J. Hydrol.* **158**, 265 (1994)
- Q. Duan, J. Schaake, V. Andreassian, S. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan, L. Oudin, S. Sorooshian, T. Wagener, E.F. Wood, Model parameter estimation experiment (MOPEX): An overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320**(1–2), 3–17 (2006)
- R.C. Eberhart, J. Kennedy, A new optimizer using particle swarm theory, in *Micro Machine and Human Science, 1995, MHS '95. Proceedings of the Sixth International Symposium on*, Nagoya, 4–6 October 1995 (IEEE, New York, 1995), pp. 39–43. <https://doi.org/10.1109/MHS.1995.494215>
- R. Eglese, Simulated annealing: A tool for operational research. *Eur. J. Oper. Res.* **46**(3), 271–281 (1990)
- C. Fen, C. Chan, H. Cheng, Assessing a response surface-based optimization approach for soil vapor extraction system design. *J. Water Resour. Plann. Manag.* **135**(3), 198–207 (2009)
- F. Francés, J.I. Vélez, J.J. Vélez, Split-parameter structure for the automatic calibration of distributed hydrological models. *J. Hydrol.* **332**(1), 226–240 (2007)
- M. Franchini, Use of a genetic algorithm combined with a local search method for the automatic calibration of conceptual rainfall-runoff models. *Hydrol. Sci. J.* **41**(1), 21–39 (1996)
- M. Franchini, G. Galeati, Comparing several genetic algorithm schemes for the calibration of conceptual rainfall-runoff models. *Hydrol. Sci. J.* **42**(3), 357–379 (1997)
- J. Friedman, Multivariate adaptive regression splines. *Ann. Stat.* **19**(1), 1–67 (1991)
- T.Y. Gan, G.F. Biftu, Automatic calibration of conceptual rainfall-runoff models: Optimization algorithms, catchment conditions, and model structure. *Water Resour. Res.* **32**(12), 3513–3524 (1996)

- Y. Gan, Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao, Z. Di, A comprehensive evaluation of various sensitivity analysis methods: A case study with a hydrological model. *Environ. Model. Softw.* **51**, 269–285 (2014)
- Y. Gao, H. Guan, Z. Qi, Y. Hou, L. Liu, A multi-objective ant colony system algorithm for virtual machine placement in cloud computing. *J. Comput. Syst. Sci.* **79**(8), 1230–1242 (2013)
- M.K. Gill, Y.H. Kaheil, A. Khalil, M. McKee, L. Bastidas, Multiobjective particle swarm optimization for parameter estimation in hydrology. *Water Resour. Res.* **42**(7), 417–431 (2006). <https://doi.org/10.1029/2005WR004528>
- D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning* (Addison Wesley, Estados Unidos, 1989), p. 102
- W. Gong, Q. Duan, An adaptive surrogate modeling-based sampling strategy for parameter optimization and distribution estimation (ASMO-PODE). *Environ. Model. Softw.* **95**, 61–75 (2017)
- W. Gong, Q. Duan, J. Li, C. Wang, Z. Di, A. Ye, C. Miao, Y. Dai, Multiobjective adaptive surrogate modeling-based optimization for parameter estimation of large, complex geophysical models. *Water Resour. Res.* **52**(3), 1984–2008 (2016)
- V. Granville, M. Krivánek, J.-P. Rasson, Simulated annealing: A proof of convergence. *IEEE Trans. Pattern Anal. Mach. Intell.* **16**(6), 652–656 (1994)
- H.V. Gupta, S. Sorooshian, P.O. Yapo, Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. *Water Resour. Res.* **34**(4), 751–763 (1998)
- H.V. Gupta, S. Sorooshian, T.S. Hogue, D.P. Boyle, Advances in automatic calibration of watershed models, in *Calibration of Watershed Models*, (American Geophysical Union, Washington, DC, 2003), pp. 9–28
- W.K. Hastings, Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**(1), 97–109 (1970)
- M.I. Hejazi, X. Cai, D.K. Borah, Calibrating a watershed simulation model involving human interference: An application of multi-objective genetic algorithms. *J. Hydroinf.* **10**(1), 97–111 (2008)
- J.H. Holland, *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence* (University of Michigan Press, Ann Arbor, 1975)
- R. Jin, W. Chen, T.W. Simpson, Comparative studies of metamodeling techniques under multiple modeling criteria. *Struct. Multidisc. Optim.* **23**, 1–13 (2001)
- D. Jones, A taxonomy of global optimization methods based on response surfaces. *J. Glob. Optim.* **21**, 345–383 (2001)
- D. Jones, M. Schonlau, W. Welch, Efficient global optimization of expensive black-box functions. *J. Glob. Optim.* **13**(4), 455–492 (1998)
- B. Kamali, S.J. Mousavi, K.C. Abbaspour, Automatic calibration of HEC-HMS using single-objective and multi-objective PSO algorithms. *Hydrol. Process.* **27**(26), 4028–4042 (2013)
- J. Kennedy, *Encyclopedia of Machine Learning* (Springer, Berlin, 2011), pp. 760–766
- J. Kennedy, J.F. Kennedy, R.C. Eberhart, Y. Shi, *Swarm Intelligence* (Morgan Kaufmann, San Francisco, 2001)
- B. Khakbaz, B. Imam, K. Hsu, S. Sorooshian, From lumped to distributed via semi-distributed: Calibration strategies for semi-distributed hydrologic models. *J. Hydrol.* **418**, 61–77 (2012)
- S. Kirkpatrick, Optimization by simulated annealing: Quantitative studies. *J. Stat. Phys.* **34**(5–6), 975–986 (1984)
- P.K. Kitanidis, R.L. Bras, Real-time forecasting with a conceptual hydrologic model: 2. Applications and results. *Water Resour. Res.* **16**(6), 1034–1044 (1980)
- J. Kollat, P. Reed, T. Wagener, When are multiobjective calibration trade-offs in hydrologic models meaningful? *Water Resour. Res.* **48**(3), 520–539 (2012). <https://doi.org/10.1029/2011WR011534>

- V. Kulandaiswamy, C. Subramanian, A nonlinear approach to runoff studies, in *Proceedings of the International Hydrology Symposium*, vol. 1, (Colorado State University, Fort Collins, 1967), pp. 72–79
- D.N. Kumar, M.J. Reddy, Ant colony optimization for multi-purpose reservoir operation. *Water Resour. Manag.* **20**(6), 879–898 (2006)
- C. Kuok, C.P. Chan, Particle swarm optimization for calibrating and optimizing Xinanjiang model parameters. *Int. J. Adv. Sci. Appl.* **3**, 115 (2012)
- F. Kursawe, Parallel Problem Solving from Nature: 1st Workshop, PPSN I Dortmund, FRG, October 1–3, 1990 Proceedings, ed. by H.-P. Schwefel, R. Männer (Springer Berlin Heidelberg, Berlin, 1991), pp. 193–197
- G.-F. Lin, C.-M. Wang, A nonlinear rainfall–runoff model embedded with an automated calibration method – Part 2: The automated calibration method. *J. Hydrol.* **341**(3–4), 196–206 (2007)
- S.Y. Liang, T.R. Gautam, S.T. Khu, V. Babovic, M. Keijzer, N. Muttgil, Genetic programming: a new paradigm in rainfall runoff modeling. *J. Am. Water Resour. Assoc.* **38**(3), 705–718 (2002)
- X. Liu, T. Yang, K. Hsu, C. Liu, S. Sorooshian, Evaluating the streamflow simulation capability of PERSIANN-CDR daily rainfall products in two river basins on the Tibetan plateau. *Hydrol. Earth Syst. Sci.* **21**(1), 169 (2017)
- H. Lü, T. Hou, R. Horton, Y. Zhu, X. Chen, Y. Jia, W. Wang, X. Fu, The streamflow estimation using the Xinanjiang rainfall runoff model and dual state-parameter estimation method. *J. Hydrol.* **480**, 102–114 (2013)
- R. Ludwig, I. May, R. Turcotte, L. Vescovi, M. Braun, J.-F. Cyr, L.-G. Fortin, D. Chaumont, S. Biner, I. Chartier, The role of hydrological model complexity and uncertainty in climate change impact assessment. *Adv. Geosci.* **21**, 63–71 (2009)
- S. Madadgar, A. Afshar, An improved continuous ant algorithm for optimization of water resources problems. *Water Resour. Manag.* **23**(10), 2119–2139 (2009)
- H. Madsen, Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. *J. Hydrol.* **235**(3), 276–288 (2000)
- H. Madsen, Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. *Adv. Water Resour.* **26**(2), 205–216 (2003)
- H. Madsen, G. Wilson, H.C. Ammentorp, Comparison of different automated strategies for calibration of rainfall-runoff models. *J. Hydrol.* **261**(1), 48–59 (2002)
- H.R. Maier, A.R. Simpson, A.C. Zecchin, W.K. Foong, K.Y. Phang, H.Y. Seah, C.L. Tan, Ant colony optimization for design of water distribution systems. *J. Water Resour. Plan. Manag.* **129**(3), 200–209 (2003)
- H.R. Maier, Z. Kapelan, J. Kasprzyk, J. Kollat, L.S. Matott, M. Cunha, G.C. Dandy, M.S. Gibbs, E. Keedwell, A. Marchi, Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. *Environ. Model Softw.* **62**, 271–299 (2014)
- R. Moussa, N. Chahinian, Comparison of different multi-objective calibration criteria using a conceptual rainfall-runoff model of flood events. *Hydrol. Earth Syst. Sci.* **13**(4), 519–535 (2009)
- J.A. Nelder, R. Mead, A simplex method for function minimization. *Comput. J.* **7**(4), 308–313 (1965)
- V. Nourani, S. Talatahari, P. Monadjemi, S. Shahradfar, Application of ant colony optimization to optimal design of open channels. *J. Hydraul. Res.* **47**(5), 656–665 (2009)
- A. O'Hagan, Bayesian analysis of computer code outputs: a tutorial. *Reliab. Eng. Syst. Saf.* **91**(10–11), 1290–1300 (2006)
- R.E. Olarte, N. Obregon, Comparison between a simple GA and an ant system for the calibration of a rainfall-runoff model, in *6th International Conference on Hydroinformatics* (in 2 volumes, with CD-ROM) (World Scientific Publishing Company, Singapore, 2004), pp. 842–849, ISBN 981-238-787-0
- A. Ostfeld, Ant colony optimization for water resources systems analysis—Review and challenges, in *Ant Colony Optimization Methods and Applications* (Technion Israel Institute of Technology, Israel, 2011), p. 147

- M.A. Panduro, C.A. Brizuela, L.I. Balderas, D.A. Acosta, A comparison of genetic algorithms, particle swarm optimization and the differential evolution method for the design of scannable circular antenna arrays. *Prog. Electromagn. Res. B* **13**, 171–186 (2009)
- D. Pilgrim, Travel times and nonlinearity of flood runoff from tracer measurements on a small watershed. *Water Resour. Res.* **12**(3), 487–496 (1976)
- J. Pintér, Continuous global optimization software: A brief review. *Optima* **52**(1–8), 270 (1996)
- N.V. Queipo, R.T. Haftka, W. Shyy, T. Goel, R. Vaidyanathan, P. Kevin Tucker, Surrogate-based analysis and optimization. *Prog. Aerosp. Sci.* **41**(1), 1–28 (2005)
- C. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006)
- S. Razavi, B.A. Tolson, D.H. Burn, Review of surrogate modeling in water resources. *Water Resour. Res.* **48**(7), 401–433 (2012). <https://doi.org/10.1029/2011WR011527>
- P.M. Reed, D. Hadka, J.D. Herman, J.R. Kasprzyk, J.B. Kollat, Evolutionary multiobjective optimization in water resources: The past, present, and future. *Adv. Water Resour.* **51**, 438–456 (2013)
- R.G. Regis, C.A. Shoemaker, A stochastic radial basis function method for the global optimization of expensive functions. *INFORMS J. Comput.* **19**, 497–509 (2007)
- D.A. Savic, G.A. Walters, J.W. Davidson, A genetic programming approach to rainfall-runoff modelling. *Water Resour. Manag.* **13**(3), 219–231 (1999)
- P. Serafini, *Multiple Criteria Decision Making* (Springer, Berlin, 1994), pp. 283–292
- M. Shafii, F.D. Smedt, Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. *Hydrol. Earth Syst. Sci.* **13**(11), 2137–2149 (2009)
- Y. Shi, *Particle Swarm Optimization: Developments, Applications and Resources* (IEEE, Seoul, 2001), pp. 81–86. <https://doi.org/10.1109/CEC.2001.934374>
- A.R. Simpson, G.C. Dandy, L.J. Murphy, Genetic algorithms compared to other techniques for pipe optimization. *J. Water Resour. Plan. Manag.* **120**(4), 423–443 (1994)
- T.W. Simpson, J.D. Peplinski, P.N. Koch, J.K. Allen, Metamodels for computer-based engineering design: Survey and recommendations. *Eng. Comput.* **17**, 129–150 (2001)
- K.P. Singh, Nonlinear instantaneous unit hydrograph theory. *J. Hydraul. Div. Am. Soc. Civ. Eng.* **90**, 313–347 (1964)
- V.P. Singh, *Computer Models of Watershed Hydrology* (Water Resources Publications, Englewood, 1995)
- B.E. Skahill, J. Doherty, Efficient accommodation of local minima in watershed model calibration. *J. Hydrol.* **329**(1), 122–139 (2006)
- E. Snelson, Flexible and efficient Gaussian process models for machine learning. Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London 2007
- A. Sóbester, S. Leary, A. Keane, On the design of optimization strategies based on global response surface approximation models. *J. Glob. Optim.* **33**(1), 31–59 (2005)
- K. Socha, M. Dorigo, Ant colony optimization for continuous domains. *Eur. J. Oper. Res.* **185**(3), 1155–1173 (2008)
- S. Sorooshian, Surface water hydrology: On-line estimation. *Rev. Geophys.* **21**(3), 706–721 (1983)
- P. Srivastava, J. Hamlett, P. Robillard, R. Day, Watershed optimization of best management practices using AnnAGNPS and a genetic algorithm. *Water Res. Res.* **38**(3), 3–1 (2002)
- B. Suman, Study of simulated annealing based algorithms for multiobjective optimization of a constrained problem. *Comput. Chem. Eng.* **28**(9), 1849–1871 (2004)
- N.R. Sumner, P.M. Fleming, B.C. Bates, Calibration of a modified SFB model for twenty-five Australian catchments using simulated annealing. *J. Hydrol.* **197**(1), 166–188 (1997)
- Q. Sun, D. Kong, C. Miao, Q. Duan, T. Yang, A. Ye, Z. Di, W. Gong, Variations in global temperature and precipitation for the period of 1948 to 2010. *Environ. Monit. Assess.* **186**(9), 5663–5679 (2014)
- A. Suppapitnarm, K. Seffen, G. Parks, P. Clarkson, A simulated annealing algorithm for multi-objective optimization. *Eng. Optim.* **33**(1), 59–85 (2000)

- Y. Tang, P. Reed, T. Wagener, How effective and efficient are multiobjective evolutionary algorithms at hydrologic model calibration? *Hydrol. Earth Syst. Sci. Discuss.* **2**(6), 2465–2520 (2005)
- M. Thyer, G. Kuczera, B.C. Bates, Probabilistic optimization for conceptual rainfall-runoff models: A comparison of the shuffled complex evolution and simulated annealing algorithms. *Water Resour. Res.* **35**(3), 767–773 (1999)
- J.A. Vrugt, H.V. Gupta, W. Bouten, S. Sorooshian, A shuffled complex evolution metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resour. Res.* **39**(8), 1201–1213 (2003a). <https://doi.org/10.1029/2002WR001642>
- J.A. Vrugt, H.V. Gupta, L.A. Bastidas, W. Bouten, S. Sorooshian, Effective and efficient algorithm for multiobjective optimization of hydrologic models. *Water Resour. Res.* **39**(8), 1214–1233 (2003b). <https://doi.org/10.1029/2002WR001746>
- Q. Wang, The genetic algorithm and its application to calibrating conceptual rainfall-runoff models. *Water Resour. Res.* **27**(9), 2467–2471 (1991)
- Q. Wang, Using genetic algorithms to optimise model parameters. *Environ. Model Softw.* **12**(1), 27–34 (1997)
- H. Wang, W. Guo, ACO Optimizing Neural Network for Macroscopic Water Distribution System Modeling (IEEE, Kuala Lumpur, 2010), pp. 367–370. <https://doi.org/10.1109/ICICCI.2010.109>
- Y.C. Wang, P.S. Yu, T.C. Yang, Comparison of genetic algorithms and shuffled complex evolution approach for calibrating distributed rainfall-runoff model. *Hydrol. Process.* **24**(8), 1015–1026 (2010)
- C. Wang, Q.Y. Duan, W. Gong, A.Z. Ye, Z.H. Di, C.Y. Miao, An evaluation of adaptive surrogate modeling based optimization with two benchmark problems. *Environ. Model. Softw.* **60**, 167–179 (2014)
- P.A. Whigham, P.F. Crapper, Time series modelling using genetic programming: An application to rainfall-runoff models. *Adv. Genet. Program* **3**, 89–104 (1999)
- D.H. Wolpert, W.G. Macready, No free lunch theorems for optimization. *IEEE Trans. Evol. Comput.* **1**(1), 67–82 (1997)
- C.F.J. Wu, M. Hamada, *Experiments: Planning, Analysis, and Optimization*, 2nd edn. (Wiley, New York, 2009)
- S.-J. Wu, H.-C. Lien, C.-H. Chang, Calibration of a conceptual rainfall-runoff model using a genetic algorithm integrated with runoff estimation sensitivity to parameters. *J. Hydroinf.* **14**(2), 497–511 (2012)
- J. Yang, P. Reichert, K.C. Abbaspour, J. Xia, H. Yang, Comparing uncertainty analysis techniques for a SWAT application to the Chaohe Basin in China. *J. Hydrol.* **358**(1–2), 1–23 (2008)
- T. Yang, X. Gao, S.L. Sellars, S. Sorooshian, Improving the multi-objective evolutionary optimization algorithm for hydropower reservoir operations in the California Oroville–Thermalito complex. *Environ. Model Softw.* **69**, 262–279 (2015)
- T. Yang, X. Gao, S. Sorooshian, X. Li, Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. *Water Resour. Res.* **52**(3), 1626–1651 (2016)
- T. Yang, A.A. Asanjan, M. Faridzad, N. Hayatbini, X. Gao, S. Sorooshian, An enhanced artificial neural network with a shuffled complex evolutionary global optimization with principal component analysis. *Inf. Sci.* **418**, 302–316 (2017a)
- T. Yang, A.A. Asanjan, E. Welles, X. Gao, S. Sorooshian, X. Liu, Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. *Water Resour. Res.* **53**(4), 2786–2812 (2017b)
- T. Yang, Y. Tao, J. Li, Q. Zhu, L. Su, X. He, X. Zhang, Multi-criterion model ensemble of CMIP5 surface air temperature over China. *Theor. Appl. Climatol.* **132**(3), 1057–1072 (2017c). <https://doi.org/10.1007/s00704-017-2143-4>
- P.O. Yapo, H.V. Gupta, S. Sorooshian, Multi-objective global optimization for hydrologic models. *J. Hydrol.* **204**(1), 83–97 (1998)

- M. Zambrano-Bigiarini, R. Rojas, A model-independent particle swarm optimisation software for model calibration. *Environ. Model Softw.* **43**, 5–25 (2013)
- A.C. Zecchin, H.R. Maier, A.R. Simpson, A. Roberts, M.J. Berrisford, M. Leonard, Max-min ant system applied to water distribution system optimization. *Proc. Int. Congr. Model. Simul. (MODSIM)* **2**, 795–800 (2003)
- A.C. Zecchin, A.R. Simpson, H.R. Maier, A. Marchi, J.B. Nixon, Improved understanding of the searching behavior of ant colony optimization algorithms applied to the water distribution design problem. *Water Resour. Res.* **48**(9), 795–800 (2012)
- Y. Zhang, F.H.S. Chiew, Relative merits of different methods for runoff predictions in ungauged catchments. *Water Res. Res.* **45**(7), 412–425 (2009). <https://doi.org/10.1029/2008WR007504>
- X. Zhang, R. Srinivasan, M. Van Liew, Approximating SWAT model using artificial neural network and support vector machine. *J. Am. Water Resour. Assoc.* **45**(2), 460–474 (2009a)
- X. Zhang, R. Srinivasan, D. Bosch, Calibration and uncertainty analysis of the SWAT model using genetic algorithms and Bayesian model averaging. *J. Hydrol.* **374**(3), 307–317 (2009b)
- X. Zhang, R. Srinivasan, K. Zhao, M.V. Liew, Evaluation of global optimization algorithms for parameter calibration of a computationally intensive hydrologic model. *Hydrol. Process.* **23**(3), 430–441 (2009c)
- X. Zhang, R. Srinivasan, M.V. Liew, On the use of multi-algorithm, genetically adaptive multi-objective method for multi-site calibration of the SWAT model. *Hydrol. Process.* **24**(8), 955–969 (2010)
- Q. Zhu, K.I. Hsu, Y.P. Xu, T. Yang, Evaluation of a new satellite-based precipitation data set for climate studies in the Xiang River basin, southern China. *Int. J. Climatol.* **37**, 4561 (2017)
- E. Zitzler, L. Thiele, Multiobjective evolutionary algorithms: A comparative case study and the strength Pareto approach. *IEEE Trans. Evol. Comput.* **3**(4), 257–271 (1999)
- E. Zitzler, K. Deb, L. Thiele, Comparison of multiobjective evolutionary algorithms: Empirical results. *Evol. Comput.* **8**(2), 173–195 (2000)



Uncertainty Quantification of Complex System Models: Bayesian Analysis

Jasper A. Vrugt and Elias C. Massoud

Contents

1	Introduction and Scope	564
2	Model Calibration	566
3	Parameter Uncertainty: First-Order Approximation	570
4	Bayesian Inference	571
4.1	The Prior Distribution, $P(\theta)$	573
4.2	The Likelihood Function, $L(\theta \tilde{Y})$	575
4.3	Generalized Likelihood Function	583
4.4	The Posterior Distribution, $P(\theta \tilde{Y})$	588
5	Monte Carlo Approximation	590
5.1	Rejection Sampling	590
5.2	Importance Sampling	592
5.3	Markov Chain Monte Carlo Simulation	593
6	Case Studies	608
6.1	Instantaneous Unit Hydrograph	608
6.2	The Rainfall-Runoff Transformation	610
6.3	Vadose Zone Hydrology	612
7	Limits of Acceptability	619
7.1	The DREAM _(LOA) Algorithm	621
7.2	Vadose Zone Hydrology Revisited	621
8	Marginal Likelihood and Model Complexity	622
9	Conclusion	626

J. A. Vrugt (✉)

Department of Civil and Environmental Engineering, University of California Irvine,
Irvine, CA, USA

Department of Earth System Science, University of California Irvine, Irvine, CA, USA
e-mail: jasper@uci.edu

E. C. Massoud

Department of Civil and Environmental Engineering, University of California Irvine,
Irvine, CA, USA

Appendix	627
A: Derivation of Bayes' Theorem	627
B: MATLAB Code DREAM	629
C: MATLAB Code DREAM _(LOA)	631
References	632

Abstract

This chapter summarizes the main elements of Bayesian probability theory to help reconcile dynamic environmental system models with observations, including prediction in space (interpolation), prediction in time (forecasting), assimilation of data, and inference of the model parameters. Special attention is given to the treatment of parameter uncertainty (first-order approximations and Bayesian intervals), the prior distribution, the formulation of the likelihood function (using first-principles), the marginal likelihood, and sampling techniques used to estimate the posterior target distribution. This includes rejection sampling, importance sampling, and recent developments in Markov chain Monte Carlo simulation to sample efficiently complex and/or high-dimensional target distributions, including limits of acceptability. We illustrate the application of Bayes' theorem and inference using three illustrative examples involving the flow and storage of water in the surface and subsurface. At least some level of calibration of these models is required to match their output with observations of system behavior and response. Algorithmic recipes of the different methods are provided to simplify implementation and use of Bayesian analysis.

Keywords

Hypothesis testing · Bayesian analysis · Prior distribution · Likelihood function · Posterior distribution · Monte Carlo sampling · Markov chain Monte Carlo simulation · Data assimilation · Hydrologic modeling

1 Introduction and Scope

The Earth is the densest planet in our solar system and the only astronomical object known to mankind to harbor life. About 71% of the Earth's surface is covered with water, and the remaining 29% constitutes land mass made up of continents and islands dissected by rivers, lakes, and other sources of water that contribute to the hydrosphere. The large-scale motion of the Earth's outermost shell (lithosphere), composed of several tectonic plates which float on a hotter, softer layer in the mantle (asthenosphere), has created mountain ranges and volcanic activity on plate boundaries. Coevolution and juxtaposition of these topographic features with climatic and geologic variations have resulted in a highly diverse landscape with large variations in soils, vegetation, geomorphology, and biota (biosphere). These landscapes can be conceived as a series of large and small ecosystems, nested within one another in a hierarchy of spatial scales.

Ecosystems constitute a complex network of living organisms, which are interconnected and linked together with the abiotic environment through a myriad of interrelated physical, chemical, and biological processes operating at or near the Earth's surface. Many of these processes are difficult, costly, labor intensive, and/or unethical to measure directly in the field, particularly at large spatial scales. This daunting complexity has stimulated researchers in many different fields of study to explore the use of mathematical modeling to mimic the behavior of complex systems. Computer models are particularly useful to gain (new) insights and understanding of system functioning and to predict behavior into the space (interpolation) and time (forecasting) domain. The capabilities of such models exceed by far traditional paper-and-pencil calculations and can involve simulations on spatial scales of individual atoms to the entire ecosystem, and temporal scales of nanoseconds to many millions of years. Examples include numerical weather prediction models, astrophysical and cosmological simulations of dark matter, computational modeling of the brain, and spatially distributed simulation of environmental systems. The CPU-time of these simulations can vary from less than a second for simple dynamic models with fixed (integration) time step up to many hours of calculation for spatially explicit models involving multidimensional numerical solution of (systems of) differential-algebraic or ordinary/partial differential equations.

The model building process is strongly influenced by perception, intuition, and prior knowledge on system functioning and reality, and colored by mental concepts (state of mind). From countless processes and mechanisms, the modeler seeks to isolate, detect, and generalize into laws those key principles that explain the observed data. Their selection and translation to a mathematical model is the most critical, difficult, and subjective part of modeling. To guard against the use of an inadequate model, statisticians advise selecting the “best” model among a set of plausible candidate models chosen and/or construed by the researcher(s). This approach rules out model selection bias and recognizes explicitly the ambiguity in the interpretation and analysis of complex natural systems. The ensemble of models, or hypotheses, constitute a finite sample of possible explanations of the data deemed plausible a-priori from the extremely large, perhaps even incomprehensible, space of alternatives. This can include black-box, conceptual (empirical), and physically based models and involve widely different mechanisms of the spatio-temporal processes that determine system behavior and response. Each of these models might be as justifiable as the other (Vrugt and Robinson 2007; Ye et al. 2008; Clark et al. 2011). In the penultimate section of this chapter, we will shortly discuss the issue of model selection. For the time being, we refer the reader to the work of Volpi et al. (2017) and references therein for hypothesis testing using a suite of competing models. Nevertheless, most of the material of this chapter is applicable to any mathematical model, in so far, measurement data are available to evaluate (falsify) the consistency of simulated output.

Figure 1 provides a schematic overview of most important sources of uncertainty that affect our ability to mimic perfectly complex dynamical systems. These sources of uncertainty have been discussed extensively in the literature, and many different (non)statistical methods have been developed to quantify parameter, calibration data,

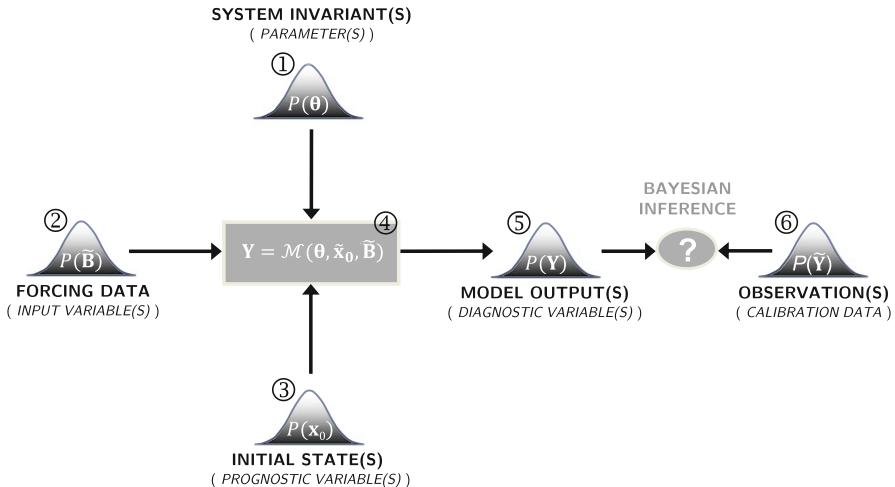


Fig. 1 Schematic illustration of the most important sources of uncertainty in environmental systems modeling, including (1) parameter, (2) input data (also called forcing or boundary conditions), (3) initial state, (4) model structural, (5) output, and (6) calibration data uncertainty. The measurement error of the calibration data is often prescribed, a rather convenient assumption in most practical situations

model output, and state variable uncertainty. Model structural errors (4: epistemic error) have received relatively little attention, yet are key to learning and scientific discovery (Gupta et al. 2008; Vrugt et al. 2005; Vrugt and Sadegh 2013).

The focus of this chapter is on the characterization of the different error sources depicted schematically in Fig. 1 with specific emphasis on a statistical representation of model parameter uncertainty. The material of this chapter is taken directly from the first author's graduate course on "Merging Models and Data" (CEE-290) at the University of California, Irvine. We refer interested readers to the lecture on Bayesian analysis (topic 4 in CEE-290) which appears on [YouTube](#) at the following link <https://www.youtube.com/watch?v=bZAhv5z6NaY>. We assume application to spatio-temporal environmental models that may be discrete in time and/or space but with processes that are continuous in both.

2 Model Calibration

Consider a n -vector of measurements, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \dots, \tilde{y}_n\}$ observed at discrete times $t = \{1, \dots, n\}$ that summarizes the response of an environmental system \mathfrak{F} to k temporally variant control inputs, $\mathbf{B} = \{\mathbf{b}_1, \dots, \mathbf{b}_n\}$, with column elements, $\mathbf{b}_t = \{b_{t1}, \dots, b_{tk}\}$. We use a computer model, $\mathcal{M}(\cdot)$, to explain the observed data

$$\tilde{\mathbf{Y}} \leftarrow \mathcal{M}(\boldsymbol{\theta}, \mathbf{x}_0, \tilde{\mathbf{B}}) + \mathbf{E}, \quad (1)$$

where $\boldsymbol{\theta} = \{\theta_1, \dots, \theta_d\}$ is the $d \times 1$ -vector of model parameters, \mathbf{x}_0 stores the values of the state variables at the start of simulation, $\tilde{\mathbf{B}}$ signifies the $k \times n$ control matrix with temporal measurements of the forcing variables, and $\mathbf{E} = \{e_1, \dots, e_n\}$ is a vector of residuals. The index t for time takes on strictly positive integer values in the remainder of this chapter, $t \in \{1, \dots, n\} \in \mathbb{N}_+$, yet may take on real values, $t \in (0, n] \in \mathbb{R}_+$ in the actual system model, $\mathcal{M}(\cdot)$, to resolve for continuous-time processes, wherein the simulated output at $t = 0$ is defined completely by \mathbf{x}_0 .

The model in Eq. (1) simplifies considerably the description of the spatially distributed real-world system, into a lumped topology consisting of much fewer, and discrete, entities. This simplification is computationally convenient in that it reduces to a finite dimension the state space of the system, and the partial differential equations of the continuous time and space domain of the physical system into ordinary differential equations with much fewer parameters. If deemed appropriate, a spatially explicit formulation can be used instead with control vector, \mathbf{b} , formulated as a two- or three-dimensional matrix to account explicitly for spatially varying boundary conditions. Without further loss of generality, we restrict the model parameters to a closed space, Θ , equivalent to a d -dimensional hypercube, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, called the feasible parameter space. The n -vector of error residuals, \mathbf{E} , thus depends on the assumed model, \mathcal{M} , and its associated parameters, initial states and forcing data, hence

$$\mathbf{E}_{\mathcal{M}}(\boldsymbol{\theta}, \mathbf{x}_0, \tilde{\mathbf{B}}) = \tilde{\mathbf{Y}} - \mathbf{Y}_{\mathcal{M}}(\boldsymbol{\theta}, \mathbf{x}_0, \tilde{\mathbf{B}}) \quad (2)$$

where $\mathbf{Y}_{\mathcal{M}}(\boldsymbol{\theta}, \mathbf{x}_0, \tilde{\mathbf{B}})$ signifies the simulated output of the model, \mathcal{M} .

For the time being, lets make the convenient assumption that \mathcal{M} mimics perfectly the underlying system, \mathfrak{F} , it is intended to represent. Lets further assume that the forcing data are observed without measurement error, $\delta(\mathbf{B}, \tilde{\mathbf{B}}) = 0$, and that errors in the initial states, \mathbf{x}_0 , pose no harm as their impact on the simulated output, \mathbf{Y} , diminishes rapidly with advancing time. This latter assumption is certainly appropriate for real-world systems controlled by negative (or degenerative) feedback. A prime example of negative feedback is the thermostat. If room temperature drops below a reference value, the furnace will supply heat and restore the temperature to “normal.” This cooling-warming cycle thus regulates room temperature. Similar degenerative interactions are found among processes that control the rainfall-runoff transformation in a watershed. Surface runoff, overland flow, evaporation, infiltration, transpiration, drainage, and recharge (among others) act together to remove excess precipitation, thereby promoting convergence of the soil moisture status to a stable state. The existence of such equilibrium state is easily verified in practice using repeated numerical simulation with a watershed model using different values of the initial states. This equilibrium state does not exist for systems whose behavior is regulated by positive feedback as small perturbations to the initial states can lead to widely different responses of the model via exponential growth, oscillation, or chaotic behavior. For systems with negative feedback, a spin-up period of Q days therefore suffices to promote stability and ameliorate the effect of state initialization errors on the model output, $\lim_{t \rightarrow Q} \delta(y_t(\tilde{\mathbf{x}}_0), y_t(\mathbf{x}_0)) \rightarrow 0$.

For systems with generative (negative) feedbacks, the error in the initial states poses no harm as its effect on system simulation rapidly diminishes when time advances. One can therefore take advantage of a spin-up period to remove sensitivity of the modeling results (and error residuals) to state value initialization.

The assumptions of perfect model, input data, and initial states (due to spin-up period) are common to environmental modeling. This so-called ideal case leaves as our only “unknowns” the model parameters, and possibly as well, the variance of the data measurement errors (more of which later), which may account implicitly for structural errors in the model. The residual vector can thus be written as

$$\mathbf{E}(\boldsymbol{\theta}) = \tilde{\mathbf{Y}} - \mathbf{Y}(\boldsymbol{\theta}) = \{e_1(\boldsymbol{\theta}), \dots, e_n(\boldsymbol{\theta})\}, \quad (3)$$

and necessitates inference on $\boldsymbol{\theta}$ to minimize the n -vector of residuals, $\mathbf{E}(\boldsymbol{\theta})$.

Figure 2 summarizes the resulting model calibration problem with explicit recognition of measurement, parameter, and epistemic uncertainty. Symbol \oplus signifies the measurement operator and provides (temporal) observations of the control variables and the response of the physical system, . The mathematical model, , is at best only an approximation of the curly shaped, physical system.

The prior values of the parameters provide a simulation (gray line) that mimics reasonably well the transient behavior of the system (blue dots), yet underestimates systematically the peaks and recession periods. A subsequent trial of the parameter values much better explains (green line) the observed data. Model calibration now involves the searching of the parameter values that mimic “best” the observed system response. The calibrated model can then serve a host of different purposes such as process analysis, evaluating different management strategies, and prediction of system behavior into the space and/or time domain. These tasks can only be completed with confidence if the physical system of interest and its control inputs (forcing variables) satisfy “constancy.” This stationarity assumption is rather convenient and opens up the wide arsenal of (multivariate) statistical and nonlinear optimization methods for inference of the model parameters (Sadegh et al. 2015).

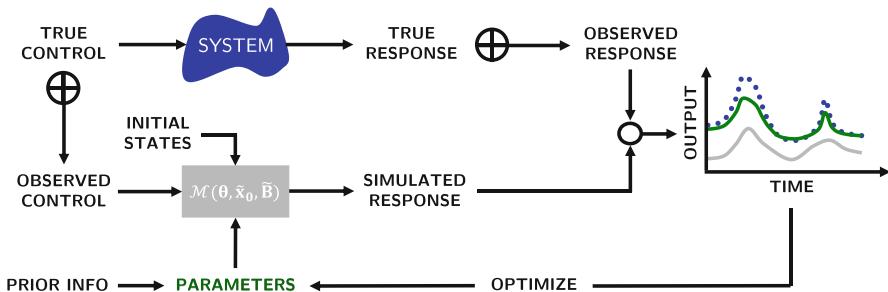


Fig. 2 Schematic overview of the model calibration problem in the presence of measurement, parameter, and epistemic uncertainty. The model parameters are adjusted iteratively so that the simulated response (solid lines) of the model, $\mathcal{M}(\boldsymbol{\theta}, \tilde{\mathbf{x}}_0, \tilde{\mathbf{B}})$, approximates as closely and consistently as possible the observed response (blue dots)

Note, per scope of this book, we will focus primarily on hydrologic models with *output* calibration targets such as river discharge and soil moisture content.

The word “best” appears purposely quoted in the previous paragraph as much research has shown that there is no unambiguously correct way in which to determine unique model parameters. Indeed, the “optimal” parameter values are dependent critically on the assumptions that are made with respect to the different error sources of Fig. 1. In the ideal case (perfect model, input data and initial states), the sum of squared residuals (SSR)

$$\min_{\boldsymbol{\theta} \in \Theta} F(\boldsymbol{\theta}) = \sum_{i=1}^n e_i(\boldsymbol{\theta})^2, \quad (4)$$

provides unbiased and minimum-variance estimates of the parameters when the measurement errors of the calibration data, $\hat{\mathbf{Y}}$, are homoscedastic (constant variance) and serially (temporally) uncorrelated. This is also referred to as the least squares solution. Visually, this solution minimizes the sum of squared vertical distances between the n -vector of data points, $\hat{\mathbf{Y}}$ and the corresponding simulated values, $\mathbf{Y}(\boldsymbol{\theta})$, of the model. The lower the value of the SSR, the better the model fits the data. When the measurement errors are believed to have a nonconstant variance, the heteroscedastic maximum likelihood estimator (HMLE) can be used (Sorooshian and Dracup 1980). Other commonly used metrics in hydrologic model calibration include the coefficient of determination, the index of agreement, and the Nash-Sutcliffe efficiency (Nash and Sutcliffe 1970), although these metrics are not rooted in statistical theory.

In Eq. (4), the function $F(\boldsymbol{\theta})$ is also called the objective function. This function lumps the n -residuals of the vector $\mathbf{E}(\boldsymbol{\theta})$ into a single aggregate measure of model-data mismatch, and this measure is subsequently minimized (or maximized, if appropriate) with an optimization algorithm. This constitutes the field of constrained optimization, that is,

$$\arg \min_{\boldsymbol{\theta} \in \Theta} F(\boldsymbol{\theta}), \quad (5)$$

where the goal is to find the optimum parameter values as rapidly as possible using the smallest number of model evaluations. The constraints are equivalent to the lower and upper bounds of the parameters and ensure that the optimum parameter values reside in the feasible parameter space, Θ . The choice of objective function, however, remains a rather intricate and difficult task, fraught with subjective assumptions regarding model structural, control data, and calibration data errors.

Optimization algorithms provide an estimate of the “best” parameter values that minimize (maximize) some predefined objective function, $F(\boldsymbol{\theta})$. It would be naive, however, to rely on such single unique estimate of the parameters in the presence of epistemic uncertainty and measurement errors of the control input and calibration data. Indeed, practical experience suggests that it is typically difficult to find a single “best” vector of parameter values, whose performance obviates consideration from

other feasible solutions. It is therefore of paramount importance to investigate and delineate properly the space of feasible solutions. This is key to (among others) analysis of parameter identifiability and quantification of the uncertainty associated with the simulated model output.

3 Parameter Uncertainty: First-Order Approximation

Per statistical theory, we can approximate the confidence intervals of the parameters by centering around the $d \times 1$ -vector of optimum parameter values, $\boldsymbol{\theta}^*$, a d -variate normal distribution, $\mathcal{N}_d(\boldsymbol{\theta}^*, \mathbf{C}(\boldsymbol{\theta}^*))$, with $d \times d$ covariance matrix \mathbf{C} evaluated at $\boldsymbol{\theta}^*$. (This section may have elements that are difficult to digest now but easier to comprehend after reading Sect. 4.). The probability density function, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, of this multivariate Gaussian is given by

$$P(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = (2\pi)^{-\frac{d}{2}} |\mathbf{C}(\boldsymbol{\theta}^*)|^{\frac{1}{2}} \exp\left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)^T \mathbf{C}(\boldsymbol{\theta}^*)^{-1} (\boldsymbol{\theta} - \boldsymbol{\theta}^*)\right), \quad (6)$$

where $|\cdot|$ signifies the determinant operator, and T denotes transpose. The symbol “|” in $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ characterizes conditional probability and conveys that the distribution of Eq. (6) is conditioned on the observed data, $\tilde{\mathbf{Y}}$. Note that Eq. (6) reduces to a univariate normal distribution if $\mathbf{C}(\boldsymbol{\theta}^*)$ is a 1×1 matrix (scalar). The $d \times d$ covariance matrix, $\mathbf{C}(\boldsymbol{\theta}^*)$, can be derived from the model parameter sensitivity matrix, $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$, as follows

$$\mathbf{C}(\boldsymbol{\theta}^*) = \sigma_E^2 \left(\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)^T \mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*) \right)^{-1}, \quad (7)$$

where $\sigma_E^2 = \text{SSR}/(n - d)$ denotes the variance of the residuals, and the symbol $^{-1}$ signifies matrix inverse. The transpose operator acting on the first of two $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$'s (term between brackets) enforces equal inner and outer dimensions of n and d , respectively, so that the matrix product produces, after inversion and multiplication with σ_E^2 , the $d \times d$ covariance matrix, $\mathbf{C}(\boldsymbol{\theta}^*)$.

The variance of each parameter is now stored in the main diagonal of $\mathbf{C}(\boldsymbol{\theta}^*)$. Indeed, the $100\alpha\%$ confidence interval of each of the d model parameters can now be calculated using

$$\boldsymbol{\theta}^* \pm t_{(n-d),(1-\alpha)/2} \sqrt{\text{diag}(\mathbf{C}(\boldsymbol{\theta}^*))}, \quad (8)$$

where $t_{(n-d),(1-\alpha)/2}$ signifies the value of Student's T cumulative distribution function at $(1 - \alpha)/2$ and $n - d$ degrees of freedom. For abundant measurements, $n \gg d$, the critical t value converges to textbook values of 1.00, 1.96, and 2.58 for $\alpha = 0.68$, 0.95, and 0.99, respectively.

The $n \times d$ Jacobian matrix, $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$, stores the first-order partial derivatives of the model output with respect to each of the parameters

$$\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*) = \frac{\partial \mathcal{M}(\boldsymbol{\theta}^*)}{\partial \boldsymbol{\theta}^*} = \begin{bmatrix} \frac{\partial \mathbf{Y}(\boldsymbol{\theta}^*)}{\partial \theta_1^*} & \dots & \frac{\partial \mathbf{Y}(\boldsymbol{\theta}^*)}{\partial \theta_d^*} \end{bmatrix} = \begin{bmatrix} \frac{\partial y_1(\boldsymbol{\theta}^*)}{\partial \theta_1^*} & \dots & \frac{\partial y_1(\boldsymbol{\theta}^*)}{\partial \theta_d^*} \\ \vdots & \ddots & \vdots \\ \frac{\partial y_n(\boldsymbol{\theta}^*)}{\partial \theta_1^*} & \dots & \frac{\partial y_n(\boldsymbol{\theta}^*)}{\partial \theta_d^*} \end{bmatrix}. \quad (9)$$

The j th column of $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$ thus stores the sensitivity of the n elements of the model output $\mathbf{Y} = \{y_1, \dots, y_n\}$ to the j th parameter. These columns are also referred to as basis functions. For models whose output \mathbf{Y} depends linearly on the values of the parameters, the d basis functions of $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$ can be derived analytically. Then the basis functions are valid and each column of $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$ will be constant and independent of $\boldsymbol{\theta}$, and the first-order approximation of \mathbf{C} in Eq. (7) will be exact. For nonlinear models, we cannot determine analytically the entries of $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$. The basis functions are then invalid and we must resort to numerical simulation to approximate the Jacobian matrix. For instance, we can approximate the first column of the sensitivity matrix, $\partial \mathbf{Y}(\boldsymbol{\theta}^*)/\partial \theta_1^*$, by perturbing the first parameter with a quantity $\Delta \theta_1$ and calculating the resulting change in model output from the default (optimal) simulation with $\boldsymbol{\theta}^*$

$$\frac{\partial \mathbf{Y}(\boldsymbol{\theta}^*)}{\partial \theta_1^*} = \frac{\Delta \mathbf{Y}(\boldsymbol{\theta}^*)}{\Delta \theta_1} \approx \frac{\mathbf{Y}(\{\theta_1^* + \Delta \theta_1, \dots, \theta_d^*\}) - \mathbf{Y}(\{\theta_1^*, \dots, \theta_d^*\})}{\Delta \theta_1}. \quad (10)$$

This recipe is repeated for the other $d - 1$ parameters (columns) of $\mathbf{J}_{\mathcal{M}}(\boldsymbol{\theta}^*)$. Eq. (10) uses a one-sided interval to approximate the partial model derivatives. More accurate results will be obtained if a two-sided interval is used with $-\Delta \theta_1$ and $\Delta \theta_1$, yet this doubles the number of simulations.

For models whose output is linearly dependent on their parameters, the first-order approximation of Eq. (6) will give an exact description of the parameter uncertainty. For all other models with invalid basis functions, this first-order approximation provides only an estimate of the actual parameter uncertainty. This approximation can be deficient if the covariance matrix varies considerably over the domain of $\boldsymbol{\theta}$ for which there is significant uncertainty. This is more the rule than the exception for nonlinear system models, particularly when the actual $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ distribution is poorly described by a multivariate normal distribution due to the presence of multimodality, local minima, and strong nonlinear parameter interactions. What is more, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, can be truncated by the prior distribution.

4 Bayesian Inference

Bayesian inference allows for an exact description of parameter uncertainty (and other sources of uncertainty) by treating the parameters (and nuisance variables) as probabilistic variables with joint posterior probability density function, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$. This multivariate distribution, the so-called posterior parameter distribution, is the

consequence of two antecedents, a prior distribution which captures our initial degree of beliefs in the values of the model parameters, and a likelihood function which quantifies by the rules of probability theory the level of confidence (= conditional belief) in the parameter values, $\boldsymbol{\theta}$, in light of the observed data, $\tilde{\mathbf{Y}}$, alone. Bayes' theorem (also referred to as Bayes' law or Bayes' rule) expresses mathematically, and in a simple formula, the fundamental relationship between the prior, conditional, and posterior (= updated) beliefs of the parameters. The theorem is named after the English statistician, philosopher, and Presbyterian minister Thomas Bayes (1701–1761), who formulated the solution in written notes. These ideas were not published until 2 years after Bayes' death, when his work on inverse probability emerged posthumously in the *Philosophical Transactions of the Royal Society of London* in the masterwork “*An Essay Towards Solving a Problem in the Doctrine of Chances*” (Bayes and Price 1763). In this publication, Bayes' relates the “direct” probability of a hypothesis conditional on some body of data to the “inverse” probability of the data conditional on the hypothesis (nowadays referred to as likelihood).

Bayes' ideas gained limited exposure at the time of publication in 1763 until they were rediscovered and enhanced further by the French scholar Pierre-Simon Laplace (1749–1827) (Literature research by Stigler (1983) and Zabell (1989) provides evidence that Thomas Bayes may have communicated his theorem to friends as early as 1748, in response to a challenge from the famous Scottish philosopher David Hume (1711–1776)). In recent decades, Bayes theorem has emerged as a corner stone of modern probability theory (hypothesis testing) and as working paradigm for the subjectivist approach to epistemology, statistics, and inductive logic. Subjectivists draw heavily from Bayes' law in their efforts to quantify, using the laws of probability, how a rational person's confidence level (or degree of belief) in a hypothesis changes when exposed to new evidence. This subjectivist (or Bayesian) viewpoint shares many common elements with the human learning process wherein existing knowledge, behaviors, skills, values, or preferences are continuously modified or reinforced in response to incoming data (information).

Bayes' theorem can be derived from the basic axioms of probability, specifically conditional probability (see Appendix A), and reads in our application

$$P(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) = \frac{P(\boldsymbol{\theta})P(\tilde{\mathbf{Y}}|\boldsymbol{\theta})}{P(\tilde{\mathbf{Y}})}, \quad (11)$$

where $P(\boldsymbol{\theta})$ and $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$ signify the prior and posterior parameter distribution, respectively and $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \equiv P(\tilde{\mathbf{Y}}|\boldsymbol{\theta})$ denotes the likelihood function. The model evidence, $P(\tilde{\mathbf{Y}})$ (or marginal likelihood), acts as a normalizing constant (scalar)

$$P(\tilde{\mathbf{Y}}) = \int_{\Theta} P(\boldsymbol{\theta})P(\tilde{\mathbf{Y}}|\boldsymbol{\theta})d\boldsymbol{\theta} = \int_{\Theta} P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})d\boldsymbol{\theta} = \int_{\Theta} P(\boldsymbol{\theta},\tilde{\mathbf{Y}})d\boldsymbol{\theta}, \quad (12)$$

so that the posterior distribution integrates to unity over the prior (feasible) parameter space, $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$. Knowledge of $P(\tilde{\mathbf{Y}})$ is strictly necessary for hypothesis testing

to select the most plausible model of the real-world system \mathfrak{F} from a set of different models deemed valid a-priori. We will briefly discuss the topic of model selection in Sect. 8 of this chapter. For now, we suffice to say that the evidence of a model is largest, if its data likelihood is high relative to other models and distributed uniformly over the prior parameter space, Θ . The evidence estimates can also serve as weights to average the simulations of the different models, as in Bayesian model averaging (Hoeting et al. 1999; Wasserman 2000; Raftery et al. 2005; Vrugt and Robinson 2007). Appendix A illustrates the application of Bayes' theorem to determine, using a hypothetical data set, the conditional probabilities of two events, rainfall and thunder.

If we rely on a single hypothesis, $\mathcal{M}(\cdot)$, of the system \mathfrak{F} of interest, then the denominator, $P(\tilde{\mathbf{Y}})$ in Eq. (11), is of no particular interest as all statistical inferences about the parameters of $\mathcal{M}(\cdot)$ can be made from the unnormalized posterior distribution

$$P(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \quad (13)$$

The proportionality sign conveys that the posterior density of the unnormalized distribution is an unknown multiple (scalar) of the normalized density. This might be confusing, yet consider a empirical histogram of $M = 1000$ independent observations of some variable of interest. Whether we use for each bin units of frequency, relative frequency (= frequency/ M) or probability density (histogram integrates to one), the 95% confidence intervals of the variable are unaffected. Thus, our inferences of the parameters are protected against linear transformations of the density.

Numerical implementation of the Bayesian paradigm in Eq. (13) requires the user to specify a prior parameter distribution, $P(\boldsymbol{\theta})$, and the likelihood function, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$. The next two sections review these two antecedents.

4.1 The Prior Distribution, $P(\boldsymbol{\theta})$

The prior distribution should encode all the “subjective” knowledge about the parameters, $\boldsymbol{\theta}$, before collection of the data, $\tilde{\mathbf{Y}}$. This distribution, often simply called the prior, expresses one’s beliefs about the parameters before the data (also referred to as evidence) is taken into account. The work by Berger (1985) describes up to ten different techniques to construct the prior distribution.

In general, a prior distribution can be construed on the basis of findings reported in the literature or other publications, past experimental data collected in the laboratory or field, or other direct and indirect information. This can include “soft” data based on qualitative knowledge and understanding of processes and/or system behavior. A prior can also be elicited from expert judgment, or guided by principles of symmetry (scale invariance), or information-theoretical arguments (maximum entropy). Examples of the latter include a Jeffreys prior (Jeffreys 1946) and the reference prior (Bernardo 1979; Berger et al. 2009). Finally, for certain choices of the prior distribution, the posterior distribution has the same algebraic form, possibly

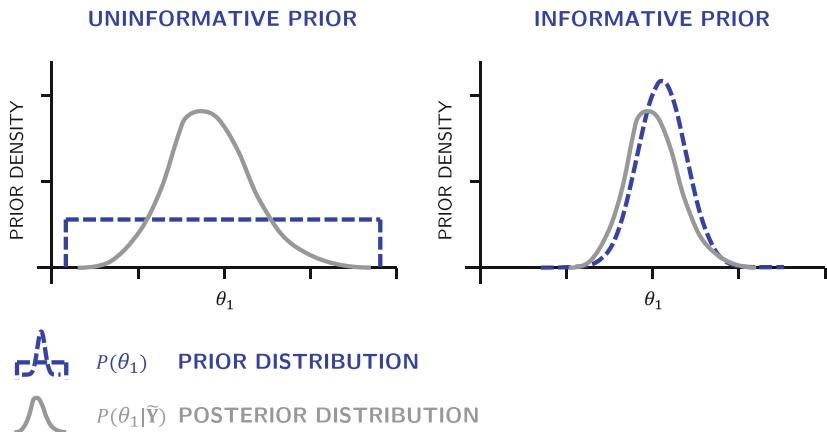


Fig. 3 The prior distribution, $P(\boldsymbol{\theta})$, reflects all knowledge about the entity $\boldsymbol{\theta}$ before collecting data, \tilde{Y} , (referred to as evidence) through field and/or laboratory experimentation. An uninformative prior (left) does not favor a-priori any particular value of the parameter, yet can elicit “objective” ranges. An informative prior distribution (right) expresses varying levels of favor to different values of the parameter. Some values of the parameter are deemed more plausible a-priori than other feasible solutions

with different parameter values. For example, if the likelihood function is Gaussian, then a normal prior over the mean will ensure that the posterior distribution is also Gaussian. Such priors, also called conjugate priors, can thus be deliberately chosen in lieu of analytical tractability. This avoids the need for numerical approximation of the posterior distribution using sampling methods (more of which later).

Prior distributions can be classified as informative or uninformative (or noninformative) depending on their information content for the model parameter or entity of interest. Figure 3 portrays graphically the two different types of prior distribution for some hypothetical model with just a handful of unknown parameters. Also shown is their anticipated effect on the marginal posterior distribution of the first parameter.

The uninformative prior expresses vague and general information about a variable. All its values are deemed equally likely a-priori. Yet, an uninformative prior can communicate objective information on the ranges of the quantity of interest. The classification “uninformative” is therefore somewhat of a misnomer, instead the wording diffuse prior might seem more appropriate. An informative prior expresses specific, or definite, information about a variable of interest. Such distribution voices preference to certain values of the parameter. The effect of this is visible on the posterior distribution as it appears more condensed with a informative prior than with a uninformative prior distribution. This is not always the case, certainly not if the prior distribution and the likelihood function are in disagreement on the statistical distribution of the parameters. The uninformative prior is also referred to as flat or uniform prior.

For the time being, we assume conveniently the use of a multivariate uniform prior distribution, $\mathcal{U}_d(\mathbf{a}, \mathbf{b})$, where \mathbf{a} and \mathbf{b} are d -vectors with lower and upper bound values of the d parameters of $\boldsymbol{\theta}$, respectively

$$P(\boldsymbol{\Theta}) \xrightarrow{\mathcal{D}} \mathcal{U}_d(\mathbf{a}, \mathbf{b}), \quad (14)$$

where the vectors \mathbf{a} and \mathbf{b} are defined by the d -dimensional hypercube, $\boldsymbol{\Theta}$, or $a_j = \min(\boldsymbol{\Theta}_j)$ and $b_j = \max(\boldsymbol{\Theta}_j)$ where $j = \{1, \dots, d\}$. The density of the uninformative prior distribution, $P(\boldsymbol{\Theta})$, in Eq. (14) is constant and independent of $\boldsymbol{\Theta}$.

The use of a multivariate prior distribution is appropriate when all the different parameters of the model have a similar marginal prior distribution. Examples of such priors include the multivariate Gaussian, $\mathcal{N}_d(\cdot)$, the multivariate uniform, $\mathcal{U}_d(\cdot)$, the multivariate Student, $\mathcal{T}_d(\cdot)$, the multivariate Gamma, and the Wishart distribution. All these constitute informative priors – with the exception of $\mathcal{U}_d(\cdot)$ (of course one can make each of these distributions almost flat by using infinitely large parameter variances). One of the advantages of a multivariate prior is that they can honor explicitly (linear) correlations among the model parameters (as off-diagonal terms in $d \times d$ covariance matrix). A disadvantage is that they assume the same type of marginal distribution for each of the parameters. This might not be appropriate for multivariate cases with differing individual priors. As alternative, we could use a different univariate distribution, $P(\theta_j) \sim \mathcal{X}(\cdot)$, for each of the parameters, $j = \{1, \dots, d\}$. This univariate approach enhances considerably our freedom in picking suitable marginal prior distributions, yet assumes parameter independence. The joint density of $P(\theta_j)$'s is therefore equivalent to the product of their individual densities, $P(\boldsymbol{\Theta}) = P(\theta_1) \times \dots \times P(\theta_d)$. This approach allows using at the same time informative and uninformative prior distributions. For the uniform case of Eq. (14), the joint density can be written as

$$P(\boldsymbol{\Theta}) = P(\theta_1) \times \dots \times P(\theta_d) = c_1 \times \dots \times c_d \propto 1 \quad (15)$$

where each constant, c_i ; $i = \{1, \dots, d\}$, is equivalent to the reciprocal of the range of each parameter. This ensures that the prior distribution, $P(\boldsymbol{\Theta})$, integrates to one, $\int_{\boldsymbol{\Theta}} P(\boldsymbol{\Theta}) = 1$, and as such is a formal probability distribution. The effect of normalization is clearly visible in Fig. 3 wherein the larger support of the uninformative (uniform) prior distribution (on right) is counteracted by a density that is much smaller than that of the informative Gaussian distribution (on left). In practice, however, we are allowed to work with “improper” priors that do not integrate to unity, as long as we are focused solely on inference of the model parameters. For multiple different working hypotheses, the prior must integrate to one, otherwise the marginal likelihood, $P(\tilde{\mathbf{Y}})$, is corrupted and we cannot proceed with model selection.

4.2 The Likelihood Function, $L(\boldsymbol{\Theta}|\tilde{\mathbf{Y}})$

Now the prior distribution has been defined, we are left with the definition of the likelihood function, $L(\boldsymbol{\Theta}|\tilde{\mathbf{Y}})$. This function summarizes, in a probabilistic sense, the compatibility of the n observed data, $\tilde{\mathbf{Y}}$, and the n model output, $\mathbf{Y} = \mathcal{M}(\boldsymbol{\Theta})$ simulated by the parameter values, $\boldsymbol{\Theta}$. Likelihood functions play a key role in statistical inference, and the word “likelihood” is used often as synonym for

“probability” – yet, in practice, the word *probability* is appropriate when describing possible future outcomes for fixed parameter values before data are available. The use of *likelihood*, on the contrary, is appropriate to describe a function of a parameter vector for a given outcome after data are available. If we want to quantify the likelihood of an outcome b in light of some observation, a , then we need to define a probability density function, $f_a(b)$, for the entity a . For example, if $f_a(b)$ is a normal distribution, $f_a(b) \sim \mathcal{N}(a, c)$, then we can calculate the density of this distribution at our outcome b once the (measurement error) variance c of the observation a is known. This computation of the likelihood is easily generalized to a vector of observations and leaves us with n likelihoods of $\mathbf{Y}(\boldsymbol{\theta})$ evaluated at $\tilde{\mathbf{Y}}$ using the different distributions of $f_{\tilde{y}_j}(y_j)$, where $j = \{1, \dots, n\}$. In practice, it is more insightful to express the likelihood as a function of the residuals, $\mathbf{E}(\boldsymbol{\theta})$, instead rather than the observed and simulated values. This does not affect the actual likelihood values, except centers $f_a(b)$ on $a = 0$ and uses as entry the residual, $b = a - b$, between the observation and outcome. We therefore use instead the notation $f(e_j(\boldsymbol{\theta}))$ in the remainder of this chapter.

We cannot proceed further without making some important assumptions regarding the dependence structure of the n different residuals (and thus likelihoods). For the time being, we assume conveniently that the residuals are serially uncorrelated. Then the joint likelihood, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$, can be written in multiplicative form as follows

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) = f(e_1(\boldsymbol{\theta})) \times \dots \times f(e_n(\boldsymbol{\theta})) = \prod_{t=1}^n f(e_t(\boldsymbol{\theta})), \quad (16)$$

where $f(b)$ signifies the zero-mean probability density function evaluated at b . If serial correlation is expected among the residuals, then this can be accounted for with a filter in the computation of the joint probability density of the n different “events” (likelihoods). Unless known, the coefficients of the filter join the inference of $\boldsymbol{\theta}$ as nuisance variables. We will revisit this topic in a later part of this section.

Now we have an expression for the joint likelihood of the residuals (and thus parameters, $\boldsymbol{\theta}$), we need to make an assumption regarding the statistical distribution of the different f s. In the ideal case with a perfect model, input data, and initial states, we certainly expect the residual distribution to match perfectly the distribution of the measurement errors of the system response, $\tilde{\mathbf{Y}}$. A typical assumption is that these measurement errors follow a Gaussian distribution, $\mathcal{N}(0, \hat{\sigma}^2)$, with constant variance $\hat{\sigma}^2$ – and thus $f(e_j(\boldsymbol{\theta})) \stackrel{\mathcal{D}}{\sim} \mathcal{N}(0, \hat{\sigma}^2) \forall j \in \{1, \dots, d\}$. If we substitute $\mathcal{N}(0, \hat{\sigma}^2)$ in Eq. (16), then the joint likelihood of the n residuals is simply equivalent to the product of n normal densities

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\sigma}^2) = \prod_{t=1}^n \frac{1}{\sqrt{2\pi\hat{\sigma}^2}} \exp \left[-\frac{1}{2} \left(\frac{\tilde{y}_t - y_t(\boldsymbol{\theta})}{\hat{\sigma}} \right)^2 \right]. \quad (17)$$

Mathematics teaches us that

$$\prod_{t=1}^n \frac{1}{a} = a^{-n} \text{ and } \prod_{t=1}^n \exp(-a_t) = \exp\left(-\sum_{t=1}^n a_t\right) \quad (18)$$

so we can simplify Eq. (17) to read

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\sigma}^2) = \left(\sqrt{2\pi\hat{\sigma}^2}\right)^{-n} \exp\left(-\frac{1}{2}\hat{\sigma}^{-2} \sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right), \quad (19)$$

where the summation term in the exponent is equivalent to the SSR used commonly as objective function, $F(\boldsymbol{\theta})$, in model calibration. The better the model fits the data, the larger the value of the joint likelihood, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\sigma}^2)$, in Eq. (19). The vector of parameter values that maximizes the likelihood function is also referred to as maximum likelihood (ML) solution.

The formulation of the likelihood function of Eq. (19) can suffer from arithmetic underflow, that is, finite multiplication can result in a number that is so close to zero that the computer cannot store this in memory. This can already happen for relatively small n , say $n = 500$, particularly if the model describes poorly the observed data, and the residuals are large compared to the measurement error standard deviation, $|e_t(\boldsymbol{\theta})| \gg \hat{\sigma}$, for many elements $t \in \{1, \dots, n\}$. For reasons of numerical stability, it is therefore convenient to work with the log-likelihood, $\mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\sigma}^2)$, instead

$$\mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\sigma}^2) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2}n\log(\hat{\sigma}^2) - \frac{1}{2}\hat{\sigma}^{-2} \sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2. \quad (20)$$

This log-likelihood formulation is arguably also easier to interpret algebraically. The value of $\hat{\sigma}^2$ can be defined a-priori by the user or alternatively its value can be inferred simultaneously with the parameters, $\boldsymbol{\theta}$. As last resort, we can “integrate out” (This wording may not characterize accurately the step-by-step mathematical procedure, yet is consistent with the wording used by Box and Tiao (1992) and used in Appendix A of Kavetski et al. (2006a).) the measurement error variance in Eq. (20) using as proxy for $\hat{\sigma}^2$ the variance s^2 of the error residuals

$$s^2 = \frac{1}{n-1} \sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2. \quad (21)$$

We can substitute for $\hat{\sigma}^2$ the sufficient statistic s^2 . This gives us

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) &= -\frac{1}{2}n\log(2\pi) - \frac{1}{2}n\log\left(\frac{1}{n-1} \sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right) \\ &\quad - \frac{1}{2} \sum_{t=1}^n \frac{(\tilde{y}_t - y_t(\boldsymbol{\theta}))^2}{\frac{1}{n-1} \sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2}, \end{aligned} \quad (22)$$

and with $\log(ab) = \log(a) + \log(b)$, this results in

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) &= -\frac{1}{2}n\log(2\pi) - \frac{1}{2}n\log\left(\frac{1}{n-1}\right) - \frac{1}{2}n\log\left(\sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right) \\ &\quad - \frac{1}{2}(n-1). \end{aligned} \quad (23)$$

A further simplification of

$$\frac{1}{2}n\log\left(\frac{1}{n-1}\right) = \frac{1}{2}n(\log(1) - \log(n-1)) = -\frac{1}{2}n\log(n-1) \quad (24)$$

leads to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) &= -\frac{1}{2}n\log(2\pi) + \frac{1}{2}n\log(n-1) - \frac{1}{2}n\log\left(\sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right) \\ &\quad - \frac{1}{2}(n-1) \end{aligned} \quad (25)$$

We can safely discard those terms from Eq. (25) that are independent of $\boldsymbol{\theta}$. These terms are normalization constant of $\mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$. This includes the first, second, and fourth term at the right hand side, respectively. Without these constants, the log-likelihood function reads

$$\mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) \propto -\frac{1}{2}n\log\left(\sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right) \quad (26)$$

and the proportionality sign is used as expression for the unnormalized likelihood. This equation is equivalent to

$$\mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) \propto \log\left(\sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right)^{-\frac{1}{2}n}. \quad (27)$$

If we are interested in the actual likelihood, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$, we end up with

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) \propto \left(\sum_{t=1}^n (\tilde{y}_t - y_t(\boldsymbol{\theta}))^2\right)^{-\frac{1}{2}n}, \quad (28)$$

which is similar to

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) \propto \sum_{t=1}^n |\tilde{y}_t - y_t(\boldsymbol{\theta})|^{-n} \quad (29)$$

If we now apply Bayes theorem, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, and assume a uniform prior distribution of the parameters, then the posterior density is equivalent to

$$P(\boldsymbol{\theta}|\tilde{\mathbf{Y}}) \propto \sum_{t=1}^n |\tilde{y}_t - y_t(\boldsymbol{\theta})|^{-n}, \quad (30)$$

which is equivalent to Eq. (23) of Thiemann et al. (2001). This concludes the derivation.

The derivation of Eq. (30) assumes that the measurement errors of the system response data, $\tilde{\mathbf{Y}}$, exhibit a constant variance, $\hat{\sigma}^2$. This assumption may be appropriate for variables such as temperature and pressure which are known to have a homoscedastic measurement error. This assumption, however, is not justified for entities such as windspeed and discharge as the variance of their measurement error increases with their magnitude. We can modify Eq. (17) to take into account a nonconstant measurement error variance

$$L(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \hat{\boldsymbol{\sigma}}^2) = \prod_{t=1}^n \frac{1}{\sqrt{2\pi\hat{\sigma}_t^2}} \exp\left[-\frac{1}{2}\left(\frac{\tilde{y}_t - y_t(\boldsymbol{\theta})}{\hat{\sigma}_t}\right)^2\right], \quad (31)$$

where $\hat{\boldsymbol{\sigma}}^2 = \{\hat{\sigma}_1^2, \dots, \hat{\sigma}_n^2\}$ denotes the n -vector of measurement error variances. The log-likelihood of Eq. (31) now becomes

$$\mathcal{L}(\boldsymbol{\theta}|\tilde{\mathbf{Y}}, \hat{\boldsymbol{\sigma}}^2) = -\frac{1}{2}n\log(2\pi) - \frac{1}{2} \sum_{t=1}^n \{\log(\hat{\sigma}_t^2)\} - \frac{1}{2} \sum_{t=1}^n \left(\frac{(\tilde{y}_t - y_t(\boldsymbol{\theta}))^2}{\hat{\sigma}_t^2}\right). \quad (32)$$

This allows for a different value of the measurement error variance for each observation of $\tilde{\mathbf{Y}}$.

Now lets imagine a situation in which the residuals $\mathbf{E}(\boldsymbol{\theta}) = \tilde{\mathbf{Y}} - \mathbf{Y}(\boldsymbol{\theta}) = \{e_1(\boldsymbol{\theta}), \dots, e_n(\boldsymbol{\theta})\}$ exhibit temporal correlation. This serial correlation, also known as autocorrelation, can exist between values at different times, as a function of the two times, or of the time lag. Lets assume conveniently that the residual mean, $\mu_{\mathbf{E}(\boldsymbol{\theta})}$, and variance, $\sigma_{\mathbf{E}(\boldsymbol{\theta})}^2$ are time invariant. Then, $\mathbf{E}(\boldsymbol{\theta})$ is a wide-sense stationary process, and the (auto)correlation of two residuals, $e_i(\boldsymbol{\theta})$ and $e_j(\boldsymbol{\theta})$, is a function only of the time lag $k = i - j$ between i and j ($i \geq j$). The (auto) correlation coefficient, $\rho(k) \in [-1, 1]$ is then mathematically defined as follows

$$\begin{aligned} \rho(k) &= \frac{\mathbb{E}\left[\left(e_t(\boldsymbol{\theta}) - \mu_{\mathbf{E}(\boldsymbol{\theta})}\right)\left(e_{t+k}(\boldsymbol{\theta}) - \mu_{\mathbf{E}(\boldsymbol{\theta})}\right)\right]}{\sigma_{\mathbf{E}(\boldsymbol{\theta})}^2} \\ &= \frac{\sum_{t=1}^{n-k} \left(e_t(\boldsymbol{\theta}) - \mu_{\mathbf{E}(\boldsymbol{\theta})}\right)\left(e_{t+k}(\boldsymbol{\theta}) - \mu_{\mathbf{E}(\boldsymbol{\theta})}\right)}{\sum_{t=1}^{n-k} \left(e_t(\boldsymbol{\theta}) - \mu_{\mathbf{E}(\boldsymbol{\theta})}\right)^2}, \end{aligned} \quad (33)$$

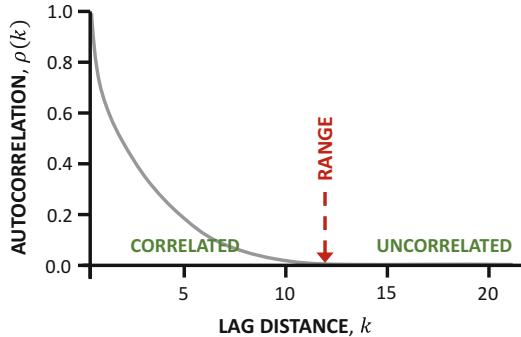


Fig. 4 Illustrative example of the autocorrelation function of some hypothetical residual time series, $E(\boldsymbol{\theta})$. The serial residual correlation, $\rho(k)$, is computed at different lags $k = \{1, \dots, 22\}$ using Eq. (33). These values are subsequently connected using the gray line. Per definition, the zeroth-order correlation is equal to unity, or $\rho(0) = 1$. This is easily shown as $e_t(\boldsymbol{\theta}) = e_{t+0}(\boldsymbol{\theta})$ and so the numerator in Eq. (33) simplifies to $\sigma_{E(\boldsymbol{\theta})}^2$, the denominator

where $\mu_{E(\boldsymbol{\theta})} = \frac{1}{n} \sum_{t=1}^n e_t(\boldsymbol{\theta})$. A value of $\rho(k) = 1$ indicates perfect positive correlation, whereas a value of $\rho(k) = -1$ signifies perfect anticorrelation. Two words of caution. First, the (auto)correlation coefficient characterizes only linear relationships between the residuals. Second, the residual variance, $\sigma_{E(\boldsymbol{\theta})}^2$, must be stable (homogeneity assumption) and larger than zero.

The value of $\rho(k)$ is also referred to as the k th-order (auto)correlation, and can be depicted graphically in a so-called autocorrelation function (see Fig. 4).

The autocorrelation is particularly significant at the first few lags but then deteriorates rapidly at larger distances between the residuals. Residuals that are more than $k = 11$ lags apart appear, on average, uncorrelated. Thus, the temporal correlation, $\rho(k)$, of the residuals depends only on the “distance” between two residuals and not on their actual position in the time series of $E(\boldsymbol{\theta})$.

We can take into explicit account serial correlation of the residuals in the derivation of the log-likelihood function. For example, let's suppose that the residuals exhibit correlation at the first lag, that is $k = 1$. We can write this serial correlation as an AR(1)-process

$$e_t(\boldsymbol{\theta}) = c + \phi_1 e_{t-1}(\boldsymbol{\theta}) + \epsilon_t, \quad (34)$$

where c signifies the bias, $\phi_1 \in [-1, 1]$ is the first-order correlation coefficient, and ϵ_t denotes the remaining errors, $\epsilon(\boldsymbol{\theta}) = \{\epsilon_1, \dots, \epsilon_n\}$, hereafter also referred to as the “decorrelated” residuals. If we assume that $\epsilon_t \stackrel{\mathcal{D}}{\sim} \mathcal{N}(0, \hat{\sigma}_t^2)$, then the expectation is $\mathbb{E}[e_t(\boldsymbol{\theta})] = c/(1 - \phi_1)$, and central dispersion is $\text{Var}[e_t(\boldsymbol{\theta})] = \hat{\sigma}^2 / (1 - \phi_1^2)$. To illustrate the effect of autocorrelation, please consider Fig. 5 which displays two different residual time series with (orange line) and without (green line) first-order serial correlation.

The differences between the two residual vectors are evident. The uncorrelated residuals have a zero-mean, jump up and down the $\epsilon = 0$ line (dotted gray line), and

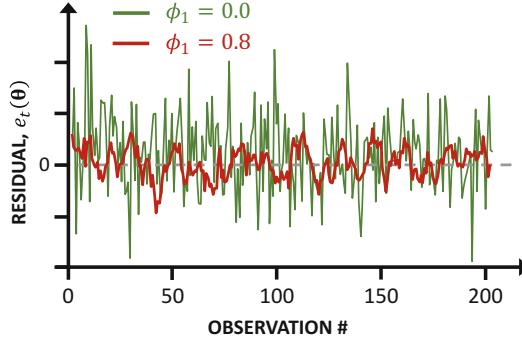


Fig. 5 The impact of first-order serial correlation on the residuals. The green line lacks serial correlation, hence $\phi_1 = 0$, whereas the orange time series of residuals exhibits strong autocorrelation at the first lag, $\phi_1 = 0.8$. The dashed gray line signifies a perfect match between the model and the data. Note $c = 0$ and $\epsilon_t = 0 \forall t \in \{1, \dots, 200\}$

do not display any “collective” memory or consciousness. The correlated residuals, on the contrary, show a much strong memory effect with neighboring residuals that take on very similar values.

Lets assume that the residuals do not exhibit a systematic bias, thus $c = 0$. Per Eq. (34) with $c = 0$, the values of ϵ_t are equivalent to

$$\begin{aligned} \epsilon_t(\boldsymbol{\theta}, \phi_1) &= e_t(\boldsymbol{\theta}) - \phi_1 e_{t-1}(\boldsymbol{\theta}) \\ &= (\tilde{y}_t - y_t(\boldsymbol{\theta})) - \phi_1 (\tilde{y}_{t-1} - y_{t-1}(\boldsymbol{\theta})), \end{aligned} \quad (35)$$

with $\text{Var}[(\tilde{y}_t - y_t(\boldsymbol{\theta})) - \phi_1 (\tilde{y}_{t-1} - y_{t-1}(\boldsymbol{\theta}))] = \text{Var}[\epsilon_t(\boldsymbol{\theta}, \phi_1)] = \hat{\sigma}^2$. As we do not have available \tilde{y}_0 , we cannot derive in a single step the log-likelihood function of the decorrelated residuals, ϵ . Instead, lets ignore for now $y_1(\boldsymbol{\theta})$ and focus on the last $n - 1$ simulated values of $\mathbf{Y}(\boldsymbol{\theta})$. Per Eq. (32), the log-likelihood of $\{\epsilon_2(\boldsymbol{\theta}, \phi_1), \dots, \epsilon_n(\boldsymbol{\theta}, \phi_1)\}$ equates to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \{\tilde{y}_2, \dots, \tilde{y}_n\}, \phi_1, \{\hat{\sigma}_2^2, \dots, \hat{\sigma}_n^2\}) &= -\frac{1}{2}(n-1)\log(2\pi) - \frac{1}{2}\sum_{t=2}^n \{\log(\hat{\sigma}_t^2)\} \\ &\quad - \frac{1}{2}\sum_{t=2}^n \left(\frac{((\tilde{y}_t - y_t(\boldsymbol{\theta})) - \phi_1 (\tilde{y}_{t-1} - y_{t-1}(\boldsymbol{\theta})))^2}{\hat{\sigma}_t^2} \right). \end{aligned} \quad (36)$$

which is identical to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \{\tilde{y}_2, \dots, \tilde{y}_n\}, \phi_1, \{\hat{\sigma}_2^2, \dots, \hat{\sigma}_n^2\}) &= -\frac{1}{2}(n-1)\log(2\pi) - \frac{1}{2}\sum_{t=2}^n \{\log(\hat{\sigma}_t^2)\} \\ &\quad - \frac{1}{2}\sum_{t=2}^n \left(\frac{\epsilon_t^2(\boldsymbol{\theta})}{\hat{\sigma}_t^2} \right). \end{aligned} \quad (37)$$

We are now left with the log-likelihood of the first simulated value, $y_1(\boldsymbol{\theta})$, of $\mathbf{Y}(\boldsymbol{\theta})$. We know that the $\text{Var}[e_1(\boldsymbol{\theta})] = \hat{\sigma}_1^2 / (1 - \phi_1^2)$ and thus the log-likelihood of $y_1(\boldsymbol{\theta})$ is

$$\mathcal{L}(\boldsymbol{\theta} | \tilde{y}_1, \phi_1, \hat{\sigma}_1^2) = -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log\left(\frac{\hat{\sigma}_1^2}{(1 - \phi_1^2)}\right) - \frac{1}{2} \left(\frac{(\tilde{y}_1 - y_1(\boldsymbol{\theta}))^2}{\hat{\sigma}_1^2 / (1 - \phi_1^2)}\right). \quad (38)$$

As $\log(a/b) = \log(a) - \log(b)$, this equation can be rearranged and simplified to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{y}_1, \phi_1, \hat{\sigma}_1^2) = & -\frac{1}{2} \log(2\pi) - \frac{1}{2} \log(\hat{\sigma}_1^2) + \frac{1}{2} \log(1 - \phi_1^2) \\ & - \frac{1}{2} (1 - \phi^2) \hat{\sigma}_1^{-2} (\tilde{y}_1 - y_1(\boldsymbol{\theta}))^2. \end{aligned} \quad (39)$$

The joint log-likelihood of the first-order correlated residuals, $\mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \phi_1, \hat{\boldsymbol{\sigma}}^2)$, is now equivalent to the sum of $\mathcal{L}(\boldsymbol{\theta} | \tilde{y}_1, \phi_1, \hat{\sigma}_1^2)$ and $\mathcal{L}(\boldsymbol{\theta} | \{\tilde{y}_2, \dots, \tilde{y}_n\}, \phi_1, \{\hat{\sigma}_2^2, \dots, \hat{\sigma}_n^2\})$ which yields

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \phi_1, \hat{\boldsymbol{\sigma}}^2) = & -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \{\log(\hat{\sigma}_t^2)\} + \frac{1}{2} \log(1 - \phi_1^2) \\ & - \frac{1}{2} (1 - \phi^2) \hat{\sigma}_1^{-2} (\tilde{y}_1 - y_1(\boldsymbol{\theta}))^2 - \frac{1}{2} \sum_{t=2}^n \left(\frac{((\tilde{y}_t - y_t(\boldsymbol{\theta})) - \phi_1(\tilde{y}_{t-1} - y_{t-1}(\boldsymbol{\theta})))^2}{\hat{\sigma}_t^2} \right). \end{aligned} \quad (40)$$

This equation reduces to the Gaussian likelihood function of Eq. (32) if $\phi_1 = 0$, that is, the residuals, $\mathbf{e}(\boldsymbol{\theta})$, do not exhibit serial correlation. We can generalize further Eq. (40) by explicit treatment of bias in the residuals. We will not present this derivation herein as it follows exactly the previous steps, except that $c \neq 0$. The log-likelihood becomes

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, c, \phi_1, \hat{\boldsymbol{\sigma}}^2) = & -\frac{1}{2} n \log(2\pi) - \frac{1}{2} \sum_{t=1}^n \{\log(\hat{\sigma}_t^2)\} + \frac{1}{2} \log(1 - \phi_1^2) \\ & - \frac{1}{2} (1 - \phi^2) \hat{\sigma}_1^{-2} \left(\tilde{y}_1 - y_1(\boldsymbol{\theta}) - \underbrace{[c/(1 - \phi)]}_{\mathbb{E}[e_1(\boldsymbol{\theta})] = c/(1 - \phi_1)} \right)^2 \\ & - \frac{1}{2} \sum_{t=2}^n \left(\frac{((\tilde{y}_t - y_t(\boldsymbol{\theta})) - c - \phi_1(\tilde{y}_{t-1} - y_{t-1}(\boldsymbol{\theta})))^2}{\hat{\sigma}_t^2} \right). \end{aligned} \quad (41)$$

This concludes the derivation of the different likelihood functions. The nuisance variables (A nuisance variable is a random variable that is fundamental to the probabilistic model but that is not of particular itself.) ϕ_1 , c , and the n -vector of measurement error variances, $\hat{\boldsymbol{\sigma}}^2$, can be defined a-priori by the user, or alternatively,

their values can be inferred jointly with the parameters, $\boldsymbol{\theta}$. If the system response data, $\tilde{\mathbf{Y}}$, exhibit heteroscedastic measurement errors, then this could be a pitfall as the number of unknown variables of $\hat{\sigma}^2$ grows linearly with n . A pragmatic remedy to this problem is to relate $\hat{\sigma}_t$ to the measured data, \tilde{y}_t , using some predefined measurement error model

$$\hat{\sigma}_t = \sigma_0 + \sigma_1 \tilde{y}_t, \quad (42)$$

where $\sigma_0 > 0$ and $\sigma_1 \in [0, 1]$ are two unknown coefficients that define the intercept and slope of the measurement error function. This approach reduces the number of nuisance variables to four, that is, $\boldsymbol{\alpha} = \{\sigma_0, \sigma_1, c, \phi_1\}$. Nonlinear measurement error models can be used as well, whatever is deemed appropriate in practice. The inference thus involves the estimation of the posterior distribution of $\{\boldsymbol{\theta}, \boldsymbol{\alpha}\}$ using the model, $\mathcal{M}(\boldsymbol{\theta})$, and observed data, $\tilde{\mathbf{Y}}$.

4.3 Generalized Likelihood Function

The most literate likelihood function of the previous section assumes Gaussian distributed residuals with a constant bias, c , and low-order serial dependence, ϕ_1 . Experience suggests that these assumptions are not particularly adequate for real-world studies involving numerical models of complex systems. Indeed, diagnostic checks often demonstrate a considerable variation in bias, variance, and serial dependence of the residuals for different parts of the model response. Such nontraditional residual distributions are a response to model structural and forcing data errors. The assumption that these errors are negligibly small or somehow absorbed into the output residual is mathematically convenient but questionable at best. Indeed, the impact of model structural and forcing (control input) data errors may, in general, be much larger than the measurement error of the calibration data, $\tilde{\mathbf{Y}}$.

Two different approaches can be found in the literature to improve the handling of nontraditional residual distributions. The first approach relaxes the common assumptions of normality and first-order serial dependence in lieu of a more flexible description of the probabilistic properties of the residuals. Examples of this approach in hydrologic modeling can be found in Sorooshian and Dracup (1980) and Kuczera (1983), and different groups of researchers have build further on this early work (Bates and Campbell 2001; Yang et al. 2007; Schoups and Vrugt 2010; Smith et al. 2010; Evin et al. 2013; Scharnagl et al. 2015). Nuisance variables (Kavetski et al. 2006a; Vrugt et al. 2008; Renard et al. 2011) or time-variable parameters (Kuczera et al. 2006; Reichert and Mieleitner 2009) can be used to treat explicitly errors in the model structure and forcing data.

The second approach abandons the classical Bayesian paradigm in lieu of a likelihood-free methodology which uses summary metrics of the data instead. This alternative approach is geared towards the detection of model structural error, to illuminate how the model should be improved for the purpose of learning and scientific discovery. This diagnostic approach will be discussed in the penultimate

section of this chapter. In this section, we briefly review work on extending the applicability of previously used likelihood functions to situations where residual errors are correlated, heteroscedastic, and non-Gaussian with varying degrees of kurtosis and skewness.

The likelihood functions discussed in the previous section assume first-order correlated, Gaussian distributed, residuals with constant bias. We can relax the assumption of normality of the residuals by using an alternative, and more flexible, statistical distribution whose functional shape is guided by the residuals, $\mathbf{E}(\boldsymbol{\theta})$. If we add a shape parameter, β , to the normal distribution, we derive the generalized normal distribution or exponential power (EP) distribution. The density of the standardized EP distribution (zero-mean and unit standard deviation) at point a is equivalent to (Box and Tiao 1992)

$$f_{\text{EP}}(a) = \omega_\beta \exp\left(-c_\beta |a|^{2/(1+\beta)}\right), \quad (43)$$

where $\beta \in (-1, 1)$ is the kurtosis parameter, which determines the peakedness of the distribution. The values of ω_β and c_β are given by (Box and Tiao 1992)

$$\omega_\beta = \frac{\Gamma^{1/2}[3(1+\beta)/2]}{(1+\beta)\Gamma^{3/2}[(1+\beta)/2]} \quad c_\beta = \left(\frac{\Gamma[3(1+\beta)/2]}{\Gamma[(1+\beta)/2]}\right)^{1/(1+\beta)}, \quad (44)$$

where $\Gamma[b]$ signifies the incomplete Gamma function evaluated at b

$$\Gamma[b] = \int_0^\infty x^{b-1} \exp(-x) dx \quad \forall b \in \mathbb{R}_+, \quad (45)$$

which satisfies the recursion $\Gamma(b+1) = b\Gamma(b)$.

The EP density, $f(a)$, in Eq. (43) is symmetric around $a = 0$. This symmetry impairs our ability to describe accurately skewed residual distributions with an upper or lower tail. Fernandez and Steel (1998) developed a general solution for the treatment of skew in symmetric distributions with closed-form mathematical expressions. They presented a template density, $f_{\text{skew}}(a)$ with skewness parameter, $\xi \in \mathbb{R}_+$

$$f_{\text{skew}}(a) = \frac{2}{\xi + \xi^{-1}} f\left(a\xi^{-\text{sgn}(a)}\right) \quad (46)$$

to introduce (heavy) tails in a symmetric probability density function, $f(\cdot)$, where “sgn” denotes the signum function

$$\text{sgn}(a) = \begin{cases} 1 & \text{if } a > 0 \\ 0 & \text{if } a = 0 \\ -1 & \text{if } a < 0 \end{cases} \quad (47)$$

The mean and standard deviation of a in Eq. (46) can be computed using (Fernandez and Steel 1998)

$$\mu_\xi = M_1(\xi - \xi^{-1}) \quad \sigma_\xi = \sqrt{(M_2 - M_1^2)(\xi^2 + \xi^{-2}) + 2M_1^2 - M_2} \quad (48)$$

wherein M_r is the r th absolute moment of the symmetric density, $f(\cdot)$,

$$M_r = 2 \int_0^\infty s^r f(s) ds \quad (49)$$

For the standardized EP density, $f_{EP}(\cdot)$, in Eq. (43), this results in the following expressions of M_1 and M_2 in Eq. (48)

$$M_1 = \frac{\Gamma[1+\beta]}{\Gamma^{1/2}[3(1+\beta)/2]\Gamma^{1/2}[(1+\beta)/2]} \quad M_2 = 1 \quad (50)$$

We can now substitute the exponential power density, $f_{EP}(\cdot)$, of Eq. (43) into the skew distribution, $f_{skew}(\cdot)$, of Eq. (46). Before so doing, we first must standardize the resulting skewed exponential power, or SEP, distribution. To obtain a zero mean and unit standard deviation, we scale the SEP distribution with σ_ξ and replace a by $\mu_\xi + \sigma_\xi a$ in $f_{skew}(\cdot)$ of Eq. (46) to yield

$$f_{SEP}(a) = \frac{2\sigma_\xi}{\xi + \xi^{-1}} \omega_\beta \exp\left(-c_\beta \left|\frac{\mu_\xi + \sigma_\xi a}{\xi^{\text{sgn}(\mu_\xi + \sigma_\xi a)}}\right|^{2/(1+\beta)}\right), \quad (51)$$

where c_β , ω_β , μ_ξ , and σ_ξ are a function of the kurtosis parameter, β , and skewness parameter, ξ , using Eqs. (44) and (48), respectively.

Figure 6 plots the SEP density of Eq. (51) for different values of the skewness, β and kurtosis, ξ , parameter.

The density is symmetric for $\xi = 1$, positively skewed for $\xi > 1$ and negatively skewed for $\xi < 1$. For a symmetric density, that is $\xi = 1$, a value of $\beta = -1$ results in a uniform distribution, $\beta = 0$ produces a normal distribution, and $\beta = 1$ portrays a double-exponential or Laplace distribution. Thus, a value of $\beta \in (0, 1]$ produces a SEP distribution with (much) heavier tails than the normal density. This is of great importance as it makes the inference of the model parameters, Θ , more robust against outliers.

The parameters β and ξ of the f_{SEP} distribution thus allow us to relax the normality assumption of the residuals and mimic accurately nontraditional densities with different levels of skew and kurtosis. Before we can proceed further with derivation of the SEP likelihood function, we need to decide how we treat serial dependence of the residuals. Schoups and Vrugt (2010) used the following recursive model of the residuals, $\mathbf{E}(\Theta)$

$$\Phi_p(B)e_t(\Theta) = \hat{\sigma}_t \varepsilon_t \quad (52)$$

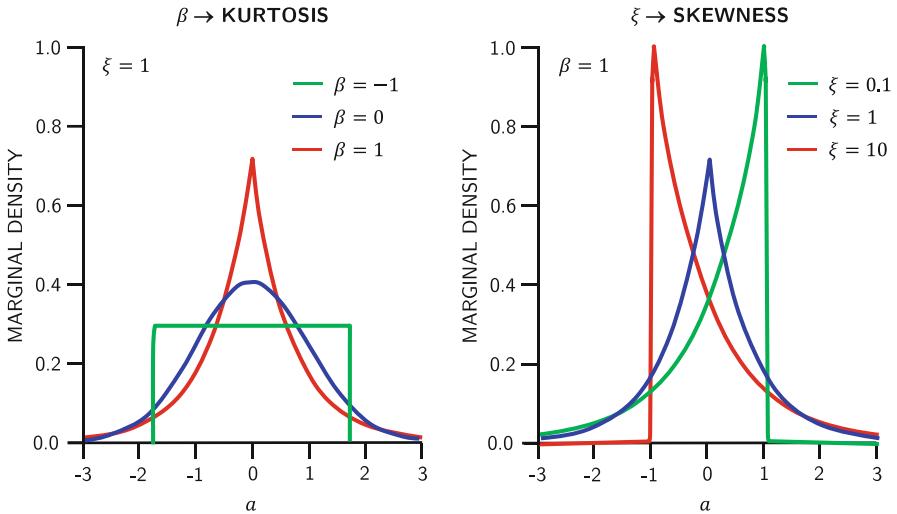


Fig. 6 Densities of the skew exponential power (SEP) distribution with zero mean and unit standard deviation for various values of the kurtosis, β , and skewness, ξ , parameters

where B represents the backward shift (“backshift”) operator, that is $B^i e_t(\boldsymbol{\theta}) = e_{t-i}(\boldsymbol{\theta})$, $\hat{\sigma}_t$ signifies the measurement error standard deviation of observation \tilde{y}_t of $\tilde{\mathbf{Y}}$, ε_t is the remaining (uncorrelated) error, and

$$\Phi_p(B) = 1 - \sum_{i=1}^p \phi_i B^i \quad (53)$$

is an autoregressive polynomial with p coefficients, $\Phi_p = \{\phi_1, \dots, \phi_p\}$, where $\phi_j \in [-1, 1]$ and $j = \{1, \dots, p\}$. Note that $B^1 e_t(\boldsymbol{\theta}) = e_{t-1}(\boldsymbol{\theta})$ and $B^2 e_t(\boldsymbol{\theta}) = B(Be_t(\boldsymbol{\theta})) = e_{t-2}(\boldsymbol{\theta})$, and so forth. If we now assume that each decorrelated residual, ε_t , is distributed according to a zero-mean and unit standard deviation SEP distribution with skewness β and kurtosis ξ , or $\varepsilon_t \stackrel{\mathcal{D}}{\sim} f_{\text{SEP}}(0, 1, \beta, \xi)$, then the expectation of $\hat{\sigma}_t \varepsilon_t$ in Eq. (52) is $\mathbb{E}[\hat{\sigma}_t \varepsilon_t(\boldsymbol{\theta}, \Phi_p)] = 0$, and the central dispersion is equivalent to $\text{Var}[\hat{\sigma}_t \varepsilon_t(\boldsymbol{\theta}, \Phi_p)] = \hat{\sigma}_t^2 \text{Var}[\varepsilon_t(\boldsymbol{\theta}, \Phi_p)] = \hat{\sigma}_t^2 \times 1 = \hat{\sigma}_t^2$.

If we assume independence of the different ε_t 's, then the joint likelihood of the n -vector of $\hat{\sigma}_t \varepsilon_t$'s in Eq. (52) can be written in multiplicative form (see Eq. 16)

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) = \prod_{t=1}^n f_{\text{SEP}}(\hat{\sigma}_t \varepsilon_t(\boldsymbol{\theta}, \Phi_p)), \quad (54)$$

where the t th SEP density is given by $f_{\text{SEP}}(0, \sigma_b, \beta, \xi)$. This formulation of $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$ poses computational difficulties as we do not have available a mathematical expression for the nonstandardized SEP distribution with non-unit variance, $\hat{\sigma}_t^2$. Fortunately, we can take advantage of a simple trick, that is the solution of $f_{\text{SEP}}(0, \hat{\sigma}_t, \beta, \xi)$ at $\hat{\sigma}_t \varepsilon_t$ is equivalent to $\hat{\sigma}_t^{-1} f_{\text{SEP}}(0, 1, \beta, \xi)$ evaluated at $\varepsilon_t / \hat{\sigma}_t$. This leads to the following formulation of the likelihood function

$$L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) = \prod_{t=1}^n \hat{\sigma}_t^{-1} f_{\text{SEP}}(\varepsilon_t(\boldsymbol{\theta}, \Phi_p)). \quad (55)$$

We can now substitute for $f_{\text{SEP}}(\cdot)$ in Eq. (51) to yield the following formulation of the likelihood function, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\boldsymbol{\sigma}}, \beta, \xi, \Phi_p)$ (Schoups and Vrugt 2010)

$$\begin{aligned} L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\boldsymbol{\sigma}}, \beta, \xi, \Phi_p) &\simeq \prod_{t=1}^n \hat{\sigma}_t^{-1} \\ &\times \frac{2\sigma_\xi}{\xi + \xi^{-1}} \omega_\beta \exp \left(-c_\beta \left| \frac{\mu_\xi + \sigma_\xi \varepsilon_t(\boldsymbol{\theta}, \Phi_p)}{\xi \text{sgn}(\mu_\xi + \sigma_\xi \varepsilon_t(\boldsymbol{\theta}, \Phi_p))} \right|^{2/(1+\beta)} \right), \end{aligned} \quad (56)$$

where c_β , ω_β , μ_ξ , and σ_ξ , are a function of the kurtosis parameter, β , and skewness parameter, ξ , using Eqs. (44) and (48), respectively, and $\boldsymbol{\alpha} = \{\beta, \xi, \Phi_p\}$ are nuisance variables subject to inference along with the model parameters, $\boldsymbol{\theta}$. The joint likelihood function of Eq. (56) is not exact as we cannot compute the values of ε_t pertaining to the first $t = \{1, \dots, p\}$ residuals of $\mathbf{E}(\boldsymbol{\theta})$ as this requires knowledge of the unobserved residuals, $\{e_{1-p}(\boldsymbol{\theta}), \dots, e_0(\boldsymbol{\theta})\}$. A similar problem was observed with $e_0(\boldsymbol{\theta})$ in the derivation of the Gaussian log-likelihood function in Eq. (40) with first-order serial correlation of the residuals. The approximation of Eq. (56) assumes that the unobserved residuals, $\{e_{1-p}(\boldsymbol{\theta}), \dots, e_0(\boldsymbol{\theta})\}$, are all zero. If $n \gg p$, then Eq. (56) is a valid approximation of the true likelihood.

The formulation of Eq. (56) was derived by Schoups and Vrugt (2010) and coined the *generalized likelihood function* (GLF) as it extends applicability of standard likelihood functions to situations wherein residuals are non-Gaussian distributed with heavy tails, skew, heteroscedasticity, and serial correlation at multiple different lags. The presented order of computational steps in the GLF may not always return the exact likelihood of the residuals. Stability analysis in Evin et al. (2013) has demonstrated that variance stabilization of the residuals should precede treatment of serial correlation. Thus, the backshift operator in Eq. (52) should act on the normalized residuals, $e_t(\boldsymbol{\theta})/\hat{\sigma}_t$, rather than $e_t(\boldsymbol{\theta})$ with $t = \{1, \dots, n\}$ (see also Bates and Campbell (2001)).

The log-likelihood function of Eq. (56) appears in Eq. (8) of Schoups and Vrugt (2010). This derivation is straightforward and leads to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\boldsymbol{\sigma}}, \beta, \xi, \Phi_p) &\simeq n \log \left(\frac{2\sigma_\xi \omega_\beta}{(\xi + \xi^{-1})} \right) - \sum_{t=1}^n \{\log(\hat{\sigma}_t)\} \\ &- c_\beta \sum_{t=1}^n \left| \frac{\mu_\xi + \sigma_\xi \varepsilon_t(\boldsymbol{\theta}, \Phi_p)}{\xi \text{sgn}(\mu_\xi + \sigma_\xi \varepsilon_t(\boldsymbol{\theta}, \Phi_p))} \right|^{2/(1+\beta)} \end{aligned} \quad (57)$$

which can be further simplified to

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\boldsymbol{\sigma}}, \beta, \xi, \Phi_p) &\simeq n \log(2\sigma_\xi \omega_\beta) - n \log(\xi + \xi^{-1}) - \sum_{t=1}^n \{\log(\hat{\sigma}_t)\} \\ &\quad - c_\beta \sum_{t=1}^n \left| \frac{\mu_\xi + \sigma_\xi \varepsilon_t(\boldsymbol{\theta}, \Phi_p)}{\xi^{\text{sgn}}(\mu_\xi + \sigma_\xi \varepsilon_t(\boldsymbol{\theta}, \Phi_p))} \right|^{2/(1+\beta)} \end{aligned} \quad (58)$$

If the measurement error standard deviations, $\hat{\boldsymbol{\sigma}} = \{\hat{\sigma}_1, \dots, \hat{\sigma}_n\}$, of the n observations, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \dots, \tilde{y}_n\}$, are known a-priori, then the log-likelihood formulation of Eq. (58) suffices. Otherwise, we can derive estimates of $\hat{\boldsymbol{\sigma}} = \{\hat{\sigma}_1, \dots, \hat{\sigma}_n\}$ using the measurement model of Eq. (42). The values of σ_0 and σ_1 are then estimated along with the other nuisance variables of the generalized likelihood functions, thus $\boldsymbol{\alpha} = \{\sigma_0, \sigma_1, \beta, \xi, \Phi_p\}$. This concludes the (sub)section on likelihood functions.

4.4 The Posterior Distribution, $P(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$

Now we have defined the prior distribution, $P(\boldsymbol{\theta})$, and the likelihood function, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$, we are left with inference of the (unnormalized) posterior distribution, $P(\boldsymbol{\theta} | \tilde{\mathbf{Y}}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$, which summarizes our updated knowledge (belief) on the parameters, $\boldsymbol{\theta}$. Unfortunately, in most applications of Bayes' theorem, the posterior distribution does not have a closed-form (compact) analytical solution, and we have to resort to sampling methods to approximate the posterior distribution. The underlying principles of this approach are explained in Fig. 7 using a scatter plot of $M = 160$ samples drawn randomly from a bivariate normal distribution, $\mathcal{N}_2(\mathbf{a}, \Sigma)$ with mean $\mathbf{a} = \{a_1, a_2\}$ and 2×2 covariance matrix,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

As the off-diagonal entries of Σ are set to zero, the two parameters of the distribution, θ_1 and θ_2 , are independent, and consequently $P(\theta_1) \sim \mathcal{N}(a_1, \sigma_1^2)$ and $P(\theta_2) \sim \mathcal{N}(a_2, \sigma_2^2)$.

The plotted samples exhibit several key features. First, the scatter of points exhibits significant variations in sampling density. The sample density is largest in the center of the cloud and decreases slowly in all radial directions away from this midpoint. Second, a circular pattern emerges of points with equal sampling density. Third, the sampled points do not express a preferred orientation of θ_1 and θ_2 . The lack of linear dependence is exemplary for uncorrelated variables (parameters). Fourth, the peaks of the inferred frequency distributions of θ_1 (green line) and θ_2 (blue line) are within the bins of maximum sample density. This coincides with the nucleus (heart) of the point cloud. Finally, the frequency distributions of the two parameters are symmetric and

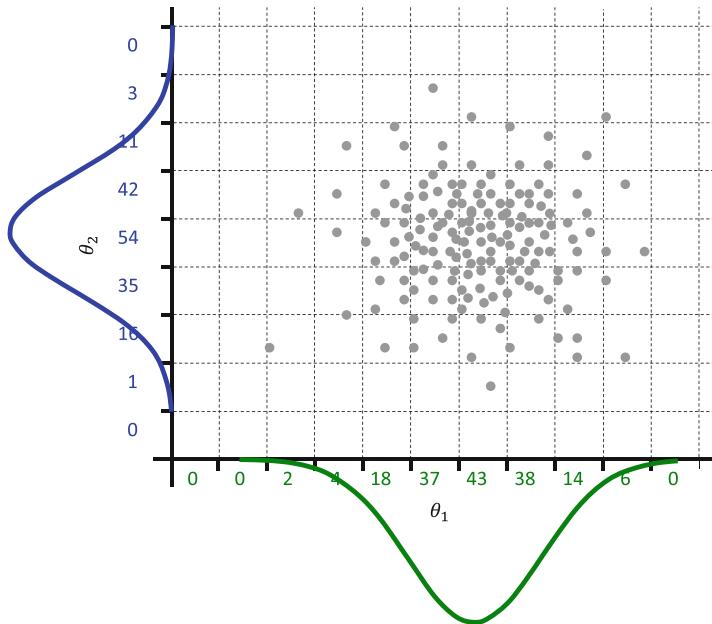


Fig. 7 Scatter plot of $M = 160$ samples (gray dots) drawn at random from a bivariate normal distribution, $\mathcal{N}_2(\mathbf{a}, \Sigma)$. The dashed lines delineate the different bins that are used to construct the frequency distribution (histogram) of θ_1 and θ_2 . The green and blue lines depict the marginal distributions of the parameters and are inferred from the sampled points

well described with a Gaussian distribution. The key notion of these findings is that (a) the sampling density is a proxy for the probability density of the target distribution, (b) the orientation of the samples is a measure of the correlation among the variables of the target distribution, and (c) the dispersion of the samples is a measure of the variances of the individual variables of the target distribution.

This simple example with the point cloud shows that we can represent any probability distribution with a large number of samples as long as the sampled points satisfy one crucial requirement and that is that they are distributed exactly according to the underlying density function of the probability distribution. In other words, the sample density at any point of the target distribution must match exactly the density of the distribution at that point. Then, the inferred marginal distributions of the parameters will match their counterparts of the target distribution of interest and the covariance (correlation) structure of the parameters is honored perfectly. Indeed, the inferred distributions of θ_1 and θ_2 in Fig. 7 match exactly their “true” counterparts, $P(\theta_1)$ and $P(\theta_2)$. The number of samples that is required to represent properly a multivariate probability distribution depends on the shape and dimensionality of this distribution, more of which later.

These results now beg the question of how we should generate the samples? Drawing them at random from some distribution will not suffice – unless this distribution is exactly equal to the target distribution of interest. In the past decades,

many different methods have been developed for sampling accurately an unknown distribution. All these methods rely in some way on Monte Carlo simulation. In the next sections, we discuss the application of these methods to approximate the d -variate posterior distribution, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$. This distribution, also referred to as the target or limiting distribution, is often high dimensional. The different methods assume a continuous parameter space $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^d$, yet with simple modifications can be used to approximate discrete target distributions.

5 Monte Carlo Approximation

Monte Carlo methods are a broad class of computational algorithms that use repeated random sampling to approximate some arbitrary d -variate distribution, $F(\mathbf{x})$ with probability density function, $f(\mathbf{x})$. This unknown multivariate distribution, $F(\mathbf{x})$, is equivalent to the posterior distribution, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$. For the time being, we use the symbol $\mathbf{x} = \{x_1, \dots, x_d\}$ to denote the d -variables of the unknown distribution, $F(\mathbf{x})$, which, in our application, constitute the model parameters, $\boldsymbol{\theta}$, possibly augmented with nuisance variables, thus $\mathbf{x} = \{\boldsymbol{\theta}, \boldsymbol{\alpha}\}$.

The basic idea of Monte Carlo methods is to use an alternative distribution, $Q(\mathbf{x})$, which is easy to sample from in practice and whose probability density function, $q(\mathbf{x})$, approximates as closely and consistently as possible the unknown density, $f(\mathbf{x})$. The “known” distribution $Q(\mathbf{x})$ is also referred to as the proposal distribution and serves as catalyst to approximate the posterior distribution.

5.1 Rejection Sampling

The earliest Monte Carlo method is the acceptance-rejection algorithm (also referred to as rejection sampling) and produces M different samples of the desired target distribution, $F(\mathbf{x})$. The various steps of this method are summarized in the numerical recipe of Algorithm 1, wherein the label Z signifies a draw from the standard uniform distribution, $Z \sim \mathcal{U}(0,1)$.

Algorithm 1 Rejection Sampling

- 1: Define a proposal distribution, $Q(\mathbf{x})$, so that $q(\mathbf{x}) \geq f(\mathbf{x})$ if $f(\mathbf{x}) > 0$.
- 2: Define M and set counter, $i = 1$.
- 3: **while** $i < M$ **do**
- 4: Sample randomly a candidate point, \mathbf{x}_p , from $Q(\mathbf{x})$, $\mathbf{x}_p \sim Q(\mathbf{x})$ and calculate $f(\mathbf{x}_p)$.
- 5: Compute the acceptance probability, $P_{\text{acc}}(\mathbf{x}_p) = f(\mathbf{x}_p)/q(\mathbf{x}_p)$, of \mathbf{x}_p .
- 6: If $Z \leq P_{\text{acc}}(\mathbf{x}_p)$ then set $\mathbf{x}_{(i)} = \mathbf{x}_p$ and counter, $i = i + 1$, otherwise reject \mathbf{x}_p .
- 7: **end while**

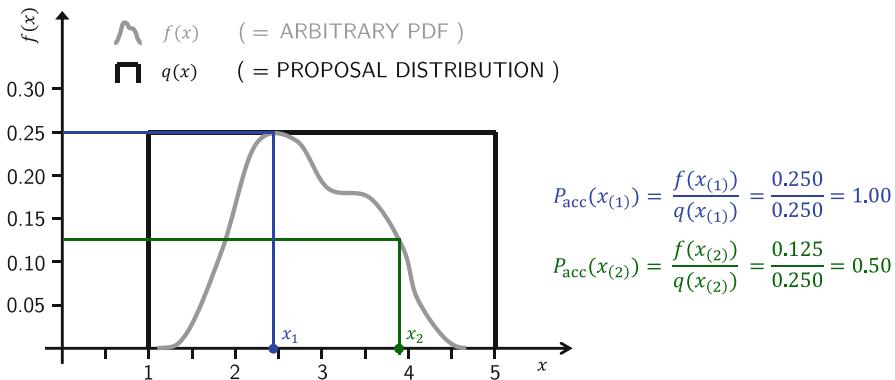


Fig. 8 Application of rejection sampling to some univariate target distribution, $F(x)$, with arbitrary density function, $f(x)$ (in gray) using as proposal density (in black) the uniform distribution, $Q(x) \sim \mathcal{U}(1, 5)$

After a sufficient number of iterations, rejection sampling produces M different samples that will be distributed exactly according to the unknown distribution, $F(\mathbf{x})$ with underlying density, $f(\mathbf{x})$. Why the accepted samples converge exactly to $F(\mathbf{x})$ is relatively easy to proof mathematically using the envelope principle. Here, instead we provide a visual explanation (see Fig. 8) and draw samples from a uniform distribution (black line) to approximate the univariate target distribution, $F(x)$, with arbitrary density, $f(x)$ (gray line).

It is not difficult to see that the acceptance probability, $P_{\text{acc}}(x_p) = f(x_p)/q(x_p)$, of each candidate point, $x_p \in [1, 5]$, is directly proportional to the density, $f(x_p)$, of the target distribution. Thus if we accept each candidate point with probability, $P_{\text{acc}}(x_p)$, we account for “bias” in the samples drawn from the proposal distribution and distribute the accepted samples exactly according to the target density, $f(x)$ of $F(x)$. This does require that the entire target distribution is sampled, that is the space of \mathbf{x} for which $f(\mathbf{x}) > 0$, and that $q(\mathbf{x}) \geq f(\mathbf{x})$, otherwise the acceptance probability can reach values larger than unity, thereby chopping off the peaks of the target distribution. In practice, a multiplier $c \in [1, \infty)$ is used to inflate the density of the proposal distribution so that it always exceeds the density of the target distribution, $cq(\mathbf{x}) \geq f(\mathbf{x})$. This multiplier enters into the denominator of the acceptance probability, and should thus be chosen wisely, otherwise (e.g., too large) the rejection algorithm can become highly inefficient as most candidate points will be dismissed. The value of c can be determined as largest value of the ratios, $f(x_p)/q(x_p)$, of the many different \mathbf{x}_p ’s. Then, the new acceptance probability, $P_{\text{acc}} = f(\mathbf{x}_p)/cq(\mathbf{x}_p)$ of each sample \mathbf{x}_p can be recomputed and the set of accepted samples reconstructed. In practice, the “best” proposal density, $q(\mathbf{x})$, minimizes the value of the constant c , that is $c = \sup_{\mathbf{x}}(f(\mathbf{x})/q(\mathbf{x}))$. A perfect agreement between the proposal and target distribution equates to a value of $c = 1$.

The efficiency of rejection sampling depends in large part on the choice of the proposal distribution that is used to generate trial points. This distribution must

satisfy two conditions, that is, (a) its support is equal to, or larger than, the target distribution and (b) its density is always equal to, or larger than, the target density. These two conditions are very difficult to satisfy in practice, without detailed knowledge of the target distribution (as in our Bayes' application). This is especially true for high-dimensional targets with complicated multivariate relationships among the variables. A poorly construed proposal distribution has profound consequences as a (very) large portion of the candidate points will be rejected and go to waste. For high-dimensional target distributions, say $d = 50$, this can lead to acceptance rates on the order of say 0.1%. Consequently, it will take a very large number of iterations to generate a sufficient sample from the target distribution.

5.2 Importance Sampling

Importance sampling is an important improvement over the acceptance-rejection algorithm. This algorithm can be written in a few lines (see Algorithm 2) and is widely used to compute raw, central, and standardized moments of the target distribution, $F(\mathbf{x})$.

Algorithm 2 Importance Sampling

- 1: Define an importance distribution, $G(\mathbf{x})$, so that $g(\mathbf{x}) > 0$ if $f(\mathbf{x}) > 0$.
- 2: Define M .
- 3: **for** $i = 1, \dots, M$ **do**
- 4: Sample randomly a point, $\mathbf{x}_{(i)}$, from $G(\mathbf{x})$, $\mathbf{x}_{(i)} \sim G(\mathbf{x})$ and calculate $f(\mathbf{x}_{(i)})$.
- 5: Compute the importance weight, $w(\mathbf{x}_{(i)}) = f(\mathbf{x}_{(i)})/q(\mathbf{x}_{(i)})$, of $\mathbf{x}_{(i)}$.
- 6: **end for**

Whereas rejection sampling produces samples with equal weight that are distributed exactly according to the target distribution, importance sampling returns a collection of weighted samples. This weighted sample cannot be used to draw marginal and or joint histograms of the target distribution but rather serves to compute moments of $F(\mathbf{x})$ such as its mean, variance, skewness, and kurtosis. To generate a sample from $F(\mathbf{x})$, we must draw *with replacement* from the importance samples, $\{\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(M)}\}$, using selection probabilities proportional to the importance weights, $\{w(\mathbf{x}_{(1)}), \dots, w(\mathbf{x}_{(M)})\}$. This process is also known as sampling importance resampling (SIR) and produces a collection of samples with equal weights distributed according to the target distribution, $F(\mathbf{x})$. Thus, points with relatively large importance weights have a much higher chance to be selected (with replacement) in this collection of equal weight points than samples with a negligible importance weight. The frequency of appearance of each importance sample in this resampled collection of points is thus proportional to the underlying density function, $f(\mathbf{x})$, of the target distribution, $F(\mathbf{x})$. If the importance sample is sufficiently large, then resampling should provide a reasonable approximation of $F(\mathbf{x})$.

Importance sampling has two main advantages over rejection sampling. First, it does not produce waste as all samples are used to approximate the moments of the

target distribution. Second, the density, $g(\mathbf{x})$, of the (importance) sampling distribution, $G(\mathbf{x})$, does not have to be equal to or larger than the target density. The only requirement is that $g(\mathbf{x}) > 0$ if $f(\mathbf{x}) > 0$. This simplifies considerable practical application. Nevertheless, the construction of a proper importance distribution is difficult without detailed knowledge of the target distribution (as in our Bayes' application) and becomes particularly cumbersome in high-dimensional parameter spaces. When the importance distribution is too wide, a large majority of the sampled points will receive negligible weights. On the other hand, when the importance distribution is too narrow, the sampled points will not characterize adequately the target distribution. Indeed, methods such as rejection sampling and importance sampling are rather frugal and inefficient for all but very low dimensional problems. Next, we therefore resort to an alternative class of methods to explore the target distribution.

5.3 Markov Chain Monte Carlo Simulation

The basis of MCMC simulation is a Markov chain that generates a random walk through the search space and successively visits solutions with stable frequencies stemming from a stationary distribution, $F(\mathbf{x})$. To explore the target distribution, $F(\mathbf{x})$, a MCMC algorithm generates trial moves from the current state of the Markov chain $\mathbf{x}_{(i-1)}$ to a candidate point, \mathbf{x}_p . The earliest MCMC approach is the random walk Metropolis (RWM) algorithm introduced by Metropolis et al. (1953). This scheme is constructed to maintain detailed balance with respect to $f(\mathbf{x})$ at each step in the chain. If $f(\mathbf{a})$ denotes the probability to find the system in state \mathbf{a} and $q(\mathbf{a} \rightarrow \mathbf{b})$ is the conditional probability, $q(\mathbf{b}|\mathbf{a})$, to perform a trial move from \mathbf{a} to \mathbf{b} , then the probability $P_{\text{acc}}(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p)$ to accept the trial move from $\mathbf{x}_{(i-1)}$ to \mathbf{x}_p is related to $P_{\text{acc}}(\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)})$ according to

$$\begin{aligned} & f(\mathbf{x}_{(i-1)}) q(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p) P_{\text{acc}}(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p) \\ &= f(\mathbf{x}_p) q(\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)}) P_{\text{acc}}(\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)}) \end{aligned} \quad (59)$$

This principle of *detailed balance* originates from the work of Boltzmann and Maxwell on collision and gas kinetics, respectively. This condition implies that there is no net inflow or outflow of probability among some closed set of possible states, and consequently, there exists a unique equilibrium distribution of the states. Detailed balance is of particular relevance to a Markov chain as it *guarantees* that the chain, under some regularity conditions, will converge to the exact equilibrium (= target) distribution. Indeed, a chain that maintains detailed balance will visit each state, \mathbf{a} , of the stationary (equilibrium) distribution with frequency proportional to its underlying probability density, $f(\mathbf{a})$. This does require the chain to be *irreducible* (it is possible to transition, in one or more steps, from any state to another configuration) to be *a-periodic* (return to a state occurs at irregular times) and to be *positive recurrent* (positive probability to return to a state) (Robert and Casella 2004). For

a properly constructed proposal distribution, these three conditions are usually satisfied in practice, except for trivial exceptions. Note that detailed balance is not a necessary condition for convergence of the Markov chain to the target distribution. A few examples will be given later.

If a symmetric jumping distribution is used, that is $q(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p) = q(\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)})$, then it follows from Eq. (59) that

$$\frac{P_{\text{acc}}(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p)}{P_{\text{acc}}(\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)})} = \frac{f(\mathbf{x}_p)}{f(\mathbf{x}_{(i-1)})} \quad (60)$$

This equation does not yet fix the acceptance probability. Metropolis et al. (1953) made the following choice

$$P_{\text{acc}}(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p) = \min \left[1, \frac{f(\mathbf{x}_p)}{f(\mathbf{x}_{(i-1)})} \right], \quad (61)$$

to determine whether to accept a trial move or not. This selection rule of candidate points has become the basic building block of MCMC algorithms. Hastings (1970) has generalized Eq. (61) to cases with nonsymmetrical proposal distributions

$$P_{\text{acc}}(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p) = \min \left[1, \frac{f(\mathbf{x}_p)q(\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)})}{f(\mathbf{x}_{(i-1)})q(\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p)} \right], \quad (62)$$

wherein the forward jump, $\mathbf{x}_{(i-1)} \rightarrow \mathbf{x}_p$, and backward jump, $\mathbf{x}_p \rightarrow \mathbf{x}_{(i-1)}$, do not have equal probability, thus $q(\mathbf{x}_p|\mathbf{x}_{(i-1)}) \neq q(\mathbf{x}_{(i-1)}|\mathbf{x}_p)$. This generalization is known as the Metropolis-Hastings (MH) algorithm and broadens significantly the type of proposal distribution that can be used for posterior inference.

Figure 9 depicts the evolution (trajectory) of a single Markov chain starting from an arbitrary initial state (black square) for some hypothetical $d = 2$ -dimensional target distribution. The gray arrows denote the different trial moves (jumps) of the chain, most of which are accepted (green dots), and some of which are declined (red dots). If a proposal is accepted then the chain moves to this new position, otherwise the chain remains at its “old” state and this position (e.g., values of \mathbf{x}) is replicated in the Markov chain. After about 13 steps, the chain has reached the stationary distribution (in orange). The subsequent positions of the chain are used to approximate the target distribution, $F(\mathbf{x})$.

The samples which are stored in the chain have equal weights and share in common with SIR that their frequency of appearance is directly proportional to the underlying density, $f(\mathbf{x})$, of the target distribution. One crucial difference with rejection and importance sampling is that the sampling (proposal) distribution, $q(\cdot)$, in MCMC simulation does not have to cover or envelope the target distribution. This proposal distribution simply travels with the state of the chain and centers on the last position to create candidate points.

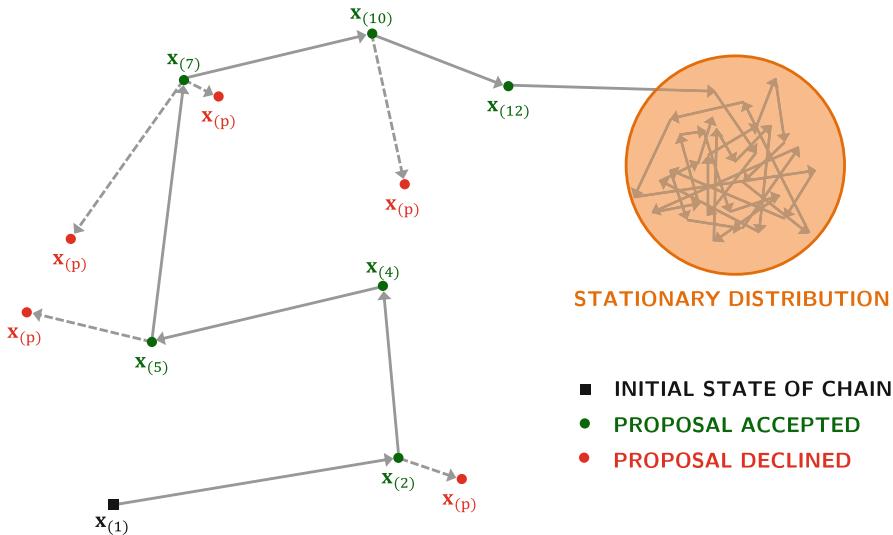


Fig. 9 Schematic illustration of a chain trajectory derived from MCMC simulation for a two-dimensional target distribution, $f(\mathbf{x})$. The black square signifies the initial state of the chain, and the gray arrows denote the different jumps. Color coding is used to differentiate between proposals (candidate points or trial moves) that have been accepted (green dots) or rejected (red dots), respectively. The orange circle signifies the area of the stationary distribution, which envelopes the target distribution. The number of times each chain position appears in the Markov chain (replicates after rejection of proposals) is directly proportional to the density of the target distribution

The core of the RWM algorithm can be written in just a few lines (see Algorithm 3) and requires a symmetric jumping distribution, $q(\cdot)$, a uniform random number generator, $Z \sim \mathcal{U}(0,1)$, and the target density, $f(\mathbf{x})$, to create a chain trajectory.

Algorithm 3 Random Walk Metropolis

- 1: Define the proposal distribution, $q(\cdot)$ (must satisfy $q(a|b) = q(b|a)$).
- 2: Define M .
- 3: Draw randomly $\mathbf{x}_{(1)}$, the initial state of the chain, and calculate $f(\mathbf{x}_{(1)})$.
- 4: **for** $i = 2, \dots, M$ **do**
- 5: Sample randomly a candidate point, \mathbf{x}_p , from the symmetric proposal distribution, $q(\mathbf{x}|\mathbf{x}_{(i-1)})$, centered on $\mathbf{x}_{(i-1)}$, thus $\mathbf{x}_p \sim q(\mathbf{x}|\mathbf{x}_{(i-1)})$.
- 6: Compute the target density, $f(\mathbf{x}_p)$, at \mathbf{x}_p .
- 7: Calculate the Metropolis ratio, $P_{\text{acc}}(\mathbf{x}_p) = \min(1, f(\mathbf{x}_p)/f(\mathbf{x}_{(i-1)}))$.
- 8: **if** $Z \leq P_{\text{acc}}(\mathbf{x}_p)$ **then**
- 9: Set $\mathbf{x}_{(i)} = \mathbf{x}_p$ and $f(\mathbf{x}_{(i)}) = f(\mathbf{x}_p)$,
- 10: **else**
- 11: Remain at “old” state, that is $\mathbf{x}_{(i)} = \mathbf{x}_{(i-1)}$ and $f(\mathbf{x}_{(i)}) = f(\mathbf{x}_{(i-1)})$.
- 12: **end if**
- 13: **end for**

In words, assume that the points $\{\mathbf{x}_0, \dots, \mathbf{x}_{(i-1)}\}$ have already been sampled, then the RWM algorithm proceeds as follows. First, a candidate point \mathbf{x}_p is sampled from a proposal (jumping) distribution, $q(\cdot)$, that depends on the present location, $\mathbf{x}_{(i-1)}$, and is symmetric, $q(\mathbf{x}_p | \mathbf{x}_{(i-1)}) = q(\mathbf{x}_{(i-1)} | \mathbf{x}_p)$. Next, the candidate point is either accepted or rejected using the Metropolis acceptance probability (Eq. 61). Finally, if the proposal is accepted, the chain moves to \mathbf{x}_p , otherwise the chain remains at its current location $\mathbf{x}_{(i-1)}$. Repeated application of these three steps results in a Markov chain which, under certain regularity conditions, has a unique stationary distribution with posterior probability density function, $f(\mathbf{x})$. In practice, this means that if one looks at the archived values of \mathbf{x} in the chain sufficiently far from the arbitrary initial state, thus after a burn-in period, then these successively generated states will be distributed according to $F(\mathbf{x})$, the unknown target distribution of \mathbf{x} . Burn-in is required to allow the chain to explore the search space and reach its stationary regime (see Fig. 9).

The RWM algorithm is relatively simple to implement, yet its efficiency is determined in large part by the choice of the proposal distribution, $q(\cdot)$, used to create trial moves (transitions) in the Markov chain. When the proposal distribution is too wide, too many candidate points are rejected, and therefore the chain will not mix efficiently and converge only slowly to the target distribution. On the other hand, when the proposal distribution is too narrow, most candidate points will be accepted, but the covered distance is so small that it will take a prohibitively large number of iterations before the chain has converged to the target distribution. The choice of proposal distribution is therefore of crucial importance and determines the computational cost and practical feasibility of MCMC simulation.

Note, that we assume a fixed computational budget of M iterations of the RWM algorithm. In practice, the chain will continue to evolve until it reaches a stationary distribution as judged by one or more convergence diagnostics (discussed later). The budget of iterations is thus determined on the fly based on the convergence properties of the sampled chain.

5.3.1 Automatic Tuning of Proposal Distribution

In the past decade, a variety of different approaches have been proposed to increase the efficiency of MCMC simulation and enhance the original RWM and MH algorithms. These approaches can be grouped into single and multiple chain methods.

Single-Chain Methods

The most common adaptive single chain methods are the adaptive proposal (AP) (Haario et al. 1999), adaptive Metropolis (AM) (Haario et al. 2001), and delayed rejection adaptive Metropolis (DRAM) algorithm (Haario et al. 2006), respectively. These methods simulate a single trajectory by drawing candidate points from a Gaussian proposal distribution, $\mathbf{x}_p \sim \mathcal{N}_d(\mathbf{x}_{(i-1)}, s_d \sum)$, with covariance matrix, Σ , which is updated periodically after every m iterations ($m \geq 1$) using all past samples stored in the chain, $\Sigma = \text{Cov}[\{\mathbf{x}_{(1)}, \dots, \mathbf{x}_{(i-1)}\}] + \varphi \mathbf{I}_d$. The variable s_d

signifies the so-called jump rate and depends on the dimensionality of the target distribution, \mathbf{I}_d , denotes the $d \times d$ identity matrix, and $\varphi = 10^{-6}$ is a small scalar that prevents the collapse of the sample covariance matrix to singularity (jumps become zero). This term also guarantees, at least in theory, that the sampled chain is irreducible because of the unbounded support of the normal jumping distribution with nonsingular (invertible) covariance matrix, Σ , that is $|\Sigma| > 0$. As a basic choice, the scaling factor is chosen to be $s_d = 2.38^2/d$ which has proven optimal for Gaussian target and proposal distributions (Gelman et al. 1996; Roberts et al. 1997) and should give an acceptance rate close to 0.44 for $d = 1$, 0.28 for $d = 5$, and 0.23 for large d .

A summary of the AM method appears below in Algorithm 4, wherein the notion $\text{mod}(a, b)$ signifies the modulo operator. This operator returns zero if the quotient k of $a > 0$ and $b > 0$, or $k = a/b$, equates to an integer, or $\text{mod}(a, b) = 0$ if $k \in \mathbb{N}_+$.

Algorithm 4 Adaptive Metropolis

- 1: Calculate $s_d = 2.38^2/d$, define $d \times d$ covariance matrix, Σ , $\varphi = 10^{-6}$, $m \geq 1$, and M .
- 2: Draw randomly $\mathbf{x}_{(1)}$, the initial state of the chain, and calculate $f(\mathbf{x}_{(1)})$.
- 3: **for** $i = 2, \dots, M$ **do**
- 4: Sample randomly a candidate point, $\mathbf{x}_p \sim N_d(\mathbf{x}_{(i-1)}, s_d\Sigma)$.
- 5: Compute the target density, $f(\mathbf{x}_p)$, at \mathbf{x}_p .
- 6: Calculate the Metropolis ratio, $P_{\text{acc}}(\mathbf{x}_p) = \min(1, f(\mathbf{x}_p)/f(\mathbf{x}_{(i-1)}))$.
- 7: **if** $Z \leq P_{\text{acc}}(\mathbf{x}_p)$ **then**
- 8: Set $\mathbf{x}_{(i)} = \mathbf{x}_p$ and $f(\mathbf{x}_{(i)}) = f(\mathbf{x}_p)$.
- 9: **Else**
- 10: Remain at “old” state, $\mathbf{x}_{(i)} = \mathbf{x}_{(i-1)}$ and $f(\mathbf{x}_{(i)}) = f(\mathbf{x}_{(i-1)})$.
- 11: **end if**
- 12: **if** $\text{mod}(i, m) = 0$ **then**
- 13: Adapt covariance matrix, $\Sigma = \text{Cov}[\{\mathbf{x}_{(1)}, \dots, \mathbf{x}_i\}] + \varphi \mathbf{I}_d$.
- 14: **end if**
- 15: **end for**

Thus, the AM algorithm is a special implementation of the RWM algorithm with a transient (multi)normal transition kernel as proposal distribution of the Markov chain. The covariance matrix of the Gaussian proposal distribution is adapted every m iterations using the archived chain samples. This adaptation enhances, sometimes dramatically, the convergence speed of the chain to the stationary distribution, as the jumps will align slowly with the orientation and scale of the target distribution. An important drawback of adaptation is, however, that the AM algorithm does not satisfy the detailed balance condition of Eq. (59) at every step in the chain. This is easy to see if we compare the (multi)normal proposal distribution of the AM algorithm immediately before and after an update to Σ at iterations $j = km$, where $k \in \mathbb{N}_+$. Then the (conditional) probability, $q(\mathbf{x}_{(j)} | \mathbf{x}_{(j-1)})$, of the forward jump, $\mathbf{x}_{(j-1)} \rightarrow \mathbf{x}_{(j)}$ (with “old” Σ), does not equate to the conditional probability, $q(\mathbf{x}_{(j-1)} | \mathbf{x}_{(j)})$, of the backward jump, $\mathbf{x}_{(j)} \rightarrow \mathbf{x}_{(j-1)}$ (with “new” Σ). Detailed balance is rapidly restored during the subsequent $m - 1$ iterations of the chain as the candidate points are created with a fixed

Σ . Nonetheless, the resulting chain simulated by the AM algorithm is not truly Markovian.

But why then does the AM algorithm converge to the appropriate limiting distribution? This is because of diminishing adaptation (Roberts and Rosenthal 2007). In words, the transition kernel (multivariate normal distribution) of the AM algorithm converges to a fixed proposal distribution with increasing length of the chain. Indeed, the distance between successive values of Σ decreases to zero as the number of samples in the chain grows without bound. Note that the chain may converge to another than the target distribution if only the recent past is used to generate trial moves (see Haario et al. (2001) for an example). Another viable adaptation strategy is to fix the covariance matrix (say identity matrix) and to tune instead the scaling factor, s_d , during a burn-in period, until a desired acceptance rate is obtained (23% for large d). If adaptation is limited to the burn-in period, then the chain transitions in the equilibrium distribution are fully Markovian. To speedup the convergence rate of the AM algorithm on high-dimensional problems (large d), Haario et al. (2005) introduced single-site updating in which one parameter is sampled at a time.

The use of a multivariate normal transition kernel (with/without adaptation) may work well for Gaussian-like target distributions but may not be adequate to characterize multimodal distributions with long tails, and possibly infinite first and second moments. Experience further suggests that single chain methods have a hard time exploring efficiently multidimensional parameter spaces, particularly when confronted with different regions of attraction and (numerous) local optima. The use of an overly dispersed proposal distribution will help to traverse difficult search spaces and/or sample disconnected modes, yet the resulting chain will converge only slowly due to an improper scaling of the jumps (excessively large). It is also particularly difficult to judge convergence of a single chain trajectory in absence of an independent benchmark against which we can compare the statistics of the sampled values. Even the most powerful diagnostics do not protect us against a chain that has converged to a local basin of attraction in the parameter space or a chain that explores only one mode of the target distribution. Indeed, single-chain MCMC methods (e.g., DRAM, RWM, AM, and AP) suffer many similar problems as local optimization methods (e.g., steepest descent, Newton method, Levenberg-Marquardt) and cannot guarantee an exhaustive exploration of the parameter space in pursuit of the target distribution.

Multiple Chain Methods: DE-MC

Multiple chain MCMC methods simulate different trajectories in parallel to explore the posterior target distribution. This approach has several important advantages, particularly for search spaces with different regions of attraction and numerous local optima, and skewed, tailed, and multimodal target distributions with complex multivariate dependencies among the variables (Gilks et al. 1994; Liu et al. 2000; ter Braak 2006; ter Braak and Vrugt 2008; Vrugt et al. 2009; Radu et al. 2009). The use of multiple different chains offers a robust protection against premature convergence and opens up an array of powerful statistical tests to assess convergence of the

sampled values to an equilibrium distribution (Gelman and Rubin 1992). One example of an efficient multichain method is the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm (Vrugt et al. 2003). This method has found application in environmental modeling and builds on the shuffling concept of the SCE-UA optimization method of Duan et al. (1992) to distribute, among the parallel chains, information gained about the search space. The use of this shuffling approach introduces a transient transition kernel which expedites convergence of SCEM-UA to the equilibrium distribution but at the expense of exact reversibility of the sampled chain trajectories. Another novelty of the SCEM-UA algorithm is its explicit treatment of outlier chains, a necessity to traverse efficiently complex search spaces in lieu of the target distribution. If adaptation of the (multi)normal proposal distribution is restricted to the burn-in period only, then the chain transitions simulated by the SCEM-UA algorithm satisfy reversibility. The method then derives an efficient Gaussian proposal distribution for the standard Metropolis algorithm.

ter Braak (2006) proposed a simple adaptive RWM algorithm called Differential Evolution Markov chain (DE-MC). DE-MC uses differential evolution as genetic algorithm for population evolution with a Metropolis selection rule to decide whether candidate points should replace their parents or not. In DE-MC, N different Markov chains are run simultaneously in parallel. If the state of a single chain is given by the d -vector \mathbf{x} , then at each generation $i - 1$, the N chains in DE-MC define a $N \times d$ matrix (population), $\mathbf{X}_{(i-1)} = \{\mathbf{x}_{(i-1)}^1, \dots, \mathbf{x}_{(i-1)}^N\}$, with each chain as a row. Then multivariate proposals, \mathbf{x}_p^j in each chain, $j = \{1, \dots, N\}$, are generated on the fly from the collection of chains, $\mathbf{X}_{(i-1)}$, using differential evolution (Storn and Price 1997; Price et al. 2005)

$$\mathbf{x}_p^j = \gamma_d \left(\mathbf{x}_{(i-1)}^a - \mathbf{x}_{(i-1)}^b \right) + \boldsymbol{\zeta}_d, \quad a \neq b \neq j, \quad (63)$$

where γ_d signifies the jump rate (dimensionality dependent), a and b are integers drawn without replacement from the natural numbers $\{1, \dots, j - 1, j + 1, \dots, N\}$, and $\boldsymbol{\zeta} \stackrel{\mathcal{D}}{\sim} \mathcal{N}_d(0, c_*)$ is drawn from a normal distribution with small standard deviation, say $c_* = 10^{-6}$. By accepting each proposal with Metropolis probability

$$P_{\text{acc}}(\mathbf{x}_p^j) = \min \left[1, f(\mathbf{x}_p^j) / f(\mathbf{x}_{(i-1)}^j) \right], \quad (64)$$

a Markov chain is obtained, the stationary or limiting distribution of which is the posterior distribution. A mathematical proof of convergence appears in ter Braak and Vrugt (2008); a graphical explanation of reversibility can be found in Vrugt and Ter Braak (2011) for discrete sampling problems.

Because the joint probability density function of the N chains factorizes to $f(\mathbf{x}^1 | \cdot) \times \dots \times f(\mathbf{x}^N | \cdot)$, the states, $\mathbf{x}_{(k)}^1 \dots \mathbf{x}_{(k)}^N$, of the individual chains are independent at any iteration (generation) k after DE-MC has become independent of its initial value. If the initial population is drawn from the prior distribution, then DE-MC translates this sample into a posterior population. From the guidelines of

s_d in RWM, the optimal choice of $\gamma_d = 2.38/\sqrt{2d}$. With a 10% probability, the value of γ is set to unity to enable the DE-MC chains to traverse rapidly large search spaces and jump directly between different (disconnected) modes of the equilibrium distribution (ter Braak 2006; ter Braak and Vrugt 2008; Vrugt et al. 2008, 2009). Mode-jumping is a desirable property of the DE-MC algorithm as evidenced by the performance of this method on multimodal target distributions.

Algorithm 5 summarizes the DE-MC algorithm in different algorithmic steps, wherein the auxiliary label Z is drawn for each chain by sampling from the standard uniform distribution, $Z \sim \mathcal{U}(0,1)$

Algorithm 5 Differential Evolution Markov Chain

```

1: Define number of chains  $N \geq 2d$ , and  $c_* = 10^{-6}$ .
2: Set iteration,  $i = 2$ .
3: for  $j = 1, \dots, N$  do
4:   Draw randomly  $\mathbf{x}_{(1)}^j$ , the initial state of the  $j$ th chain, and calculate  $f(\mathbf{x}_{(1)}^j)$ .
5: end for
6: while chains not converged do
7:   for  $j = 1, \dots, N$  do
8:     Draw without replacement integers  $a$  and  $b$  from  $\{1, \dots, j-1, j+1, \dots, N\}$ 
9:     Select the jump rate; if  $\mathcal{U}(0,1) \leq 0.9$  then  $\gamma_d = 2.38/\sqrt{2d}$  otherwise  $\gamma_d = 1$ .
10:    Create a candidate point,  $\mathbf{x}_p^j$ , in the  $j$ th chain using Eq. (63)
11:    Compute the target density,  $f(\mathbf{x}_p^j)$ , at  $\mathbf{x}_p^j$ .
12:    Calculate the Metropolis ratio,  $P_{\text{acc}}(\mathbf{x}_p^j) = \min[1, f(\mathbf{x}_p^j)/f(\mathbf{x}_{(i-1)}^j)]$ 
13:    if  $Z \leq P_{\text{acc}}(\mathbf{x}_p^j)$  then
14:      Set  $\mathbf{x}_{(i)}^j = \mathbf{x}_p^j$  and  $f(\mathbf{x}_{(i)}^j) = f(\mathbf{x}_p^j)$ .
15:    else
16:      Remain at “old” state,  $\mathbf{x}_{(i)}^j = \mathbf{x}_{(i-1)}^j$  and  $f(\mathbf{x}_{(i)}^j) = f(\mathbf{x}_{(i-1)}^j)$ .
17:    end if
18:  end for
19:  Compute convergence diagnostics
20:  Update iteration,  $i = i + 1$ .
21: end while

```

The DE-MC method remedies an important practical problem of the RWM algorithm, namely that of choosing an appropriate scale and orientation for the proposal distribution. Earlier approaches such as (parallel) adaptive direction sampling (Gilks et al. 1994; Roberts and Gilks 1994; Gilks and Roberts 1996) solved the orientation problem but not the scale problem. The DE-MC method suffers one critical deficiency, however, and that its performance is impaired if one or more of the sampled chains are trapped in an unproductive area of the parameter space in

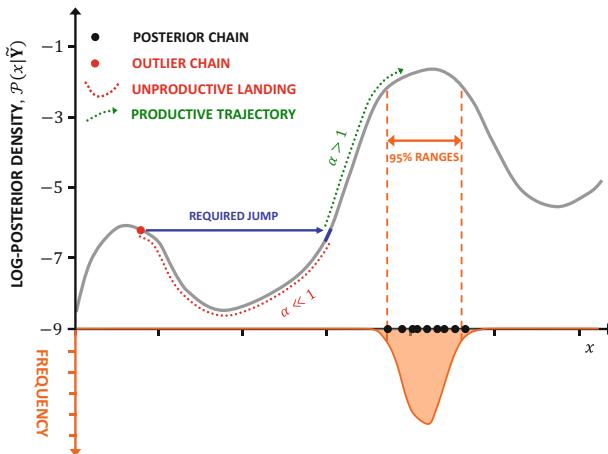


Fig. 10 Schematic illustration of a dissident chain (red) that is mired in a local basin of attraction to the search space of some univariate probability distribution. This outlier chain cannot transition to the target distribution demarcated in orange. The reasons are as follows. First, the aberrant chain cannot travel in multiple iterations the valley as almost all trial moves will exhibit a negligible acceptance rate. Second, a direct move to a point with equal probability density at the other side of the valley is implausible as the variation among the target chains is insufficient to warrant a large enough jump size with Eq. (63). Third, a direct move to the target with unit jump rate is impossible as the outlier chain cannot sample its own position ($a \neq b \neq i$) – a requirement for each chain trajectory to satisfy detailed balance. As a consequence, the dissident chain will be trapped forever

pursuit of the target distribution. This problem is well understood and explained in Fig. 10 using some arbitrary univariate probability density function.

Thus, chains that populate local optima can continue to persist forever if the jumps are insufficient to move the chain outside the space spanned by this optima (see also Fig. 2 of ter Braak and Vrugt (2008)). The state of this outlier chain not only contaminates the jumping distribution of Eq. (63) thereby slowing down unnecessarily the evolution and mixing of the “good” chains but also impairs convergence to a limiting distribution. By sampling a disjoint part of the parameter space, the N chains will not reach consensus on the limiting distribution as the mean and/or standard deviation of the parameter values sampled by the outlier chain will differ substantially from their counterparts simulated by the other $N - 1$ chains. This disagreement will continue to persist, perhaps even if we sample indefinitely. Consequently, the \hat{R} -statistic of Gelman and Rubin (1992) cannot reach its stipulated threshold of 1.2 to officially declare convergence.

The chance of a dissident chain increases rapidly with dimensionality of the target distribution (larger number of chains, $N \geq 2d$) and complexity of the underlying density function. A patch is therefore of crucial importance to remedy the searching behavior of the DE-MC algorithm and to expedite convergence on nonsmooth density functions with different regions of attraction and numerous local optima. Such a patch has important implications, however, as any efforts to remedy outlier

chains will violate detailed balance. The treatment of dissident chains should therefore be restricted to a burn-in period (see the SCEM-UA algorithm).

The next section concludes the section on MCMC simulation and presents the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm (Vrugt et al. 2008, 2009). This multichain algorithm does not suffer from outlier chains and uses subspace sampling with more than one chain pair to enhance, sometimes dramatically, the convergence rate to the target distribution. Many published papers have confirmed the efficiency and robustness of this algorithm for a large array of target distributions involving (among others) complex and/or high-dimensional search spaces, with one or multiple modes and intricate multivariate parameter dependencies. This explains why the DREAM algorithm has found widespread application and use in many different fields of study to reconcile, using Bayesian inference, system models with data. In fact, several publications have shown that DREAM even provides better solutions than commonly used optimization algorithms [e.g., Vrugt and Laloy (2014)]. Whereas optimization algorithms are subject to population degeneration, the Metropolis selection rule promulgates chain diversity necessary to explore difficult search spaces in pursuit of the stationary distribution.

Multichain Methods: DREAM

The DREAM algorithm has its roots within DE-MC but uses subspace sampling and outlier chain correction to speed up convergence to the target distribution. Subspace sampling is implemented in DREAM by only updating randomly selected variables (coordinates) of \mathbf{x} each time a proposal is generated. If A is a subset of d^* -dimensions of the original search space, $\mathbb{R}^{d^*} \subseteq \mathbb{R}^d$, then a jump, $\Delta \mathbf{x}_{(i-1)}^j$ in the j th chain, $j = \{1, \dots, N\}$ at some iteration $i - 1$ is calculated from the collection of chains, $\mathbf{X}_{(i-1)} = \{\mathbf{x}_{(i-1)}^1, \dots, \mathbf{x}_{(i-1)}^N\}$, using differential evolution (Storn and Price 1997; Price et al. 2005)

$$\begin{aligned} \Delta \mathbf{x}_{(i-1),A}^j &= \zeta_{d^*} + (\mathbf{1}_{d^*} + \boldsymbol{\lambda}_{d^*}) \gamma_{(\delta, d^*)} \sum_{k=1}^{\delta} \left(\mathbf{x}_{(i-1),A}^{\mathbf{a}_k} - \mathbf{x}_{(i-1),A}^{\mathbf{b}_k} \right) \\ \Delta \mathbf{x}_{(i-1),\neq A}^j &= 0, \end{aligned} \quad (65)$$

where $\gamma_{(\delta, d^*)} = 2.38/\sqrt{2\delta d^*}$ is the jump rate, δ signifies the number of chain pairs that is used to compute the jump, and \mathbf{a} and \mathbf{b} are δ -vectors with integers drawn (without replacement) from $\{1, \dots, j-1, j+1, \dots, N\}$. The default value of $\delta = 3$ and results, in practice, in one-third of the proposals being created with $\delta = 1$, another one-third with $\delta = 2$, and the remaining one-third using $\delta = 3$. The values of λ and ζ are sampled independently from $\mathcal{U}_{d^*}(-c, c)$ and $\mathcal{N}_{d^*}(0, c_*)$, respectively, the multivariate uniform and normal distribution with, typically, $c = 0.1$ and c_* small compared to the width of the target distribution, $c_* = 10^{-6}$ say. To expedite sampling of multimodal distributions, the default jump rate is switched to unity, $\gamma_{(\delta, d^*)} = 1$, with probability 0.2. The candidate point of the j th chain at iteration i then becomes

$$\mathbf{x}_p^j = \mathbf{x}_{(i-1)}^j + \Delta \mathbf{x}_{(i-1)}^j, \quad (66)$$

and the Metropolis ratio of Eq. (64) is used to determine whether to accept this proposal or not. If $P_{\text{acc}}(\mathbf{x}_p^j) \geq \mathcal{U}(0,1)$, the candidate point is accepted and the j th chain moves to the new position, that is $\mathbf{x}_{(i)}^j = \mathbf{x}_p^j$; otherwise $\mathbf{x}_{(i)}^j = \mathbf{x}_{(i-1)}^j$. The default equation for γ should, for Gaussian and Student target distribution, result in optimal acceptance rates close to 0.44 for $d = 1$, 0.28 for $d = 5$, and 0.23 for large d (please refer to Sect. 7.84 of Robert and Casella (2004) for a cautionary note on these references acceptance rates).

The d^* -members of the subset A are sampled at random from the entries $\{1, \dots, d\}$ (without replacement) and define the dimensions of the parameter space to be sampled by each proposal. This subspace spanned by A is construed in DREAM with the help of a crossover operator as follows. First, a value $\eta \in (0, 1]$ is sampled randomly from a geometric series of r different crossover values, $\boldsymbol{\eta} = \left\{ \frac{1}{r}, \frac{2}{r}, \dots, \frac{r}{r} \right\}$, with selection probabilities, $\mathbf{p}_\eta = \left\{ \frac{1}{r}, \dots, \frac{1}{r} \right\}$. Then, a d -vector \mathbf{z} with random labels is drawn from the multivariate uniform distribution, $\mathbf{z} \sim \mathcal{U}_d(0,1)$. All those coordinates l which satisfy $z_l \leq \eta$ are stored in the subset A and span the subspace that will be sampled using Eq. (65). If A is an empty set, then one dimension of the target distribution will be sampled at random to avoid the jump vector, $\Delta \mathbf{x}_{(i-1)}^j$, to have zero entries everywhere.

We now provide an algorithmic recipe of the DREAM algorithm (see Algorithm 6). The auxiliary label Z is drawn for each chain by sampling from the standard uniform distribution, $Z \sim \mathcal{U}(0,1)$, and $\mathcal{U}\{a,b\}$ denotes the discrete uniform distribution with support $\{a, a+1, \dots, b-1, b\}$, where $a, b \in \mathbb{N}_+$ and $b \geq a$. The variable $\mathcal{F}(\boldsymbol{\eta} | \mathbf{p}_\eta)$ signifies the discrete multinomial distribution on the crossover values, $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_r\}$, with selection probabilities, $\mathbf{p}_\eta = \{p_{\eta_1}, \dots, p_{\eta_r}\}$, and the symbol $|A|$ signifies the cardinality, or number of elements, of the set A .

Algorithm 6 DiffeRential Evolution Adaptive Metropolis

- 1: Define number of chains $N \geq \lfloor d/2 \rfloor$.
- 2: Define algorithmic variables, $r, c = 0.1$ and $c_* = 10^{-6}$.
- 3: Compute r crossover values, $\boldsymbol{\eta} = \{\eta_1, \dots, \eta_r\}$ with $\mathbf{p}_\eta = \left\{ \frac{1}{r}, \dots, \frac{1}{r} \right\}$.
- 4: Set iteration, $i = 2$.
- 5: **for** $j = 1, \dots, N$ **do**
- 6: Draw randomly $\mathbf{x}_{(1)}^j$, the initial state of the j th chain, and calculate $f(\mathbf{x}_{(1)}^j)$.
- 7: **end for**
- 8: **while** chains not converged **do**
- 9: **for** $j = 1, \dots, N$ **do**
- 10: Draw the crossover value, η , from $\mathcal{F}(\boldsymbol{\eta} | \mathbf{p}_\eta)$.
- 11: Draw a d -vector \mathbf{z} from $\mathcal{U}_d(0,1)$.
- 12: Store in subset A the indexes l of \mathbf{z} that satisfy $\mathbf{z}_l \leq \eta$, where $l = \{1, \dots, d\}$.
- 13: If $A = \emptyset$ (empty set), then fill A with random draw from integers $\{1, \dots, d\}$.

```

14: Compute the cardinality of  $A$ , that is  $d^* = |A|$ .
15: Draw the value of  $\delta$  at random from integers,  $\{1, 2, 3\}$ , thus  $\delta \sim \mathcal{U}\{1,3\}$ .
16: Sample  $\delta$ -vectors  $\mathbf{a}$  and  $\mathbf{b}$  without replacement from
    $\{1, \dots, j - 1, j + 1, \dots, N\}$ .
17: Draw a label  $R \sim \mathcal{U}(0,1)$ , if  $R \leq 0.8$  set  $\gamma_{(\delta, d^*)} = 2.38/\sqrt{2\delta d^*}$  otherwise
    $\gamma_{(\delta, d^*)} = 1$ .
18: Create a candidate point,  $\mathbf{x}_p^j$ , in the  $j$ th chain using Eqs. (65) and (66).
19: Compute the target density,  $f(\mathbf{x}_p^j)$ , at  $\mathbf{x}_p^j$ .
20: Calculate the Metropolis ratio,  $P_{\text{acc}}(\mathbf{x}_p^j) = \min \left[ 1, f(\mathbf{x}_p^j)/f(\mathbf{x}_{(i-1)}^j) \right]$ .
21: if  $Z \leq P_{\text{acc}}(\mathbf{x}_p^j)$  then
22:   Set  $\mathbf{x}_{(i)}^j = \mathbf{x}_p^j$  and  $f(\mathbf{x}_{(i)}^j) = f(\mathbf{x}_p^j)$ .
23: else
24:   Remain at “old” state,  $\mathbf{x}_{(i)}^j = \mathbf{x}_{(i-1)}^j$  and  $f(\mathbf{x}_{(i)}^j) = f(\mathbf{x}_{(i-1)}^j)$ .
25: end if
26: end for
27: Compute convergence diagnostics.
28: Patch for dissident chains.
29: Update iteration,  $i = i + 1$ .
30: end while

```

The use of a vector of crossover probabilities enables single-site Metropolis (A contains one element), Metropolis-within-Gibbs (A has one or more elements) and regular Metropolis sampling ($A = \{1, \dots, d\}$), and constantly introduces new directions in the parameter space that chains can take outside the subspace spanned by their current positions. What is more, the use of subspace sampling allows using $N < d$ in DREAM, an important advantage over DE-MC that requires $N = 2d$ chains to be run in parallel (ter Braak 2006). This randomization of the search space introduces one additional algorithmic variable to the algorithm, namely, the desired number of crossover values, r . The default setting of $r = 3$ has shown to work well in practice, but larger values of this algorithmic variable might enhance the convergence rate on high-dimensional target distributions, say $d > 50$, to preserve the frequency of low-dimensional jumps. Note, more intelligent subspace selection methods can be devised for target distributions with complex multivariate dependencies among the parameters. Correlated parameters should preferably be sampled jointly in tandem, otherwise too many of the (subspace) proposals will be rejected and the search can stagnate.

To enhance search efficiency, the selection probabilities, $\mathbf{p}_\eta = \{p_{\eta_1}, \dots, p_{\eta_r}\}$, of the different crossover values, $\{\eta_1, \dots, \eta_r\}$ are adapted during a burn-in period to maximize the distance traveled (Eulicdean norm) by the N chains. This adaptation is described in detail by Vrugt et al. (2008, 2009). Here, we also discuss the patch for dissident chains as described in detail by Vrugt et al. (2008).

We provide in Appendix B a numerical recipe of the DREAM algorithm in MATLAB. This code summarizes in about 30 lines DREAM's computational engine, yet has several important restrictions for reasons of brevity and clarity. Among others, the code does not adapt the crossover selection probabilities, assumes by default a uninformative prior distribution, requires the user to implement their desired likelihood function, does not enforce parameter boundaries, and does not monitor convergence of the sampled chain trajectories (a fixed budget of M iterations is assumed). Nevertheless, the MATLAB recipe in Appendix B should satisfy as template of DREAM for own applications for those readers proficient in statistics, computer coding, and numerical computation. Other readers are referred to the MATLAB toolbox of DREAM developed by the first author of this chapter Vrugt (2016). This package enjoys many built-in utilities and functions that simplify considerable the widespread application and use of MCMC simulation for Bayesian analysis and inverse modeling.

In the past years, several other MCMC algorithms have appeared in the literature which use DREAM as their basic building block but include special extensions to simplify inference (among others) of discrete and combinatorial search spaces, and high-dimensional and CPU-intensive system models. This includes the DREAM_(ZS) (Laloy and Vrugt 2012), DREAM_(D) (Vrugt et al. 2011), DREAM_(DZS), DREAM_(ABC) (Sadegh and Vrugt 2014), DREAM_(LOA) (Vrugt and Beven 2018), and MT-DREAM_(ZS) (Laloy and Vrugt 2012) algorithms. Most of these methods are discussed in Vrugt (2016) and have their own MATLAB toolbox. The different DREAM algorithms are also available in DREAM Suite, an easy to use Windows program.

Convergence Monitoring

Per theory, the chains that are simulated by a MCMC algorithm are expected to eventually converge to a stationary distribution, which should be the desired target distribution. But, how do we actually assess that convergence has been achieved in practice, without knowledge of the actual target distribution?

One way to check for convergence is to see how well the chains are mixing, or moving around the parameter space. For a properly converged MCMC sampler, the chains should sample, for a sufficiently long period, the approximate same part of the parameter space, and mingle readily and in harmony with one another around some fixed mean value. This can be inspected visually for each dimension of \mathbf{x} separately, and used to diagnose convergence informally.

Another diagnostic that can be used to monitor convergence is the acceptance rate. A value between 15 – 30% is usually indicative of good performance of a MCMC simulation method. Much lower values usually convey that the posterior surface is difficult to traverse in pursuit of the target distribution. A low acceptance rate can have different reasons, for instance poor model numerics, or the presence of multimodality and local optima. Yet, the acceptance rate can only diagnose whether a MCMC method such as DREAM is achieving an acceptable performance, it cannot be used to determine when exactly convergence has been achieved.

Several non-parametric and parametric statistical tests can be used to determine when convergence of the sampled chains to a limiting distribution has been achieved. The most powerful of these convergence tests is the multi-chain \hat{R} -statistic of Gelman and Rubin (1992). This diagnostic compares for each parameter $l = \{1, \dots, d\}$ the within-chain

$$W_l = \frac{2}{N(T-2)} \sum_{j=1}^N \sum_{i=\lfloor T/2 \rfloor}^T \left(\mathbf{x}_{(i),l}^j - \bar{\mathbf{x}}_l^j \right)^2 \quad \bar{\mathbf{x}}_l^j = \frac{2}{T-2} \sum_{i=\lfloor T/2 \rfloor}^T \mathbf{x}_{(i),l}^j \quad (67)$$

and between-chain variance

$$B_l/T = \frac{1}{2(N-1)} \sum_{j=1}^N \left(\bar{\mathbf{x}}_l^j - \bar{\bar{\mathbf{x}}}_l \right)^2 \quad \bar{\bar{\mathbf{x}}}_l = \frac{1}{N} \sum_{j=1}^N \bar{\mathbf{x}}_l^j \quad (68)$$

using

$$\hat{R}_l = \sqrt{\frac{N+1}{N} \frac{\hat{\sigma}_+^{2(l)}}{W_l} - \frac{T-2}{NT}}, \quad (69)$$

where T signifies the number of samples in each chain, $\lfloor \cdot \rfloor$ is the integer rounding operator, and $\hat{\sigma}_+^{2(l)}$ is an estimate of the variance of the l th parameter of the target distribution

$$\hat{\sigma}_+^{2(l)} = \frac{T-2}{T} W_l + \frac{2}{T} B_l. \quad (70)$$

To officially declare convergence, the value $\hat{R}_l \leq 1.2$ for each parameter, $l \in \{1, \dots, d\}$, otherwise the value of T should be increased and the chains run longer. As the N different chains are launched from different starting points, the \hat{R} -diagnostic is a relatively robust estimator.

A related, but more powerful convergence diagnostic is the multivariate variant of the \hat{R} -statistic of Gelman and Rubin (1992). This statistic, hereafter referred to as \hat{R}^d -diagnostic, is defined in Brooks and Gelman (1998) and assesses convergence of the d parameters simultaneously by comparing their within and between-sequence covariance matrix. Convergence is achieved when a rotationally invariant distance measure between the two matrices indicates that they are “sufficiently” close. Then, the multivariate \hat{R}^d -statistic achieves a value close to unity, otherwise its value is much larger. In fact, the \hat{R} and \hat{R}^d -statistic take on a very similar range of values, hence simplifying analysis of when convergence has been achieved. The \hat{R}^d -statistic is particularly useful for high-dimensional target distributions involving complicated multi-dimensional parameter interactions.

Other statistics include the autocorrelation function, and the Geweke (1992), and Raftery and Lewis (1992)-diagnostics. The autocorrelation function for each parameter $l = \{1, \dots, d\}$ is defined as

$$\rho_{l,k}^j = \frac{\sum_{i=1}^{T-k} (\mathbf{x}_{(i),l}^j - \bar{\mathbf{x}}_l^j)(\mathbf{x}_{(i+k),l}^j - \bar{\mathbf{x}}_l^j)}{\sum_{i=1}^T (\mathbf{x}_{(i),l}^j - \bar{\mathbf{x}}_l^j)^2}, \quad (71)$$

and returns the correlation between two samples k iterations apart in the j th chain, $j = \{1, \dots, N\}$. Compared to rejection sampling which, per construction, produces uncorrelated samples, MCMC chain trajectories exhibit autocorrelation as the current state of the chain is derived from its previous state. This correlation is expected to decrease with increasing lag k . The autocorrelation function is a useful proxy to assess sample variability and mixing, but does not convey when convergence has been achieved. A high autocorrelation, say $|\rho| > 0.8$, at lags, say $k \geq 5$, demonstrates a poor mixing of the individual chains.

The Geweke (1992)-diagnostic compares the means of two nonoverlapping parts of the Markov chain using a standard Z-score adjusted for autocorrelation. The Raftery and Lewis (1992)-statistic calculates the number of iterations, T and length of burn-in necessary to satisfy the condition that some posterior quantile of interest, say q has a probability, p of lying within interval $[q - r, q + r]$. Default values are $q = 0.025$, $p = 0.95$, and $r = 0.01$, respectively. Details of how to compute and interpret these two statistics is found in the cited references.

Altogether, joint interpretation of the different diagnostics should help assess convergence of the sampled chain trajectories. Of all these metrics, the \hat{R}^d -statistic is most conservative and strict and provides the best guidance on exactly when convergence has been achieved. This happens as soon as this statistic drops below the critical threshold of 1.2. Suppose this happens at T^* iterations (generations), then the first $(T^* - 1)$ samples of each chain are simply discarded as burn-in and the remaining $N(T - T^*)$ samples from the joint chains are used for posterior analysis. Note, we always recommend to verify convergence of DREAM by visually inspecting the mixing of the different chain trajectories.

In practice, one has to make sure that a sufficient number of chain samples is available for the inference, otherwise the posterior estimates can be biased. For convenience, we list here the total number of posterior samples, $N(T - T^*)$ (in brackets) one would need for a reliable inference with DREAM for a given dimensionality of the target distribution: $d = 1$ (500); $d = 2$ (1,000); $d = 5$ (5,000); $d = 10$ (10,000); $d = 25$ (50,000); $d = 50$ (200,000); $d = 100$ (1,000,000); and $d = 250$ (5,000,000). These listed numbers are only a rough guideline and based on several assumptions such as a reasonable acceptance rate ($>10\%$) and not too complicated shape of the posterior distribution. In general, the number of posterior samples required increases with rejection rate and complexity of the target distribution.

6 Case Studies

In this section, we illustrate the application of Bayesian inference to four different case studies involving modeling of the instantaneous unit hydrograph, the rainfall-runoff transformation, and water movement in the variably saturated zone. These examples focus on the movement and distribution of water in the surface and subsurface and satisfy the overarching theme of this book. The last two case studies have appeared in literature publications; therefore, we only present herein a short summary of the experimental data, models, numerical setup, and findings. Interested readers are referred to the original publications for further details.

6.1 Instantaneous Unit Hydrograph

The first case study considers the modeling of the instantaneous unit hydrograph using the ordinates of Nash (1960) defined as

$$Q_t = \frac{1}{r_c \Gamma(h)} \left(\frac{t}{r_c} \right)^{(h-1)} \exp\left(-\frac{t}{r_c}\right), \quad (72)$$

where Q_t (mm day^{-1}) signifies the simulated streamflow at time t (days), h (–) denotes the number of reservoirs, r_c (days) signifies the recession constant, and $\Gamma(\cdot)$ is the gamma function in Eq. (45).

A $n = 25$ - day period with synthetic daily streamflow data was generated by driving Eq. (72) with an artificial precipitation record using $h = 2$ reservoirs and a recession constant of $r_c = 4$ days. This artificial data set is subsequently perturbed with a heteroscedastic measurement error (nonconstant variance) with standard deviation equal to 10% of the original simulated discharge values. We now use the n -record of corrupted discharge values, $\tilde{\mathbf{Y}} = \{\tilde{y}_1, \dots, \tilde{y}_n\}$, to estimate the posterior distribution, $P(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$, of the Nash parameters, $\boldsymbol{\theta} = \{h, r_c\}$, using Bayes' theorem in Eq. (13). We assume a flat prior distribution for the two Nash parameters, $P(\boldsymbol{\theta}) \sim \mathcal{U}_2[1, 10]$, and use the likelihood function, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}}, \hat{\sigma}^2)$, of Eq. (31). The n -vector of measurement error variances, $\hat{\sigma}^2 = \{\hat{\sigma}_1^2, \dots, \hat{\sigma}_n^2\}$, is derived from Eq. (42) using the nuisance variables, $\boldsymbol{\alpha} = \{\sigma_0, \sigma_1\}$. We estimate the joint posterior distribution of the Nash model parameters and the nuisance variables, $P(\boldsymbol{\theta}, \boldsymbol{\alpha} | \tilde{\mathbf{Y}})$, using MCMC simulation with DREAM (Algorithm 6).

Figure 11 summarizes the results of our analysis and presents a scatter plot matrix of the posterior samples derived with the DREAM algorithm. The main diagonal displays histograms of the marginal distribution of the two Nash model parameters h and r_c and the nuisance variables σ_0 and σ_1 , whereas the off-diagonal graphs display bivariate scatter plots of the posterior samples. The x -axes matches exactly the posterior ranges of the parameters and their “true” values, $h = 4$, $r_c = 2$, $\sigma_0 = 0$, and $\sigma_1 = 0.1$, are separately indicated in each histogram with the red cross.

The posterior histograms of the Nash model parameters follow closely a Gaussian distribution with mean that is in excellent agreement (as should be!) with the true

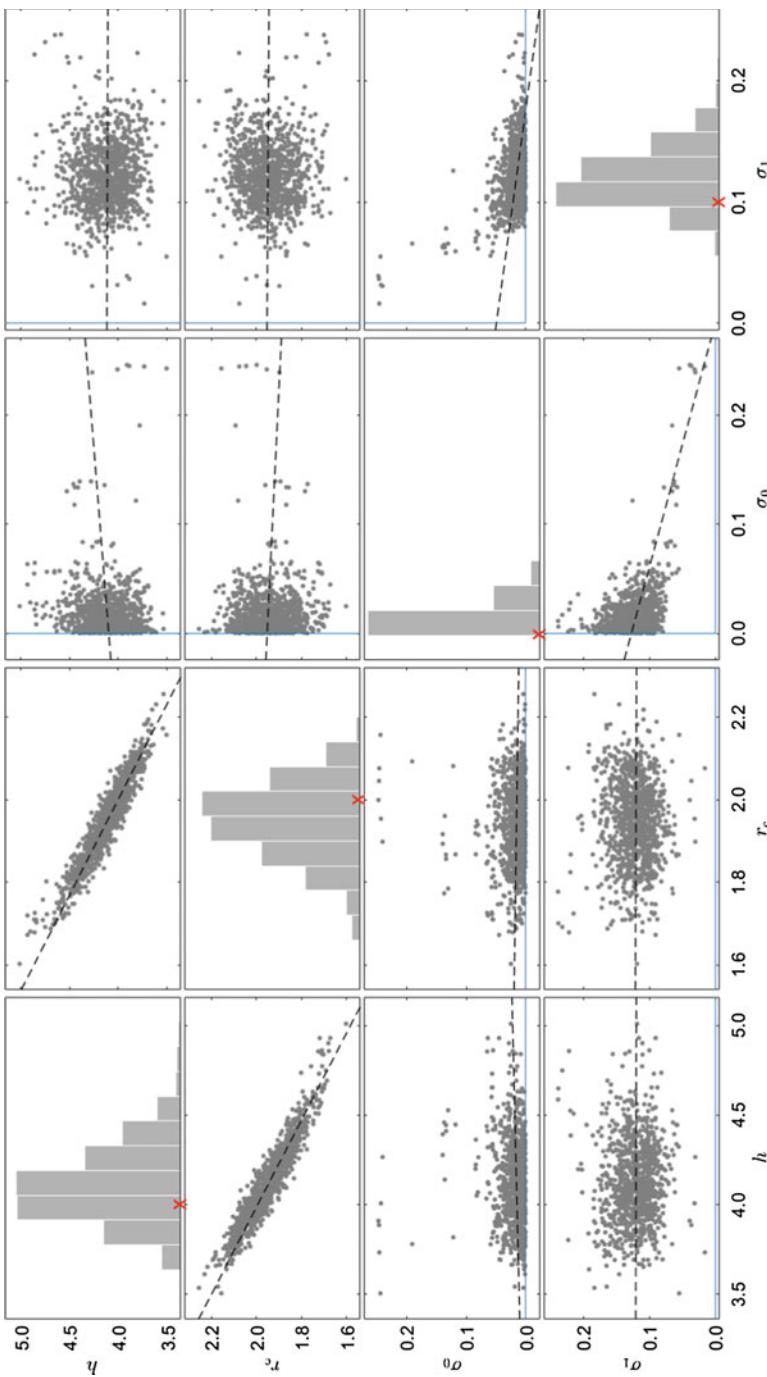


Fig. 11 Unit hydrograph: scatter-plot matrix of the posterior samples created the DREAM algorithm. The main diagonal presents histograms of the marginal posterior distributions of the Nash model parameters, h and r_c , and nuisance variables, σ_0 and σ_1 . The off-diagonal graphs display bivariate scatter plots of the posterior samples of the different parameter pairs. The true parameter values are separately indicated in each histogram using the red cross

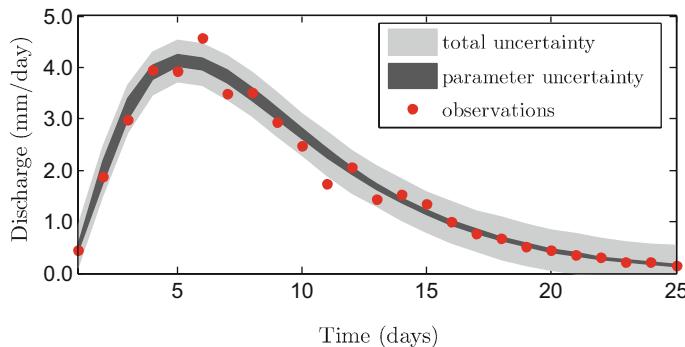


Fig. 12 Unit hydrograph: comparison of the observed (red dots) and posterior simulated hydrographs. The light and dark gray regions demarcate the 95% simulation intervals of the Nash model due to parameter and total uncertainty, respectively

values used to create the synthetic 25-day hydrograph. The marginal distributions exhibit a relatively small dispersion (in comparison to prior ranges), which demonstrates that h and r_c are well defined by calibration against the observed (synthetic) streamflow data. The nuisance variables of Eq. (42) are also well resolved, yet the marginal posterior distribution of the slope of the linear measurement error model, σ_1 is somewhat biased towards larger values. Indeed, the true value of the slope appears to the left of the maximum a-posteriori (MAP) density solution of σ_1 in the left tail of the distribution. The intercept, σ_1 , of Eq. (42) is truncated at zero by the uniform prior distribution and consequently follows a log-normal distribution. Such distribution is difficult to approximate accurately with the first-order approximation of Eq. (6) but poses no problems for sampling methods. The bivariate scatterplots of the posterior samples (off-diagonal plots) demonstrate a strong linear (negative) correlation between the Nash model parameters, h and r_c . The nuisance variables and Nash parameters appear uncorrelated, a testament to the use of an adequate likelihood function.

We next investigate how the posterior parameter uncertainty translates into simulation uncertainty of the Nash model. Figure 12 presents a time series plot of the observed data (red dots) and 95% simulation uncertainty ranges of the hydrograph due to parameter (dark gray region) and total uncertainty (light gray region). The Nash model tracks closely the observed data (as is to be expected with synthetic data) with simulation intervals that appear reasonably small. Further analysis of the residuals confirms our assumptions that the residuals are temporally uncorrelated and follow an approximately Gaussian distribution. We next present an example in which the residuals do not satisfy our assumptions.

6.2 The Rainfall-Runoff Transformation

The second case study involves simulation of the rainfall-discharge relationship. The present case study has appeared in Schoups and Vrugt (2010) and readers are

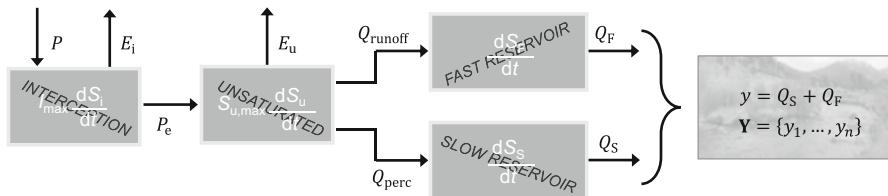


Fig. 13 The rainfall-runoff transformation: schematic representation of the hmodel conceptual watershed model

referred to this work for a detailed description of the watershed model and experimental data used herein.

We use the 7-parameter hmodel of Schoups and Vrugt (2010) to simulate daily discharge time series of the French Broad River at Asheville, North Carolina, USA. A schematic overview of the hmodel appears in Fig. 13. This rather parsimonious conceptual watershed model transforms daily estimates of mean areal precipitation and potential evapotranspiration into discharge emanating from the catchment outlet. The hmodel uses four different control volumes to characterize the water storage and distribution within the basin. These reservoirs define the state of the watershed and tie the atmospheric forcing to processes such as canopy interception, throughfall, interception evaporation, actual evapotranspiration, surface runoff, percolation, and surface and subsurface routing, which deplete or replenish the tanks. The climate of the French Broad River basin is rather mild and does not demand simulation of snow accumulation and/or melt.

The seven parameters of the hmodel and their feasible ranges are listed in Table 1.

We estimate the hmodel parameters, $\boldsymbol{\theta} = \{I_{\max}, S_{\max}, Q_{\max}, \alpha_E, \alpha_F, K_F, \alpha_S\}$, using a 7-year record of historical data from the French Broad River basin using daily measurements of the discharge and estimates of the mean areal precipitation and potential evapotranspiration. All variables have units of mm/day. A 2-year spin-up period is used to reduce sensitivity of the model to state-value initialization.

We assume a uninformative prior distribution, $P(\boldsymbol{\theta})$, for the parameters of the hmodel using the ranges listed in Table 1. As a first attempt, we expect conveniently that the discharge residuals are uncorrelated, normally distributed, and with a constant variance. These assumptions lead to the Gaussian likelihood function, $L(\boldsymbol{\theta} | \tilde{\mathbf{Y}})$, in Eq. (29) – and with our uniform prior leads to the density function of Eq. (30).

Figure 14 displays the results of the DREAM algorithm. The top panel plots the simulation of the maximum likelihood solution for the 1800-day calibration data record, whereas the bottom panel presents a diagnostic check of the corresponding discharge residuals. The hmodel simulation tracks closely the observed discharge data (dots), with 95% intervals (not shown) that appear rather small. Unfortunately, the residuals do not satisfy the three main underlying assumptions of our likelihood function (Gaussian, uncorrelated, constant variance). Indeed, the variance of the residuals increases with flow level, the residual distribution deviates from normality

Table 1 Model parameters and their prior uncertainty ranges

Parameter	Symbol	Minimum	Maximum	Units
Maximum interception	I_{\max}	1	10	mm
Soil water storage capacity	S_{\max}	10	1000	mm
Maximum percolation rate	Q_{\max}	0.1	100	mm/d
Evaporation parameter	α_E	0.1	100	—
Runoff parameter	α_F	-10	10	—
Time constant, fast reservoir	K_F	0.1	10	days
Time constant, slow reservoir	K_S	0.1	150	days

(much more peaked), and the residuals display considerable serial correlation at the first lag.

We now relax the residual assumptions and use the generalized likelihood function of Eq. (58) to derive the parameters of the hmodel. This requires inference as well of the nuisance variables, $\boldsymbol{\alpha} = \{\sigma_0, \sigma_1, \beta, \xi, \Phi_p\}$. We use $p = 4$ and derive the posterior distribution of the hmodel parameters and nuisance variables, $P(\boldsymbol{\theta}, \boldsymbol{\alpha} | \tilde{Y})$ using MCMC simulation with the DREAM algorithm. Figure 15 presents the main results of this analysis.

It is evident that the residuals of the maximum likelihood simulation (top plot) now are consistent with the properties of the likelihood function. The residual variance is constant and independent of flow level, the residual distribution is well-described with a Laplace distribution, and the residuals are temporally uncorrelated.

Figure 16 summarizes the effect of the choice of likelihood function on the posterior marginal distribution of the hmodel parameters I_{\max} , S_{\max} , Q_{\max} , α_F , K_F , and K_S . The blue histograms display the results of the Gaussian likelihood function (Eq. 29) and their green counterparts pertain to the generalized likelihood function (Eq. 58).

The results in Fig. 16 highlight several important findings. First, the marginal distributions of the hmodel parameters differ substantially between the two different likelihood functions. Second, the posterior histograms derived from the generalized likelihood function exhibit a larger dispersion. Finally, the marginal distributions derived from the generalized likelihood functions are not truncated by the prior distribution. This is particularly evident for the maximum interception, I_{\max} , which hits its upper bound of 10 mm for the Gaussian likelihood of Eq. (29) and takes on much more realistic values of 1–4 mm with the generalized likelihood function.

This concludes the second case study. Interested readers are referred to Schoups and Vrugt (2010) for a much more detailed interpretation and treatment of the results.

6.3 Vadose Zone Hydrology

The third and last case study considers the modeling of the soil moisture regime of an agricultural field near Jülich, Germany. Soil moisture content was measured with

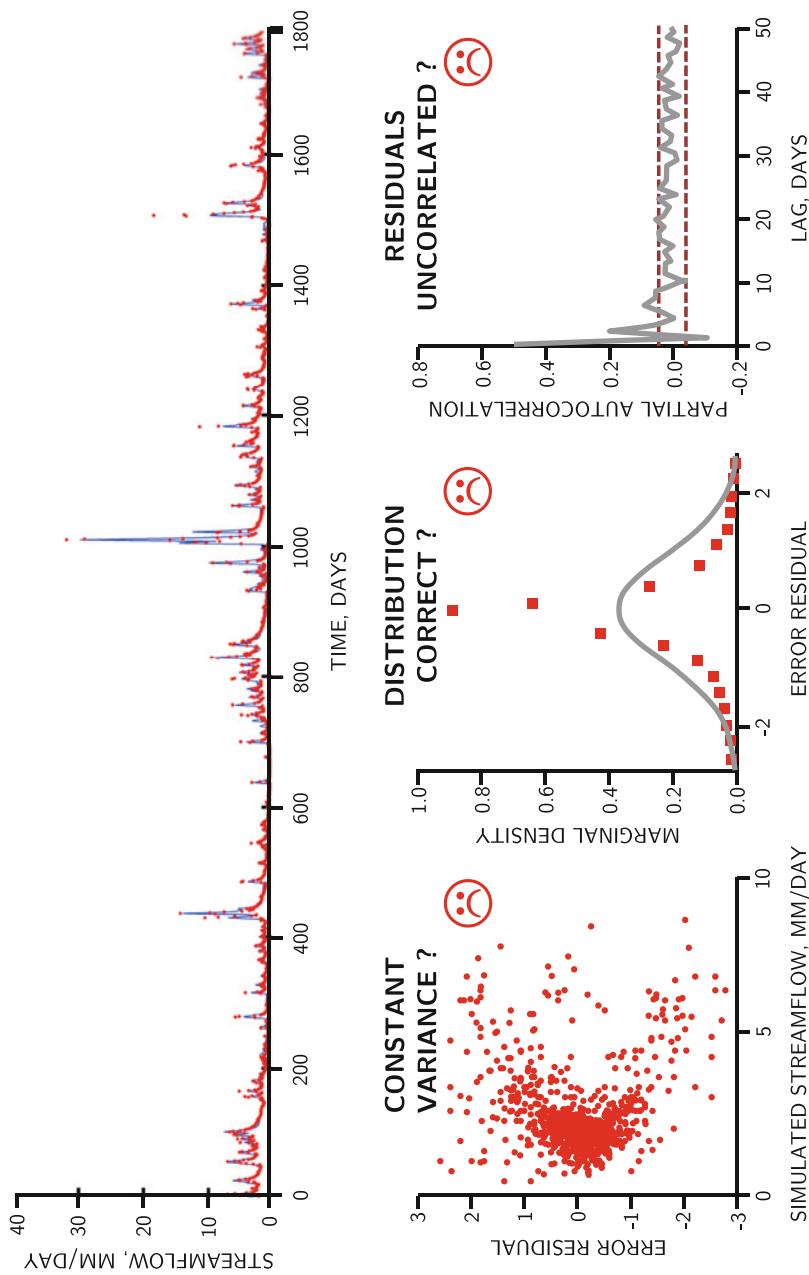


Fig. 14 The rainfall-runoff transformation: least-squares calibration with Gaussian likelihood of Eq. (29) for the French Broad River basin. Time series plot (top panel) of maximum likelihood streamflow simulation (solid blue line) and observations (red dots). The bottom panel analyzes the corresponding residuals. The left plot displays the residuals as a function of simulated streamflow. The middle plot compares the assumed (solid line) and actual (red squares) histogram of the residuals. The right plot summarizes the autocorrelation function of the residuals. The 95% significance levels of white noise are separately indicated in this graph with the dashed lines

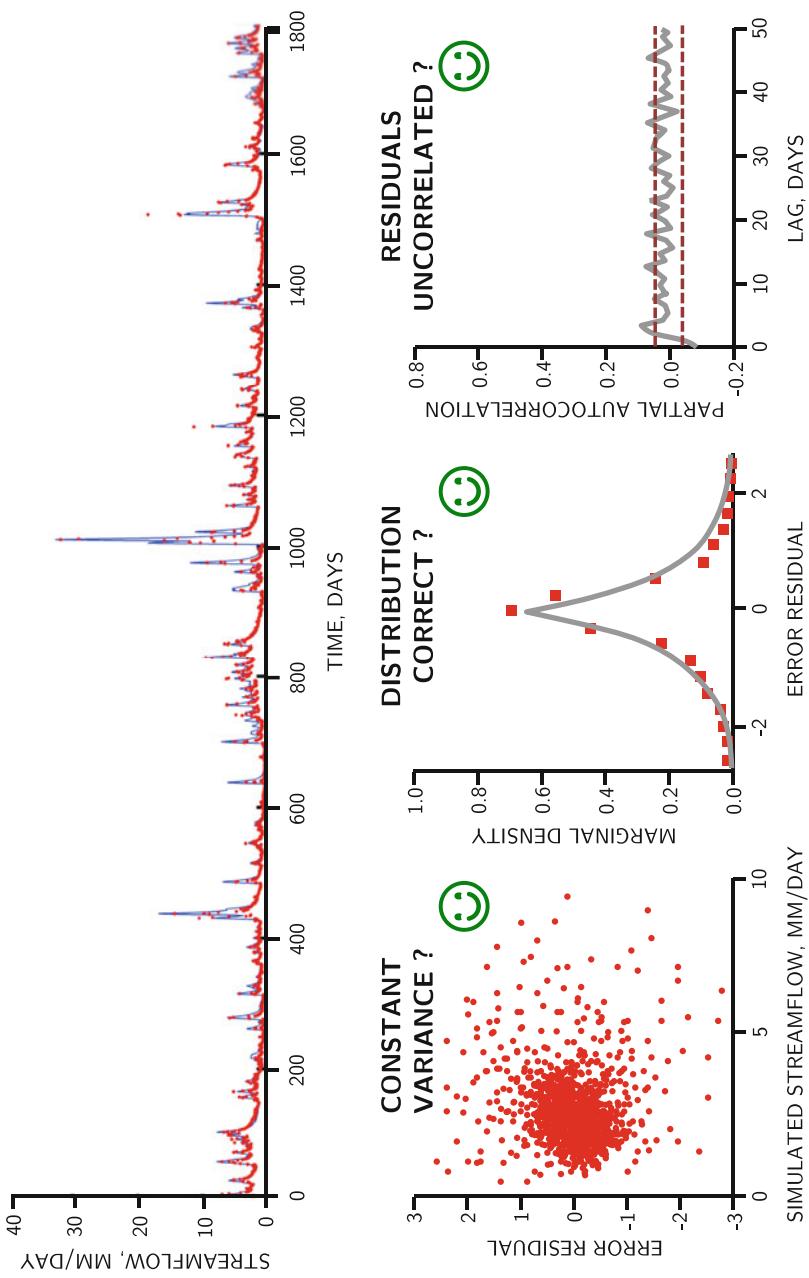
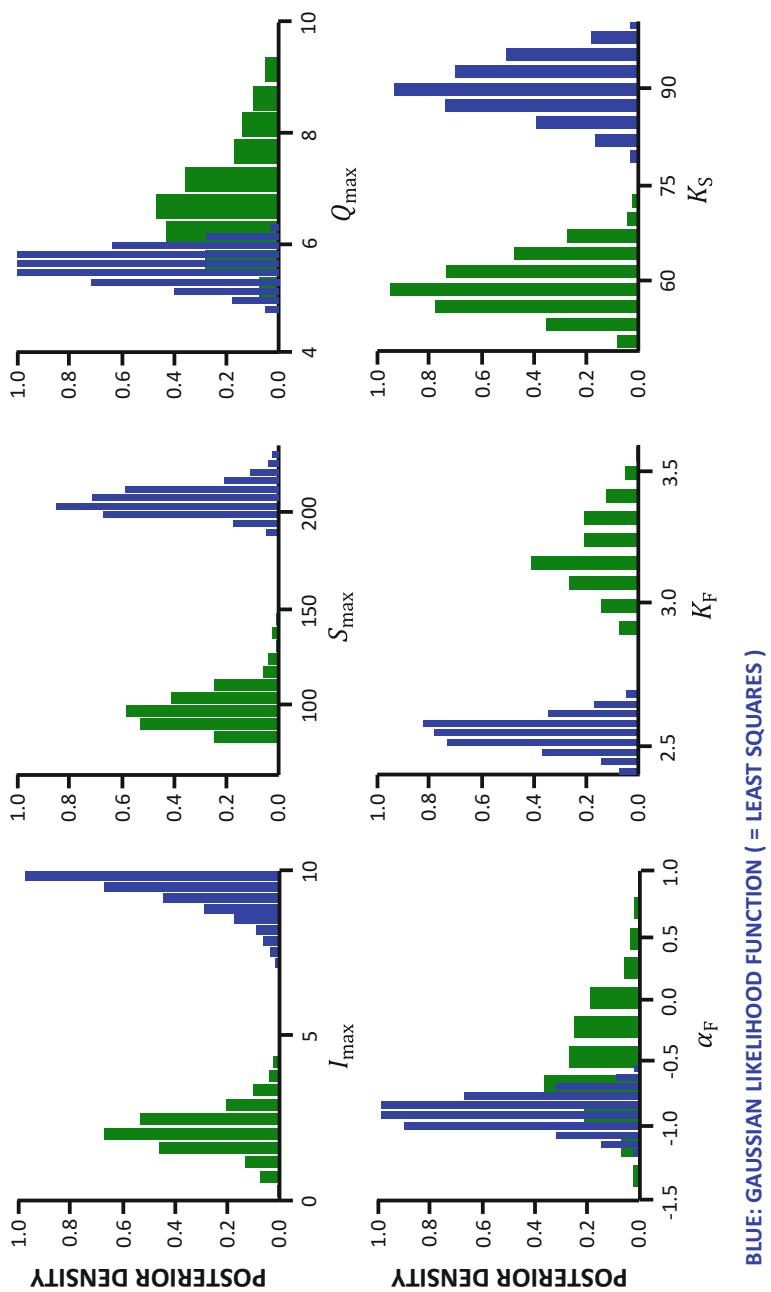


Fig. 15 The rainfall-runoff transformation: calibration with the generalized likelihood function of Eq. (58) for the French Broad River basin. Time series plot (top panel) of maximum likelihood streamflow simulation (solid blue line) and observations (red dots). The bottom panel analyzes the corresponding residuals. The left plot displays the residuals as a function of the simulated flow level. The middle plot shows the assumed (solid line) and actual (red squares) histogram of the residuals. The right plot summarizes the autocorrelation function of the residuals. The dashed lines in this graph signify the 95% significance levels of white noise



BLUE: GAUSSIAN LIKELIHOOD FUNCTION (= LEAST SQUARES)

GREEN: GENERALIZED LIKELIHOOD FUNCTION

Fig. 16 The rainfall-runoff transformation: histograms of the marginal posterior distribution of the hmodel parameters for the French Broad River basin using (in blue) the Gaussian likelihood of Eq. (29) and (in green) the generalized likelihood of Eq. (58)

time domain reflectometry (TDR) probes at 6 cm deep at 61 locations in a 50×50 m experimental plot. The TDR data were analyzed using the algorithm described in Heimovaara and Bouten (1990) and the measured apparent dielectric permittivities were converted to soil moisture values using the empirical relationship of Topp et al. (1980). Measurements were taken on $n = 29$ days between 19 March and 14 October 2009, comprising a measurement campaign of 210 days. For the purpose of the present study, the soil moisture observations were averaged per day to yield a single plot-mean water content time series. Precipitation and other meteorological variables were recorded at a meteorological station located 100 m west of the measurement site. Details of the site, soil properties, experimental design, and measurements are given by Scharnagl et al. (2011) and interested readers are referred to this publication for further details.

The HYDRUS-1D model of Šimůnek et al. (2008) was used to simulate variably saturated water flow in the agricultural field. This model uses the finite element method to solve numerically Richards' equation

$$\frac{\partial \theta}{\partial t} = \frac{\partial}{\partial z} \left[K(h) \left(\frac{\partial h}{\partial z} + 1 \right) \right], \quad (73)$$

for given (measured) initial and boundary conditions, where θ ($\text{cm}^3 \text{ cm}^{-3}$) signifies moisture content, t (days) denotes time, z (cm) is the vertical (depth) coordinate, h (cm) is the pressure head, and $K(h)$ (cm day^{-1}) represents the corresponding (unsaturated) soil hydraulic conductivity. The symbol θ ($\text{cm}^3 \text{ cm}^{-3}$) is used in the hydrologic sciences to denote volumetric soil moisture content. In this chapter, we also use this symbol to denote the parameter vector.

Numerical solution of Eq. (73) requires knowledge of the soil hydraulic properties. We use the van Genuchten-Mualem (VGM) model (van Genuchten 1980)

$$\begin{aligned} \theta(h) &= \theta_r + (\theta_s - \theta_r) [1 + (\alpha|h|)^n]^{-m} \\ K(h) &= K_s S_e(h)^\lambda \left[1 - \left(1 - S_e(h)^{1/m} \right)^m \right]^2, \end{aligned} \quad (74)$$

where θ_s and θ_r ($\text{cm}^3 \text{ cm}^{-3}$) signify the saturated and residual soil water content, respectively, α (cm^{-1}), n (–) and $m = 1 - 1/n$ (–) are shape parameters, K_s (cm day^{-1}) denotes the saturated hydraulic conductivity, and λ (–) represents a pore-connectivity parameter. The effective saturation, S_e (–), is defined as

$$S_e(h) = \frac{\theta(h) - \theta_r}{\theta_s - \theta_r}. \quad (75)$$

Observations of daily precipitation and daily potential evapotranspiration are used to characterize the upper boundary condition of our experimental plot. In the absence of detailed knowledge of the lower boundary condition, we assume a constant head, h_{bot} (cm), at the bottom of our modeled soil domain and estimate its value along with the six soil hydraulic parameters of Eq. (74).

Fig. 17 Vadose zone hydrology: schematic representation of the HYDRUS-1D model setup for the experimental field plot near Jülich, Germany

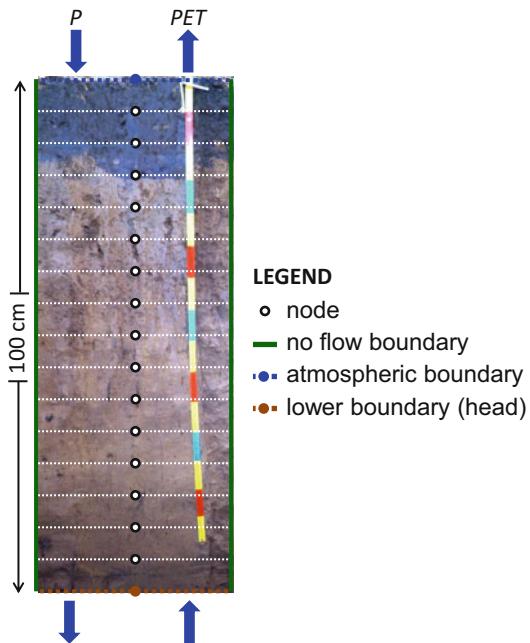


Figure 17 illustrates the setup of the HYDRUS-1D model for our experimental plot. The vertical discretization of the 100 cm deep soil profile was determined via trial-and-error to minimize the CPU-time of the model simulations. A total of 81 nodes were deemed sufficient for an accurate numerical solution with negligible mass balance errors. Nodal distance was closest at the soil surface (0.05 cm) in response to the highly dynamic atmospheric boundary conditions and increased gradually with depth to 3.5 cm at the bottom of the modeled soil domain.

Table 2 lists the parameters of the HYDRUS-1D model, their units, upper and lower bounds for the soil domain under investigation, and respective prior distributions. These parameters, $\Theta = \{\theta_r, \theta_s, \alpha, n, K_s, \lambda, h_{bot}\}$, are subject to inference using the soil moisture measurements. The marginal priors of the six soil hydraulic parameters (first six in table) are construed from surrogate data (soil texture) in Scharnagl et al. (2011) using the Rosetta toolbox of hierarchical pedotransfer functions (Schaap et al. 1998, 2001). A flat prior with ranges between -250 and -10 cm is used for the bottom head, h_{bot} . Furthermore, we use the folding method as explicated by Vrugt (2016, p. 289) to enforce strictly the ranges of Table 2. This guarantees the parameters to take on values that are deemed physically realistic. Folding is the only efficient boundary treatment procedure that does not destroy detailed balance of the sampled chain trajectories.

The initial state of each chain is sampled from the prior distribution, and the $N = 10$ different chains ran in parallel using the MATLAB parallel computing toolbox.

Table 2 Parameters of the HYDRUS-1D model and their prior uncertainty ranges

Parameter	Symbol	Lower	Upper	Units	Prior
Residual moisture content	θ_r	0.043	0.091	$\text{cm}^3 \text{ cm}^{-3}$	$\mathcal{N}(0.067, 0.006)$
Saturated moisture content	θ_s	0.409	0.481	$\text{cm}^3 \text{ cm}^{-3}$	$\mathcal{N}(0.445, 0.009)$
Reciprocal of air-entry value	α	0.003	0.009	cm^{-1}	$\mathcal{N}(0.005, 6.90 \cdot 10^{-4})$
Curve shape parameter	n	1.510	1.849	—	$\mathcal{N}(1.671, 0.042)$
Conductivity at saturation	K_s	0.138	19.962	cm day^{-1}	$\mathcal{N}(1.660, 1.386)$
Tortuosity parameter	λ	-5.490	6.270	—	$\mathcal{N}(0.390, 1.470)$
Pressure head at bottom	h_{bot}	-250	-50	cm	$\mathcal{U}(-250, 50)$

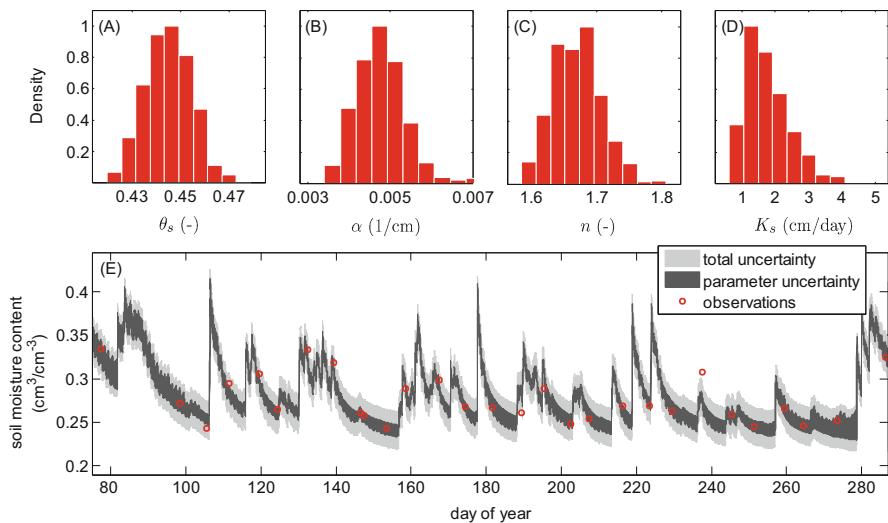
**Fig. 18** Vadose zone hydrology: histograms of the marginal posterior distribution of the soil hydraulic parameters, (a) θ_s , (b) α , (c) n , and (d) K_s , and (e) HYDRUS-1D 95% simulation uncertainty intervals due to parameter (dark region) and total uncertainty (light gray). The observed soil moisture values are indicated with a red circle

Figure 18 presents histograms of the marginal posterior distributions sampled by DREAM of the HYDRUS parameters θ_s , α , n , and K_s of the 100 cm deep soil of our 50×50 experimental plot in Germany. The bottom panel presents a time series plot of simulated soil moisture contents. The dark gray region constitutes the 95% HYDRUS-1D simulation uncertainty due to parameter uncertainty, whereas the light gray region denotes the total simulation uncertainty (parameter + randomly sampled additive error). The observed soil moisture values are indicated with a red circle.

The soil hydraulic parameters appear well defined by calibration against the observed soil moisture measurements. Their marginal distributions follow closely their respective normal prior distributions (Table 2), yet exhibit somewhat less dispersion. The HYDRUS-1D model closely tracks the observed soil moisture contents with Root Mean Square Error (RMSE) of the posterior mean simulation of about $0.01 \text{ cm}^3/\text{cm}^{-3}$. About 95% of the observations lie within the gray region, an indication that the simulation uncertainty ranges are statistically adequate. The acceptance rate of DREAM averages about 12.6% – about half of its theoretical optimal value of 22–25% (for Gaussian and Student target distributions). This deficiency is explained in part by the high nonlinearity of retention and hydraulic conductivity functions and numerical errors of the implicit, time-variable, solver of the Richards' equation. This introduces irregularities (e.g., local optima) in the posterior response surface and makes the journey to and sampling from the target distribution more difficult.

7 Limits of Acceptability

The Bayesian approach as demonstrated in action in the previous section requires users to make explicit assumptions about the nature and properties of the measurement and modeling errors. Such assumptions are often easily criticized in practical applications, in lieu of an alternative, quasi-Bayesian approach. For example, Beven (2006) suggested that a more rigorous approach to model evaluation would involve the use of limits of acceptability for each individual observation against which model simulated values are compared. Within this framework, behavioral models are defined as those that satisfy the limits of acceptability for each observation. This approach may be more objective than the standard GLUE approach advocated in Beven and Binley (1992) as the limits are defined before running the model on the basis of best available knowledge. Ideally, the limits of acceptability should reflect the observational error of the variable being compared, together with the effects of input error and commensurability errors resulting from time or space scale differences between observed and predicted values (Beven and Binley 2014). The limits of acceptability approach has been used by various authors (Blazkova and Beven 2009; Dean et al. 2009; Krueger et al. 2009; Liu et al. 2009; McMillan et al. 2010; Westerberg et al. 2011), although earlier publications used similar ideas within GLUE based on fuzzy measures (Page et al. 2003; Freer et al. 2004; Page et al. 2004, 2007; Pappenberger et al. 2005, 2007). Next, we interpret the limits of acceptability approach within the context of set theory, and then we demonstrate that much of what we have learned thus far is still of relevance and use in this quasi-Bayesian approach.

For now, let's assume that the prior distribution, $P(\boldsymbol{\theta}) \sim \mathcal{U}_d(\mathbf{a}, \mathbf{b})$, is multivariate uniform between some d -vector of values \mathbf{a} and \mathbf{b} . For a proposal, $\boldsymbol{\theta}_j$, to be deemed acceptable, $\mathbf{Y}(\boldsymbol{\theta}_j)$, should be contained exclusively within the interval $[\tilde{y}_t - \Delta_t, \tilde{y}_t + \Delta_t]$ at each time $t = \{1, \dots, n\}$. This so-called “behavioral simulation

space” belongs to the set $\widehat{\Omega}_{(\mathbf{Y})}$ and can be defined as (Keesman 1990)

$$\widehat{\Omega}_{(\mathbf{Y})} = \left\{ \mathbf{Y} \in \mathbb{R}^n : \mathbf{Y} = \mathcal{M}(\boldsymbol{\theta}, \mathbf{x}_0, \tilde{\mathbf{B}}) ; \boldsymbol{\theta} \in \widehat{\Omega}_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})} \right\}, \quad (76)$$

where $\widehat{\Omega}_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}$ constitutes the posterior (behavioral) parameter set

$$\widehat{\Omega}_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})} = \Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}. \quad (77)$$

The conditional parameter set, $\Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}$, is defined as follows

$$\Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})} = \left\{ \boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d : \tilde{\mathbf{Y}} - \mathcal{F}(\boldsymbol{\theta}, \mathbf{x}_0, \tilde{\mathbf{B}}) = \mathbf{E}(\boldsymbol{\theta}); e_t(\boldsymbol{\theta}) \in [-\Delta_t, \Delta_t], t = 1, \dots, n \right\}, \quad (78)$$

and contains solutions, $\boldsymbol{\theta}_j \in \widehat{\Omega}_{(\boldsymbol{\theta} | \mathbf{Y})}$, that satisfy the limits of acceptability of each observation.

If an informative prior distribution is used, then the behavioral (posterior) parameter set is computed as the intersection of the prior parameter set, $\Omega_{(\boldsymbol{\theta})}$, and conditional parameter set

$$\widehat{\Omega}_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})} = \Omega_{(\boldsymbol{\theta})} \cap \Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}. \quad (79)$$

Figure 19 summarizes graphically four different outcomes of the limits of acceptability framework. The behavioral solution space exists, if and only if, the conditional parameter set, $\Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}$, intersects the prior parameter set, $\Omega_{(\boldsymbol{\theta})}$. If an informative prior distribution is used, then a sufficient condition for the posterior (behavioral) parameter set to exist is that the conditional parameter set, $\Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}$, is non-empty.

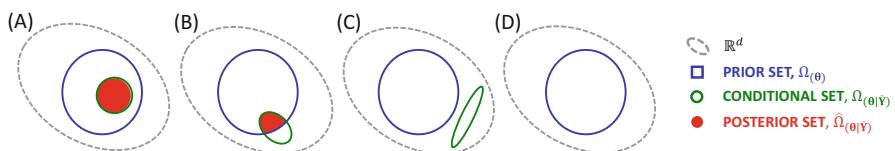


Fig. 19 Set-theoretic approach to quantification of parameter uncertainty. The blue, green, and red colors delineate the prior, $\Omega_{(\boldsymbol{\theta})}$, conditional, $\Omega_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}$, and posterior, $\widehat{\Omega}_{(\boldsymbol{\theta} | \tilde{\mathbf{Y}})}$ parameter set, respectively, whereas the gray ellipsoidal defines the feasible parameter space, $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq \mathbb{R}^d$. The four examples each portray a different outcome: (a) the conditional parameter set intersects fully the prior parameter set, (b) the conditional parameter set intersects only partially the prior parameter set, (c) the conditional and prior parameter set are disjoint (have no elements in common), and (d) the conditional parameter set is empty (no solutions exist that satisfy the limits of acceptability). For the last two examples, there does not exist a behavioral solution space

7.1 The DREAM_(LOA) Algorithm

Application of the limits of acceptability approach requires the availability of a sampling method that can efficiently search the parameter space in pursuit of behavioral solution set, $\widehat{\Omega}_{(\theta|\tilde{Y})}$. Commonly used (population Monte Carlo) rejection sampling methods are rather inefficient in locating behavioral solutions. The chance that a random sample from the prior distribution satisfies the limits of acceptability of each observation is disturbingly small, particularly if the prior parameter space is large compared to the posterior (behavioral) solution space and the number of observations, n is large. Fortunately, we can use MCMC simulation with the DREAM_(LOA) algorithm of Vrugt and Beven (2018) to explore efficiently set-theoretic functions such as Eq. (78).

This selection rule is defined as

$$P_{\text{acc}}(\boldsymbol{\theta}_{(i-1)} \rightarrow \boldsymbol{\theta}_p) = \begin{cases} I(f(\boldsymbol{\theta}_p) \geq f(\boldsymbol{\theta}_{(i-1)})) & \text{if } f(\boldsymbol{\theta}_p) < n \\ 1 & \text{if } f(\boldsymbol{\theta}_p) = n \end{cases}, \quad (80)$$

where $I(a)$ is an indicator function that returns one if the condition a is satisfied and zero otherwise, and the fitness function, $f(\cdot)$, is calculated as follows

$$f(\boldsymbol{\theta}) = \sum_{t=1}^n I(|\tilde{y}_t - y_t(\boldsymbol{\theta})| \leq \Delta_t). \quad (81)$$

If the proposal is accepted, then the Markov chain moves to this new position, $\boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}_p$, otherwise it remains at its current location, that is $\boldsymbol{\theta}_{(i)} = \boldsymbol{\theta}_{(i-1)}$.

The fitness of the proposal, $\boldsymbol{\theta}_p$, is equivalent to the number of observations its simulation satisfies within the limits of acceptability. We accept the proposal, $P_{\text{acc}}(\boldsymbol{\theta}_{(i-1)} \rightarrow \boldsymbol{\theta}_p) = 1$, if the fitness of $\boldsymbol{\theta}_p$ is larger than that of the current state of the chain, $\boldsymbol{\theta}_{(i-1)}$, or if the simulation of the proposal is consistently within $\Delta = \{\Delta_1, \dots, \Delta_n\}$ of the observed values, and thus $f(\boldsymbol{\theta}_p) = n$, otherwise the candidate point is rejected. After a burn-in period in which $f(\cdot) < n$, the convergence of DREAM_(ABC) can be monitored with the \hat{R} diagnostic of Gelman and Rubin (1992). A full description of DREAM_(LOA) appears in Vrugt and Beven (2018) and interested readers are referred to this publication for further details. Appendix C presents a stripped down MATLAB implementation of the DREAM_(LOA) algorithm.

7.2 Vadose Zone Hydrology Revisited

We now revisit the third case study and use instead limits of acceptability for each soil moisture observation. The values of Δ_t ; $t = \{1, \dots, n\}$ in Eq. (81) are derived from Fig. 8 (p. 3055) in Scharnagl et al. (2011) and, thus, match the 95% intervals of the distributed moisture content observations. This equates to an average value of the limits of acceptability of $0.047 \text{ (cm}^3 \text{ cm}^{-3}\text{)}$.

Figure 20 presents histograms of the DREAM_(LOA)-derived marginal posterior distribution of the six HYDRUS-1D model parameters of this study. The bottom panel presents a time series plot of the behavioral simulation set, $\hat{\Omega}_{(\mathbf{Y})}$. The observed soil moisture data are indicated separately with red dots.

The behavioral HYDRUS-1D model nicely tracks the observed soil moisture measurements with behavioral simulation space, $\hat{\Omega}_{(\mathbf{Y})}$, that encapsulates consistently the observed data. The root mean square error (RMSE) of the behavioral (posterior) mean simulation equates to about $0.0149 \text{ cm}^3/\text{cm}^{-3}$, a value somewhat larger than derived separately using the formal Gaussian likelihood function. The behavioral parameter space of most parameters extends a large part of their respective prior ranges with marginal distributions that deviate markedly from normality. The prior ranges are rather narrow and derived from Monte Carlo simulation with the ROSETTA pedotransfer toolbox using textural data (percentages of sand, silt, and clay) as main input variables.

Finally, Fig. 21 shows how the posterior parameter set translates into uncertainty of the soil water retention (left) and unsaturated soil hydraulic conductivity (right) functions. The light gray region corresponds to the prior parameter set, $\Omega_{(\mathbf{Y})}$, whereas the dark gray is used to denote the behavioral (posterior) solution set, $\hat{\Omega}_{(\mathbf{Y})}$. Note, we use here the variable \mathbf{Y} to denote the functional space rather than moisture content values. The posterior mean soil hydraulic functions are indicated with the solid black line. The posterior uncertainty of the soil hydraulic functions appears rather large in response to the observed spatial variability of the soil moisture data. This uncertainty can now be used to simulate the soil moisture variability in the 50×50 experimental plot, simply by drawing soil hydraulic functions from the posterior ranges. Thus, the limits of acceptability framework provides a way to account explicitly for spatial variability.

This concludes our section with case studies. A list with some key applications of the DREAM algorithm can be found in Vrugt (2016, p. 276). This includes studies in chemistry, ecology, geomorphology, physics, structural engineering, to name a few, and involves inference problems that are much more complex than presented herein.

8 Marginal Likelihood and Model Complexity

Thus far in this chapter, we have focused our inferences on a single model but without recourse to the denominator, $P(\tilde{\mathbf{Y}})$, in Eq. (11), the so-called marginal likelihood. This normalizing constant in Bayes' theorem ensures that the posterior distribution, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, integrates to unity. The marginal likelihood, or Bayesian model evidence, is of no particular interest for parameter estimation, yet of imminent importance for hypothesis testing. The hypothesis (model), \mathcal{M}_k , where $k = \{1, \dots, K\}$, with largest evidence, $P(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, is most supported by the available data, $\tilde{\mathbf{Y}}$.

Bayesian model selection encodes a natural preference for simpler and more constrained models. This approach provides a rigorous justification to the parsimony

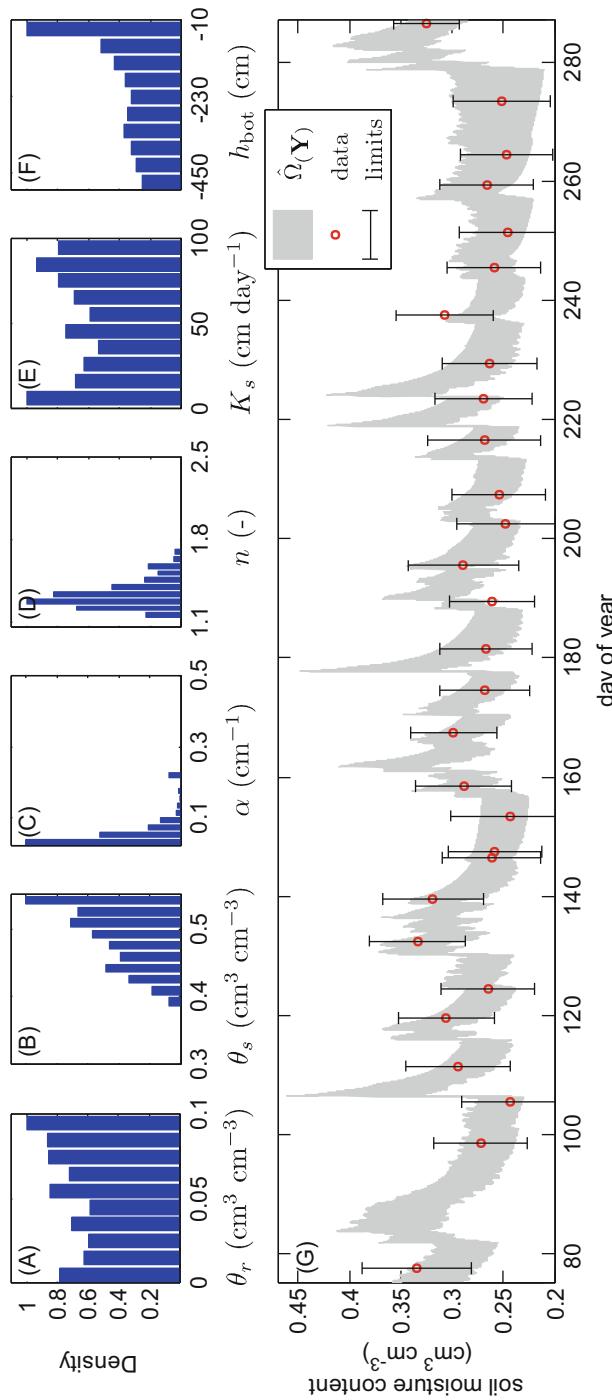


Fig. 20 Vadose zone hydrology. Top panel: histograms of the behavioral set, $\hat{\Omega}_{\theta}(\cdot|\tilde{\mathbf{Y}})$, of the soil hydraulic parameters: (a) θ_r , (b) θ_s , (c) α , (d) n , (e) K_s , and (f) h_{bot} . The x-axis matches exactly the (uniform) prior distribution. Bottom panel: comparison of observed (red dots) and posterior simulated, $\hat{\Omega}_{\mathbf{Y}}(\cdot)$, (gray region) soil moisture contents

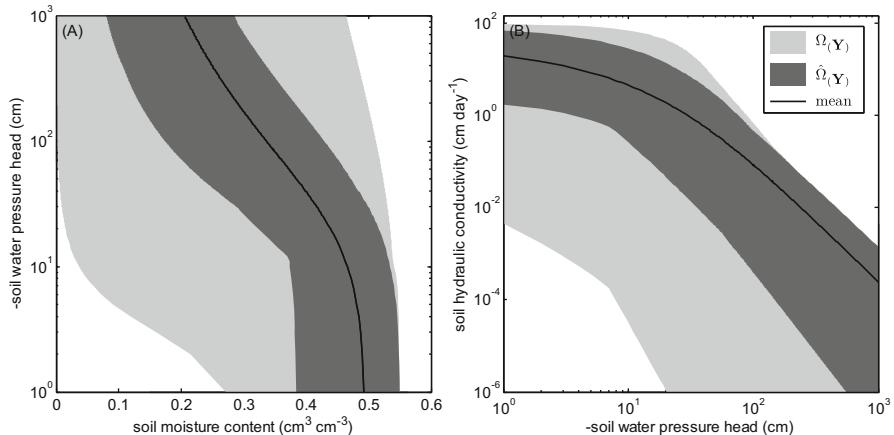


Fig. 21 Vadose zone hydrology: comparison of the prior (dark gray) and posterior (light gray) ranges of the (left) soil water retention, and (right) unsaturated soil hydraulic conductivity function. The posterior (or behavioral) mean hydraulic functions are indicated separately with the solid black line

principle of William of Ockham (1287–1347), an English Franciscan friar, philosopher, and theologian, who stated that “...Entities must not be multiplied beyond necessity.” This principle of parsimony, also known as Occam’s azor, is traceable to the works of philosophers such as Aristotle (384–322 BC), Ptolemy (c. AD 90 – c. AD 168), and consistent with requirements of falsifiability in the scientific method. Indeed, simpler hypotheses (theories) are preferred as they involve fewer assumptions and are therefore easier testable. Thus, a “good” model selection technique must necessarily balance goodness of fit with complexity (number of “free” parameters). Unfortunately, analytical solutions of $P(\tilde{\mathbf{Y}}|\mathcal{M}_k)$ are available only for certain special cases, which are too limiting to be of practical value in environmental modeling (Schöniger et al. 2014). We therefore have to resort to sampling methods which approximate numerically the integral of the posterior distribution.

In the Monte Carlo approach, the marginal likelihood can be approximated by the arithmetic mean of the likelihood function, $L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, of a large sample of points drawn randomly from the prior distribution. This approximation, however, is not particularly efficient, as many of the random samples drawn from the prior parameter distribution will exhibit insufficient density to contribute to the evidence. A more efficient approach constitutes importance sampling (see Algorithm 2). The importance distribution, $G(\boldsymbol{\theta})$, has a known integral of unity and should satisfy that $g(\boldsymbol{\theta}) > 0$ whenever $P(\boldsymbol{\theta}) > 0$, otherwise certain parts of the target distribution, $P(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, are possibly dismissed. The ratio of the density of the unnormalized posterior, $P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})$, and the density, $g(\boldsymbol{\theta})$, of the importance distribution, $G(\boldsymbol{\theta})$,

now details the contribution of some importance sample, $\boldsymbol{\theta}_j$, to the marginal likelihood. The integral of the unnormalized posterior distribution, $P(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, is thus equivalent to the expected value of $P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})/g(\boldsymbol{\theta})$, or $\mathbb{E}[P(\boldsymbol{\theta})L(\boldsymbol{\theta}|\tilde{\mathbf{Y}})/g(\boldsymbol{\theta})]$ and can be approximated numerically

$$P(\tilde{\mathbf{Y}}|\mathcal{M}_k) \simeq \frac{1}{M} \sum_{j=1}^M \frac{P(\boldsymbol{\theta}_j|\mathcal{M}_k)L(\boldsymbol{\theta}_j|\mathcal{M}_k, \tilde{\mathbf{Y}})}{g(\boldsymbol{\theta}_j)} \quad (82)$$

using M different samples $\{\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_M\}$ drawn randomly from the importance distribution, $G(\boldsymbol{\theta})$.

The importance estimator of Eq. (82) is accurate and robust but not without practical problems. Indeed, the efficiency of importance sampling depends critically on the choice of the importance distribution, $G(\boldsymbol{\theta})$. This becomes particularly relevant for CPU-intensive system models and high-dimensional target distributions. To enhance the computational efficiency of the estimator of Eq. (82), Volpi et al. (2017) has introduced a two-step approach in which samples from the target distribution (step 1) are used to construct an adequate importance distribution (step 2). Benchmark experiments on target distributions with variable dimensionality (up to 100), one or two (disconnected) modes, and variably correlated, twisted, and/or truncated dimensions show that this approach, called GAussian Mixture importancE Sampling (GAME), is unbiased, robust, and efficient and can provide accurate estimates of the evidence, $P(\tilde{\mathbf{Y}}|\mathcal{M}_k)$, at a relatively small computational cost outperforming commonly used estimators. The GAME sampler is implemented in the MATLAB package of DREAM (Vrugt 2016) and simplifies considerably scientific inquiry through hypothesis testing and model selection.

We conclude this chapter with Fig. 22 which displays hypothetical relationships between model complexity and the within and out of sample prediction error (left plot) and the maximum and marginal likelihoods (right plot).

The two graphs depict idealized relationships between the entities of interest, henceforth may not always go up in practice. Nevertheless, the conclusions we can draw from both schematics are generally valid and supported by much empirical evidence. First, the larger the complexity of a model, the smaller its within-sample prediction error. Thus, a more complex model should be able to better fit the data. The benefit of using more parameters decreases quite rapidly with model complexity. Second, the out-of-sample residuals (evaluation period) will decrease with increasing model complexity but not indefinitely. Indeed, the prediction error can increase again due to over-parameterization. In other words, the model has become too complex and is no longer supported by the calibration data. Third, the larger the complexity of a model, the larger will be the likelihood maximum (= in sample). There appears to be a strong ceiling effect, however, that is, beyond a certain complexity the likelihood maximum will hardly increase. Fourth, the marginal likelihood displays a characteristic U-shaped curve, much alike the prediction error but then using calibration data only. This underlines the practical significance

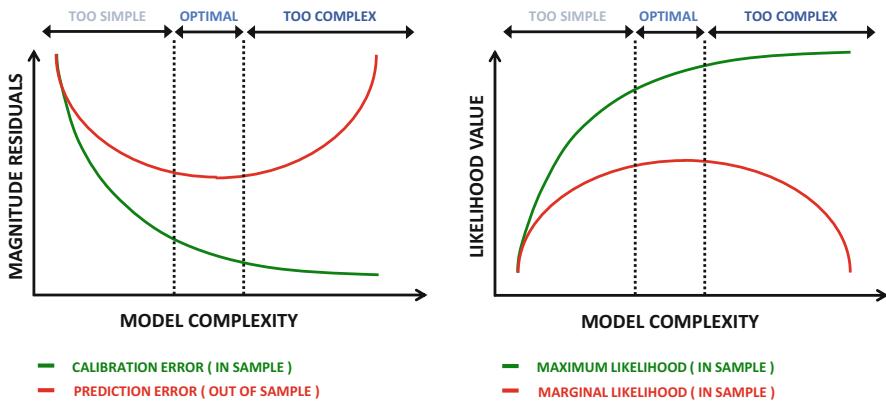


Fig. 22 Schematic overview of the relationship between model complexity (number of parameters) and (on left) the magnitude of the error residuals in the calibration (in blue) and evaluation (in red) period, and (on right) the value of the maximum likelihood, $L(\theta^* | \tilde{Y})$ (in blue), and the marginal likelihood, $P(\tilde{Y})$ (in red) of the calibration data set

of the marginal likelihood. Traditional metrics of fit require use of an independent evaluation period to back out the optimal model complexity (see left plot). The marginal likelihood necessitates only observations in the calibration data set to draw similar conclusions.

9 Conclusion

This chapter has reviewed the main elements of Bayesian inference to reconcile dynamic environmental system models with observations, to facilitate prediction in time (forecasting) and space (interpolation), data assimilation, and inference of the model parameters. The prior distribution, formulation of the likelihood function, and marginal likelihood have been discussed extensively with special emphasis on numerical techniques suited to approximate the posterior distribution. This includes rejection sampling, importance sampling, and recent developments in Markov chain Monte Carlo simulation to sample complex target distributions. We also have highlighted their application to sampling limits of acceptability. Three different case studies with surface and subsurface models were presented to illustrate the application of Bayesian inference to quantification of parameter and model predictive uncertainty. Numerical recipes were provided for each of the numerical techniques to facilitate implementation and use of Bayes analysis.

We close this chapter with a few words on how to treat other sources of uncertainty besides the parameters of the model (with/without nuisance variables). One simple and pragmatic approach is to parameterize each error source separately

and to augment the likelihood functions with these additional unknown or latent variables. Examples of this approach have appeared in the hydrologic literature in the works of Kavetski et al. (2006a, b) and Vrugt et al. (2008).

Acknowledgments The first author is supported by funding from the UC-Lab Fees Research Program Award 237285. The material presented in this chapter is part of the first author's graduate course on "Merging Models and Data" (CEE-290) taught at the University of California, Irvine. An animated presentation of this material can be found online at <https://www.youtube.com/watch?v=bhA9vtiHxZ0>. The DREAM family of algorithms discussed in this chapter are implemented in DREAM Suite, an easy to use, plug-and-play, Windows program. This program can be found online at www.dreamsuite.eu and simplifies considerably Bayesian analysis and its application to uncertainty quantification of mathematical models.

Appendix

A: Derivation of Bayes' Theorem

Bayes' theorem (also referred to as Bayes' law or Bayes' rule) is a relatively simple but fundamental result of probability theory that allows for the calculation of certain conditional probabilities. The theorem specifies the relationship between the probability of two entities, A and D , or $P(A)$ and $P(D)$, and their respective conditional probabilities, $P(A|D)$ and $P(D|A)$. This theorem follows logically from Kolmogorovs (1903–1987) axiomatic definition of probability and is consistent with the frequentist and subjectivist approach to epistemology.

If $P(A) > 0$ and $P(D) > 0$ denote the probability of two different events A and D , then the conditional probability of A given event D is equivalent to

$$P(A|D) = \frac{P(A \cap D)}{P(D)}, \quad (83)$$

where $P(A \cap D)$ signifies the probability of the union of events A and D . Similarly, the conditional probability of D given event A is

$$P(D|A) = \frac{P(D \cap A)}{P(A)}. \quad (84)$$

Per definition, $P(A \cap D) = P(D \cap A)$ which implies that

$$P(A \cap D) = P(D)P(A|D) = P(A)P(D|A), \quad (85)$$

which after simple rearrangement leads to Bayes' theorem in Eq. (11)

$$P(A|D) = \frac{P(A)P(D|A)}{P(D)}, \quad (86)$$

where A signifies the parameter values, and B denotes the data.

The different probabilities in Eq. (86) may have different interpretations, depending on the intended goal of application. Within the context of statistical inference, Bayes' theorem expresses mathematically how a subjective initial degree of belief, $P(a)$, in a proposition a , changes rationally to $P(a|d)$ in response to new data, d . The evidence d is not to be confused with the evidence, $P(d)$, or marginal likelihood. The term $P(a)$ is called the prior distribution, and $P(a|d)$ denotes the posterior probability density function, or the degree of belief having accounted for the evidence d . This subjectivist approach is the cornerstone of Bayesian inference, yet Bayes' theorem has much wider applicability. In our application of Bayes theorem, we can replace the conditional probability, $P(d|a)$ with the data likelihood, $L(a|d)$ as all our inferences are drawn from the residuals of a and d , that is $a - d$. Whether we center the measurement error distribution on the proposition, a (simulated data), or the “evidence,” b (observed data), this does not change the degree of belief in a .

To illustrate the application of Bayes theorem, let us consider a simple thought experiment in which we are trying to infer the probability that someone has a disease, X , given that they have some symptom, S . That the person has this symptom is clearly visible to the eye, but whether they have the decease is not evident. Bayes' law tells us that $P(X|S)$ can be derived from

$$P(X|S) = \frac{P(X)P(S|X)}{P(S)}. \quad (86)$$

So to compute $P(X|S)$, we need to know the prior probability, $P(S)$ and $P(X)$, of the symptom and the disease, respectively (how common are the symptom and disease), and $P(S|X)$, the probability that someone has symptom S given that he/she has the disease, X (via lab tests).

We can confirm Bayes' law with a simple practical example (see Fig. 23) wherein the occurrence of two, presumably correlated, events, *Rain* and *Thunder*, is observed at some place on our planet for a period of $n = 20$ -days. The prior probabilities of both events, $P(R)$ and $P(T)$, and their respective conditional probabilities, $P(R|T)$ and $P(T|R)$, are easy to calculate from the data.

We can now use these (conditional) probabilities to benchmark Bayes' law. Indeed, per Bayes' theorem, $P(R|T) = P(R)P(T|R)/P(T)$, which gives $P(R|T) = (11/20 \times 4/11)/(5/20) = (4/20)/(5/20) = 4/5$. This confirms the value of $P(R|T) = 4/5$ computed directly from the data. The same result is found for $P(T|R)$.

DAY	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
RAIN	0	1	0	0	1	0	1	1	1	0	0	1	1	0	0	1	0	1	1	1
THUNDER	1	1	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0

0: EVENT NOT OBSERVED, 1: EVENT OBSERVED

$$P(R) = \frac{n_R}{n} = \frac{11}{20}$$

$$P(T|R) = \frac{P(T \cap R)}{P(R)} = \frac{\frac{4}{20}}{\frac{11}{20}} = \frac{4}{11} \quad \textcolor{red}{\sim 44\% \text{ CHANCE OF THUNDER WHEN IT RAINS}}$$

$$P(T) = \frac{n_T}{n} = \frac{5}{20}$$

$$P(R|T) = \frac{P(R \cap T)}{P(T)} = \frac{\frac{4}{20}}{\frac{5}{20}} = \frac{4}{5} \quad \textcolor{red}{80\% \text{ CHANCE OF PRECIPITATION WITH THUNDER}}$$

$$P(R \cap T) = \frac{4}{20}$$

Fig. 23 A simple data set of two discrete events, Rain and Thunder. A value of one (zero) on a given day means that the event has (not) been observed. The data can be used to calculate the prior probabilities of R and T , and their conditional probabilities, $P(R|T)$ and $P(T|R)$. These conditional probabilities can be confirmed with Bayes' law

B: MATLAB Code DREAM

The core of the DREAM algorithm can be written in MATLAB in about 30 lines of code (see below). Based on input arguments, `prior` (function handle that draws samples from prior distribution), `f` (function handle that returns target density), `N` (number of chains), `M` (desired number of iterations), `d` (number of parameters) and `problem` (structure that allows user to pass additional input arguments to `f` or `prior`), the DREAM function returns as output to the user the multidimensional array `x` which stores the sampled states and corresponding target densities of each chain (third dimension). Example function handles for some d -variate uniform prior distribution, `prior = @(N,d) unifrnd(-10,10,N,d)`, and standard normal target distribution, `f = @(x,d) mvnpdf(x,zeros(1,d),eye(d))`.

```

function [x] = dream(prior,f,N,M,d,problem)
% Differential Evolution Adaptive Metropolis (DREAM) algorithm

[delta,c,c_star,r,p_g] = deal(3,0.1,1e-6,3,0.2);
x = nan(M,d+1,N);
for j = 1:N, R(j,1:N-1) = setdiff(1:N,j); end
eta = (1:r)/r; p_eta = ones(1,r)/r;

X = prior(N,d);
for j = 1:N, x(1,1:d+1,j) = [X(j,1:d) f(X(j,1:d))]; end
% Create initial population
% Store initial states and density

for i = 2:M, % Dynamic part: Evolution of N chains
    [~,draw] = sort(rand(N-1,1));
    dx = zeros(N,d);
    lambda = unifrnd(-c,C,N,1);
    for j = 1:N,
        D = randsample(1:delta,1,'true');
        a = R(j,draw(1:D,j)); b = R(j,draw(D+1:2*D,j));
        id = randsample(1:r,1,'true',p_eta);
        z = rand(1,d);
        A = find(z < eta(id));
        d_star = numel(A);
        if d_star == 0, [~,A] = min(z); d_star = 1; end
        gamma_d = 2.38/sqrt(2*d_star);
        g = randsample([gamma_d 1],1,'true',[1-p_g p_g]);
        dx(j,A) = c_star*randn(1,d_star) + ...
            (1+lambda(j))*g*sum(x(i-1,A,a)-x(i-1,A,b),3);
        x_p(j,1:d) = x(j,1:d) + dx(j,1:d);
        f_xp(j,1) = f(xp(j,1:d));
        p_acc = min(1,f_xp(j,1)./x(i-1,d+1,j));
        if p_acc > rand
            x(i,1:d+1,j) = [xp(j,1:d) f_xp(j,1)];
        else
            x(i,1:d+1,j) = x(i-1,1:d+1,j);
        end
    end
    [x(1:i,1:d+1,1:N)] = check(x(1:i,1:d+1,1:N));
    % Patch outlier chains
end
% End dynamic part

```

The variables in the DREAM function are chosen carefully to match, insofar possible, their symbols used in the main text. Built-in functions are highlighted with a *low dash* and information about their respective input and output arguments is provided by the MATLAB “help” utility. The jump vector, $dx(j, 1:d)$, of the j th chain contains the desired information about the scale and orientation of the proposal distribution and is derived from the remaining $N-1$ chains. The function `check()` is used as a patch for outlier chains, a critical vulnerability of multi-chain MCMC methods such as SCEM-UA, DE-MC, and DREAM (Vrugt et al. 2003; ter Braak and Vrugt 2008; Vrugt et al. 2008, 2009). Note, vectorization of the inner (proposal) loop would enhance significantly computational efficiency, yet affects negatively readability.

The MATLAB code of DREAM presented in this Appendix has several important restrictions (e.g., uniform prior, fixed crossover selection probabilities, convergence is not monitored, lack of built-in likelihood functions) – all of which are addressed in the MATLAB toolbox of DREAM developed by Vrugt (2016). This toolbox also allows the user to evaluate the chains in parallel using the distributed computing toolbox of MATLAB (see Vrugt (2016) for a multicore DREAM implementation). This parallel implementation violates detailed balance of the sampled chain

trajectories, nonetheless, benchmark experiments on a diverse set of problems have shown that this violation hardly affects the results.

C: MATLAB Code DREAM_(LOA)

Basic MATLAB implementation of the DREAM_(LOA) algorithm for limits of acceptability sampling. This code is identical to DREAM in Appendix B except for lines 29–31 which accommodate the revised acceptance rule of Eq. (80). Input argument f is an anonymous function handle that returns the value of the fitness in Eq. (81), and the structure problem allows the user to pass to f the observed data and corresponding limits of acceptability.

```

function [x] = dream_loa(prior,f,N,M,d,problem)
% DiffeRential Evolution Adaptive Metropolis (DREAM) ABC algorithm

[delta,c,c_star,r,p_g] = deal(3,0.1,1e-6,3,0.2); % Default of algorithmic parameters
x = nan(M,d+1,N); % Preallocate chains and density
for j = 1:N, R(j,1:N-1) = setdiff(1:N,j); end % R-matrix: index of chains for DE
eta = (1:r)/r; p_eta = ones(1,r)/r; % Crossover values and select. prob.

X = prior(N,d); % Create initial population
for j = 1:N, x(1,1:d+1,j) = [X(j,1:d) f(X(j,1:d))]; end % Store initial states and density

for i = 2:M, % Dynamic part: Evolution of N chains
    [~,draw] = sort(rand(N-1,N)); % Permute [1,...,N-1] N times
    dx = zeros(N,d); % Set N jump vectors to zero
    lambda = unifrnd(-c,c,N,1); % Draw N lambda values
    for j = 1:N,
        D = randsample(1:delta,1,'true'); % Create proposals + accept/reject
        a = R(j,draw(1:D,j)); b = R(j,draw(D+1:2*D,j)); % Select delta (equal probability)
        id = randsample(1:r,1,'true',p_eta); % Extract vectors a and b unequal j
        z = rand(1,d); % Select index of crossover value
        A = find(z < eta(id)); % Draw d values from U[0,1]
        d_star = numel(A); % Subset A dimensions to update
        if d_star == 0, [~,A] = min(z); d_star = 1; end % How many dimensions sampled?
        gamma_d = 2.38/sqrt(2*D+d_star); % A must contain one dimension
        g = randsample([gamma_d 1],1,'true',[1-p_g p_g]); % Calculate jump rate
        dx(j,A) = c_star*randn(1,d_star) + ... % Select gamma: 80/20 mix [def: 1]
        (1+lambda(j))*g*sum(x(i-1,A,a)-x(i-1,A,b),3); % Compute jth jump using DE
        x_p(j,1:d) = x(j,1:d) + dx(j,1:d); % Compute jth proposal
        f_xp(j,1) = f(xp(j,1:d),problem); % Calculate density jth proposal
        p_acc = f_xp(j,1) >= x(i-1,d+1,j); % Compute acceptance probability
        if p_acc % p_acc equal to 1?
            x(i,1:d+1,j) = [xp(j,1:d) f_xp(j,1)]; % True: Accept proposal
        else
            x(i,1:d+1,j) = x(i-1,1:d+1,j); % False: Maintain "old" position
        end
    end
    [x(1:i,1:d+1,1:N)] = check(x(1:i,1:d+1,1:N)); % Patch outlier chains
end % End dynamic part

```

The variables in the DREAM_(LOA) function are chosen carefully to match, insofar possible, their symbols used in the main text. Built-in functions are highlighted with a *low dash*. Detailed documentation on each of these functions can be found in the MATLAB “help” system.

The fitness has to be defined as an anonymous function handle as follows, $f = @(x,problem) \text{fitness}(x,problem)$, wherein problem is a structure with fields Yobs and Delta that store in a n -vector the measurements and limits of acceptability, respectively. A template fitness function, f, is given below.

```
function f = fitness(x,problem)
% This function computes the fitness for d-vector of parameter values, x

Ysim = own_model_script(x);
f = sum(abs(problem.Yobs - Ysim)) <= problem.Delta; % Run model and return simulation
% Compute fitness per Equation (81)
```

The fitness function includes a call to `own_model_script`, which executes the forward model and returns simulated values for a given d -vector of parameter values, x . This function should be written by the user.

References

- B.C. Bates, E.P. Campbell, A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. *Water Resour. Res.* **37**(4), 937–947 (2001)
- T. Bayes, R. Price, An essay towards solving a problem in the doctrine of chance. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, A.M.F.R.S. *Philos. Trans. R. Soc. Lond.* **53**(0), 370–418 (1763). <https://doi.org/10.1098/rstl.1763.0053>
- J.O. Berger, *Statistical Decision Theory and Bayesian Analysis* (Springer, New York, 1985)
- J.O. Berger, J.M. Bernardo, D. Sun, The formal definition of reference priors. *Ann. Stat.* **37**(2), 905–938 (2009). <https://doi.org/10.1214/07-AOS587>
- J.M. Bernardo, Reference posterior distributions for Bayesian inference (with discussion). *J. R. Stat. Soc. Ser. B* **41**, 113–147 (1979)
- K. Beven, A manifesto for the equifinality thesis. *J. Hydrol.* **320**(1), 18–36 (2006)
- K.J. Beven, A.M. Binley, The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.* **6**, 279–298 (1992)
- K.J. Beven, A.M. Binley, GLUE: 20 years on. *Hydrol. Process.* **28**, 5879–5918 (2014). <https://doi.org/10.1002/hyp.10082>
- S. Blazkova, K.J. Beven, A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. *Water Resour. Res.* **45**, W00B16 (2009). <https://doi.org/10.1029/2007WR006726>
- G.E.P. Box, G.C. Tiao, *Bayesian Inference in Statistical Analysis* (Wiley, New York, 1992), 588 pp
- S.P. Brooks, A. Gelman, General methods for monitoring convergence of iterative simulations. *J. Comput. Graph. Stat.* **7**, 434–455 (1998)
- M. Clark, D. Kavetski, F. Fenicia, Pursuing the method of multiple working hypotheses for hydrological modeling. *Water Resour. Res.* **47**(9), 1–16 (2011). <https://doi.org/10.1029/2010WR009827>
- S. Dean, J.E. Freer, K.J. Beven, A.J. Wade, D. Butterfield, Uncertainty assessment of a process-based integrated catchment model of phosphorus (INCA-P). *Stoch. Env. Res. Risk A.* **23**, 991–1010 (2009). <https://doi.org/10.1007/s00477-008-0273-z>
- Q. Duan, S. Sorooshian, V. Gupta, Effective and efficient global optimization for conceptual rainfall-runoff models. *Water Resour. Res.* **28**(4), 1015–1031 (1992)
- G. Evin, D. Kavetski, M. Thyre, G. Kuczera, Pitfalls and improvements in the joint inference of heteroscedasticity and autocorrelation in hydrological model calibration. *Water Resour. Res.* **49**, 4518–4524 (2013). <https://doi.org/10.1002/wrcr.20284>
- C. Fernandez, M.J.F. Steel, On Bayesian modeling of fat tails and skewness. *J. Am. Stat. Assoc.* **93**, 359–371 (1998)

- J. Freer, H. McMillan, J.J. McDonnell, K.J. Beven, Constraining dynamic TOPMODEL responses for imprecise water table information using fuzzy rule based performance measures. *J. Hydrol.* **291**, 254–277 (2004)
- A.G. Gelman, D.B. Rubin, Inference from iterative simulation using multiple sequences. *Stat. Sci.* **7**, 457–472 (1992)
- A.G. Gelman, G.O. Roberts, W.R. Gilks, *Bayesian Statistics* (Oxford University Press, Oxford, 1996), pp. 599–608
- J. Geweke, Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments, in *Bayesian Statistics 4*, ed. by J.M. Bernardo, J.O. Berger, A.P. Dawid, A.F.M. Smith (Oxford Oxford University Press, 1992), pp. 169–193
- W.R. Gilks, G.O. Roberts, Strategies for improving MCMC, in *Markov Chain Monte Carlo in Practice*, ed. by W.R. Gilks, S. Richardson, D.J. Spiegelhalter (Chapman & Hall, London, 1996), pp. 89–114
- W.R. Gilks, G.O. Roberts, E.I. George, Adaptive direction sampling. *Underst. Stat.* **43**, 179–189 (1994)
- H.V. Gupta, T. Wagener, Y. Liu, Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. *Hydrol. Process.* **22**(18), 3802–3813 (2008)
- H. Haario, E. Saksman, J. Tamminen, Adaptive proposal distribution for random walk Metropolis algorithm. *Comput. Stat.* **14**, 375–395 (1999)
- H. Haario, E. Saksman, J. Tamminen, An adaptive Metropolis algorithm. *Bernoulli* **7**, 223–242 (2001)
- H. Haario, E. Saksman, J. Tamminen, Componentwise adaptation for high dimensional MCMC. *Stat. Comput.* **20**, 265–274 (2005)
- H. Haario, M. Laine, A. Mira, E. Saksman, DRAM: Efficient adaptive MCMC. *Stat. Comput.* **16**, 339–354 (2006)
- H. Hastings, Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57**, 97–109 (1970)
- T.J. Heimovaara, W. Bouten, A computer-controlled 36-channel time domain reflectometry system for monitoring soil water contents. *Water Resour. Res.* **26**, 2311–2316 (1990). <https://doi.org/10.1029/WR026i010p02311>
- J. Hoeting, D. Madigan, A. Raftery, C. Volinsky, Bayesian model averaging: A tutorial. *Stat. Sci.* **14**(4), 382–417 (1999)
- H. Jeffreys, An invariant form for the prior probability in estimation problems. *Proc. R. Soc. Lond. A Math. Phys. Sci.* **186**(1007), 453–461 (1946). <https://doi.org/10.1098/rspa.1946.0056>
- D. Kavetski, G. Kuczera, S.W. Franks, Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. *Water Resour. Res.* **42**(3), W03407 (2006a). <https://doi.org/10.1029/2005WR004368>
- D. Kavetski, G. Kuczera, S.W. Franks, Bayesian analysis of input uncertainty in hydrological modeling: 2. Application. *Water Resour. Res.* **42**(3), W03408 (2006b). <https://doi.org/10.1029/2005WR004376>
- K. Keesman, Membership-set estimation using random scanning and principal component analysis. *Math. Comput. Simul.* **32**, 535–543 (1990)
- T. Krueger, J.N. Quinton, J. Freer, C.J. Macleod, G.S. Bilotta, R.E. Brazier, P.M. Haygarth, Uncertainties in data and models to describe event dynamics of agricultural sediment and phosphorus transfer. *J. Environ. Qual.* **38**(3), 1137–1148 (2009)
- G. Kuczera, Improved parameter inference in catchment models, 1. Evaluating parameter uncertainty. *Water Resour. Res.* **19**(5), 1151–1162 (1983). <https://doi.org/10.1029/WR019i005p01151>
- G. Kuczera, D. Kavetski, S. Franks, M. Thyer, Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. *J. Hydrol.* **331**(1), 161–177 (2006)
- E. Laloy, J.A. Vrugt, High-dimensional posterior exploration of hydrologic models using multiple-try DREAM_(ZS) and high-performance computing. *Water Resour. Res.* **48**, W01526 (2012). <https://doi.org/10.1029/2011WR010608>
- J.S. Liu, F. Liang, W.H. Wong, The multiple-try method and local optimization in metropolis sampling. *J. Am. Stat. Assoc.* **95**(449), 121–134 (2000). <https://doi.org/10.2307/2669532>

- Y. Liu, J.E. Freer, K.J. Beven, P. Matgen, Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. *J. Hydrol.* **367**, 93–103 (2009). <https://doi.org/10.1016/j.jhydrol.2009.01.016>
- H. McMillan, J. Freer, F. Pappenberger, T. Krueger, M. Clark, Impacts of uncertain river flow data on rainfall-runoff model calibration and discharge predictions. *Hydrol. Process.* **24**(10), 1270–1284 (2010)
- N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, E. Teller, Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087–1092 (1953)
- J.E. Nash, A unit hydrograph study with particular reference to British catchments. *Proc. Inst. Civ. Eng.* **17**, 249–282 (1960)
- J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models part I – A discussion of principles. *J. Hydrol.* **10**(3), 282–290 (1970)
- T. Page, K.J. Beven, J. Freer, A. Jenkins, Investigating the uncertainty in predicting responses to atmospheric deposition using the model of acidification of groundwater in catchments (MAGIC) within a generalised likelihood uncertainty estimation (GLUE) framework. *Water Soil Air Pollut.* **142**, 71–94 (2003)
- T. Page, K.J. Beven, D. Whyatt, Predictive capability in estimating changes in water quality: Long-term responses to atmospheric deposition. *Water Soil Air Pollut.* **151**, 215–244 (2004)
- T. Page, K.J. Beven, J. Freer, Modelling the chloride signal at the Plynlimon catchments, Wales using a modified dynamic TOPMODEL. *Hydrol. Process.* **21**, 292–307 (2007)
- F. Pappenberger, K. Beven, M. Horritt, S. Blazkova, Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations. *J. Hydrol.* **302**, 46–69 (2005)
- F. Pappenberger, K. Frodsham, K.J. Beven, R. Romanovicz, P. Matgen, Fuzzy set approach to calibrating distributed flood inundation models using remote sensing observations. *Hydrol. Earth Syst. Sci.* **11**(2), 739–752 (2007)
- K.V. Price, R.M. Storn, J.A. Lampinen, *Differential Evolution, A Practical Approach to Global Optimization* (Springer, Berlin, 2005)
- V.C. Radu, J. Rosenthal, C. Yang, Learn from the thy neighbor: Parallel-chain and regional adaptive MCMC. *J. Am. Stat. Assoc.* **104**(488), 1454–1466 (2009)
- A.E. Raftery, S.M. Lewis, One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Stat. Sci.* **7**, 493–497 (1992)
- A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174 (2005)
- P. Reichert, J. Mieleitner, Analyzing input and structural uncertainty of nonlinear dynamic models with stochastic, time-dependent parameters. *Water Resour. Res.* **45**, W10402 (2009). <https://doi.org/10.1029/2009WR007814>
- B. Renard, D. Kavetski, E. Leblois, M. Thyre, G. Kuczera, S.W. Franks, Toward a reliable decomposition of predictive uncertainty in hydrological modeling: Characterizing rainfall errors using conditional simulation. *Water Resour. Res.* **47**(11), W11516 (2011). <https://doi.org/10.1029/2011WR010643>
- C.P. Roberts, G. Casella, *Monte Carlo Statistical Methods*, 2nd edn. (Springer, New York, 2004)
- G.O. Roberts, W.R. Gilks, Convergence of adaptive direction sampling. *J. Multivar. Anal.* **49**, 287–298 (1994)
- G.O. Roberts, J.S. Rosenthal, Coupling and ergodicity of adaptive Markov chain Monte Carlo algorithms. *J. Appl. Probab.* **44**, 458–475 (2007)
- G.O. Roberts, A. Gelman, W.R. Gilks, Weak convergence and optimal scaling of random walk Metropolis algorithms. *Ann. Appl. Probab.* **7**, 110–120 (1997)
- M. Sadegh, J.A. Vrugt, Approximate Bayesian computation using Markov chain Monte Carlo simulation: DREAM_(ABC). *Water Resour. Res.* **50** (2014). <https://doi.org/10.1002/2014WR015386>
- M. Sadegh, J.A. Vrugt, C. Xu, E. Volpi, The stationarity paradigm revisited: Hypothesis testing using diagnostics, summary metrics, and DREAM_(ABC). *Water Resour. Res.* **51**, 9207–9231 (2015). <https://doi.org/10.1002/2014WR016805>

- M.G. Schaap, F.J. Leij, M.T. van Genuchten, Neural network analysis for hierarchical prediction of soil water retention and saturated hydraulic conductivity. *Soil Sci. Soc. Am. J.* **62**, 847–855 (1998)
- M.G. Schaap, F.J. Leij, M.T. van Genuchten, Rosetta: A computer program for estimating soil hydraulic parameters with hierarchical pedotransfer functions. *J. Hydrol.* **251**, 163–176 (2001)
- B. Scharnagl, J.A. Vrugt, H. Vereecken, M. Herbst, Bayesian inverse modeling of soil water dynamics at the field scale: Using prior information about the soil hydraulic properties. *Hydrol. Earth Syst. Sci.* **15**, 3043–3059 (2011). <https://doi.org/10.5194/hess-15-3043-2011>
- B. Scharnagl, S.C. Iden, W. Durner, H. Vereecken, M. Herbst, Inverse modelling of in situ soil water dynamics: Accounting for heteroscedastic, autocorrelated, and non-Gaussian distributed residuals. *Hydrol. Earth Syst. Sci. Discuss.* **12**, 2155–2199 (2015)
- A. Schöniger, T. Wöhling, L. Samaniego, W. Nowak, Model selection on solid ground: Rigorous comparison of nine ways to evaluate Bayesian model evidence. *Water Resour. Res.* **50**(12), W10530, 9484–9513 (2014). <https://doi.org/10.1002/2014WR016062>
- G. Schoups, J.A. Vrugt, A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic and non-Gaussian errors. *Water Resour. Res.* **46**, W10531 (2010). <https://doi.org/10.1029/2009WR008933>
- J. Šimunek, M. Šejna, H. Saito, M. Sakai, M.T. van Genuchten, *The HYDRUS-1D Software Package for Simulating the One-Dimensional Movement of Water, Heat and Multiple Solutes in Variably-Saturated Media (Version 4.0)* (Department of Environmental Sciences, University of California Riverside, Riverside, 2008)
- T. Smith, A. Sharma, L. Marshall, R. Mehrotra, S. Sisson, Development of a formal likelihood function for improved Bayesian inference of ephemeral catchments. *Water Resour. Res.* **46**, W12551 (2010). <https://doi.org/10.1029/2010WR009514>
- S. Sorooshian, J.A. Dracup, Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases. *Water Resour. Res.* **16**(2), 430–442 (1980)
- S.M. Stigler, Who discovered Bayes's theorem? *Am. Stat.* **37**(4 Part 1), 290–296 (1983)
- R. Storn, K. Price, A simple and efficient heuristic for global optimization over continuous spaces. *J. Glob. Optim.* **11**, 341–359 (1997)
- C.J.F. ter Braak, A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces. *Stat. Comput.* **16**, 239–249 (2006)
- C.J.F. ter Braak, J.A. Vrugt, Differential evolution Markov chain with snooker updater and fewer chains. *Stat. Comput.* **18**(4), 435–446 (2008). <https://doi.org/10.1007/s11222-008-9104-9>
- M. Thiemann, M. Trosset, H. Gupta, S. Sorooshian, Bayesian recursive parameter estimation for hydrologic models. *Water Resour. Res.* **37**(10), 2521–2535 (2001)
- G.C. Topp, J.L. Davis, A.P. Annan, Electromagnetic determination of soil water content: Measurements in coaxial transmission lines. *Water Resour. Res.* **16**, 574–582 (1980). <https://doi.org/10.1029/WR016i003p00574>
- M.T. van Genuchten, A closed-form equation for predicting the hydraulic conductivity of unsaturated soils. *Soil Sci. Soc. Am. J.* **44**(5), 892–898 (1980). <https://doi.org/10.2136/sssaj1980.03615995004400050002x>
- E. Volpi, G. Schoups, G. Firmani, J.A. Vrugt, Sworn testimony of the model evidence: Gaussian Mixture Importance (GAME) sampling. *Water Resour. Res.* **53**, 6133–6158 (2017). <https://doi.org/10.1002/2016WR020167>
- J.A. Vrugt, Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environ. Model. Softw.* **75**, 273–316 (2016). <https://doi.org/10.1016/j.envsoft.2015.08.013>
- J.A. Vrugt, K.J. Beven, Embracing equifinality with efficiency: Limits of acceptability sampling using the DREAM_(LOA) algorithm. *J. Hydrol.* **559**, 954–971 (2018). <https://doi.org/10.1016/j.jhydrol.2018.02.026>, In Press
- J.A. Vrugt, E. Laloy, Reply to comment by Chu et al. on High-dimensional posterior exploration of hydrologic models using multiple-try DREAM_{ext}(ZS) and high-performance computing. *Water Resour. Res.* **50**, 2781–2786 (2014). <https://doi.org/10.1002/2013WR014425>

- J.A. Vrugt, B.A. Robinson, Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resour. Res.* **43**, W01411 (2007). <https://doi.org/10.1029/2005WR004838>
- J.A. Vrugt, M. Sadegh, Toward diagnostic model calibration and evaluation: Approximate Bayesian computation. *Water Resour. Res.* **49** (2013). <https://doi.org/10.1002/wrcr.20354>
- J.A. Vrugt, C.J.F. ter Braak, DREAM_(D): An adaptive Markov chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems. *Hydrol. Earth Syst. Sci.* **15**, 3701–3713 (2011). <https://doi.org/10.5194/hess-15-3701-2011>
- J.A. Vrugt, H.V. Gupta, W. Bouten, S. Sorooshian, A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resour. Res.* **39**(8), 1201 (2003). <https://doi.org/10.1029/2002WR001642>
- J.A. Vrugt, C.G.H. Diks, W. Bouten, H.V. Gupta, J.M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resour. Res.* **41**(1), W01017 (2005). <https://doi.org/10.1029/2004WR003059>
- J.A. Vrugt, C.J.F. ter Braak, M.P. Clark, J.M. Hyman, B.A. Robinson, Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.* **44**, W00B09 (2008). <https://doi.org/10.1029/2007WR006720>
- J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, D. Higdon, B.A. Robinson, J.M. Hyman, Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. *Int. J. Nonlinear Sci. Numer. Simul.* **10**(3), 273–290 (2009)
- L. Wasserman, Bayesian model selection and model averaging. *J. Math. Psychol.* **44**(1), 92–107 (2000). <https://doi.org/10.1006/jmps.1999.1278>
- I.K. Westerberg, J.-L. Guerrero, P.M. Younger, K.J. Beven, J. Seibert, S. Halldin, J.E. Freer, C.-Y. Xu, Calibration of hydrological models using flow-duration curves. *Hydrol. Earth Syst. Sci.* **15**, 2205–2227 (2011). <https://doi.org/10.5194/hess-15-2205-2011>
- J. Yang, P. Reichert, K.C. Abbaspour, Bayesian uncertainty analysis in distributed hydrologic modeling: A case study in the Thur River basin (Switzerland). *Water Resour. Res.* **43**, W10401 (2007). <https://doi.org/10.1029/2006WR005497>
- M. Ye, P. Meyer, S.P. Neuman, On model selection criteria in multimodel analysis. *Water Resour. Res.* **44**, 1–12 (2008). <https://doi.org/10.1029/2008WR006803>
- S.L. Zabell, The rule of succession. *Erkenntnis* **31**(2–3), 283–321 (1989)



Sensitivity Analysis Methods

Yanjun Gan and Qingyun Duan

Contents

1	Introduction	638
2	Methodologies and Applications	639
2.1	Gradient-Based Methods	640
2.2	Variance-Based Methods	646
2.3	Regression-Based Methods	654
3	Which SA Methods to Use?	663
4	Summary	666
	References	666

Abstract

Sensitivity analysis (SA) is an important tool for assessing and reducing uncertainties in computer-based models. This chapter presents a comprehensive review of some commonly used SA methods, including gradient-based, variance-based, and regression-based methods. Features and applicability of those methods are described and illustrated with some examples. Merits and limitations of different methods are explained, and the criteria of choosing appropriate SA methods for different applications are suggested.

Y. Gan (✉)

State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, Beijing, China

e-mail: yjgan@cma.gov.cn

Q. Duan

Faculty of Geographical Science, Beijing Normal University, Beijing, China

e-mail: qyduan@bnu.edu.cn

Keywords

Uncertainty quantification · Sensitivity analysis · Uncertainty analysis · Parameter estimation · Design of experiment · Sampling · Parameter screening · Variance decomposition

1 Introduction

Computer-based models are used to predict real-world processes and are important tools for facilitating understanding of complex, real-world phenomena or solving challenging engineering design problems. They have become indispensable in many fields of science and engineering, from finance to life sciences, from quantum physics to earth sciences and environmental engineering (Gan et al. 2014). Different sources of uncertainties, such as forcing data, observational data, and model structure and parameters, exert great influences on model performance (Renard et al. 2010; Walker et al. 2003). Consequently, output uncertainties should be assessed and attributed to different sources to increase our understanding and confidence of the model based predictions.

Uncertainty analysis (UA) and sensitivity analysis (SA) are two of the fundamental steps in assessing and reducing model uncertainties, a discipline known as “uncertainty quantification (UQ),” with the former a forward propagation, whereas the latter an inverse assessment of model uncertainties (Jakeman et al. 2006). UA focuses on quantifying uncertainty in model outputs, while SA refers to the study of how the uncertainty in the model outputs can be apportioned to different input uncertainty sources (Cariboni et al. 2007; Saltelli et al. 2008). In general, output uncertainty can be numerically represented by statistical measures such as means, standard deviations, skewness, kurtosis, and confidence intervals or pictorially described by probability density functions, cumulative distribution functions, and box plots. Although UA should be run in tandem with SA, it is beyond our scope to give a full review of the UA methods, and hence we refer the readers to Uusitalo et al. (2015) for an overview of UA methods. Yet most commonly, SA is focused on evaluating the influences of model parameters (also called experimental factors) on model outputs (also called responses) (Rakovec et al. 2014). This is because parameters govern many aspects of a model and are of great uncertainty, even if the model structure is correct and the data errors are assumed negligible. However, focusing on SA of model parameters does not mean, for example, the model structure or data are not important, but should help understand the roles played by them.

SA has been employed by many researchers to evaluate the influence of each parameter on model performance and then screen out insensitive parameters from analyses (Borgonovo et al. 2012; Campolongo et al. 2007). It yields key insights into model parameter behaviors and would help reduce parameter dimensionality for subsequent analyses such as parameter estimation (PE), a process for calibrating model simulations to historical observations by tuning influential parameters (Duan et al. 2006). On the other hand, a full understanding of parameter behaviors would facilitate model verification and validation throughout the course of model

development and refinement (Frey and Patil 2002; Sieber and Uhlenbrook 2005). In addition, SA can identify critical regions of parameter space that can aid in model calibration.

Numerous SA approaches have been developed over the years, and they can be classified in a variety of ways. Usually, they are categorized into two groups as local and global methods according to their action ranges (Saltelli et al. 2008). Local SA methods explore the changes of model response by varying one parameter at a time while keeping other parameters constant, using partial derivatives or finite differences at a fixed parameter location as the measure of parametric sensitivity. Though simple and intuitive, local SA methods measure only local sensitivity whose value is obviously location dependent. Consequently, they are applicable only for linear and monotonic problems. On the other hand, global SA methods examine the changes of model response by varying all or a subset of the parameters simultaneously over the entire parameter space, allowing them to provide robust measures in the presence of nonlinearity and interactions among the parameters (Wainwright et al. 2014). Other classifications include methodological categories as mathematical, statistical, and graphical methods by Frey and Patil (2002) and capability categories as qualitative and quantitative methods by Saltelli et al. (1999). Qualitative SA methods aim to screen out a subset of non-influential parameters using a small number of model evaluations, whereas quantitative SA methods aim to measure each parameter's contribution to the response variance, a process that requires a large number of model evaluations (Campolongo et al. 2011; Cariboni et al. 2007). There are many different software packages which include a variety of different SA methods. An excellent review of available software packages that could be adopted for SA, as well as UA and PE, is available in Matott et al. (2009) and Wang et al. (2016).

The selection of the appropriate SA methods for specific problem is not a trivial issue in practice but a potentially tricky task to those who have a minimal amount of experience in mathematical and statistical theories. We later review a series of commonly used SA methods by applying them to a few illustrative examples and discuss their strengths and limitations. The objectives are to present a systematic introduction and illustrative application of those methods, as well as to provide guidance on choosing the appropriate techniques for specific applications.

The remainder of this chapter is arranged as follows: Methodologies of some commonly used SA methods are presented in Sect. 2. An overview of literature on comparison of different SA methods is given in Sect. 3. Finally, we discuss the criteria for selecting appropriate SA methods for specific applications in Sect. 4.

2 Methodologies and Applications

A comprehensive review of different SA methods is given in this section. The SA methods are divided into three categories according to their mathematical approaches used to compute the sensitivity indices: (1) gradient-based, (2) variance-based,

and (3) regression-based. The features of different SA categories are summarized in Table 1. The methodologies of those SA methods are presented below with some illustrative examples.

2.1 Gradient-Based Methods

Gradient-based methods compute the sensitivity indices based on the change in response gradient to the variation of an input factor. If the response gradient induced by a varying factor is larger than that of other factors, it indicates that the varying factor is more sensitive than the other factors. There are numerous gradient-based SA methods. Some of them are reviewed below.

2.1.1 One-At-a-Time

The one-at-a-time (OAT) method (Daniel 1958) is perhaps the most fundamental and intuitive SA method, which assesses parameter sensitivity by sequentially perturbing one parameter at a time while keeping the other parameters at their baseline values. A schematic diagram of three-parameter OAT design is given in Fig. 1.

Assume that we have an n -dimensional parameter space, sensitivity index of the OAT method for parameter X_i ($i = 1, 2, \dots, n$) is

$$S_i = \frac{Y(X_1, \dots, X_{i-1}, X_i \pm \Delta X_i, X_{i+1}, \dots, X_n) - Y(\mathbf{X})}{\Delta X_i} \quad (1)$$

where ΔX_i is the increment for the i th parameter. The OAT method requires only $n + 1$ experiments for an n -dimensional problem, i.e., an experiment for the base point plus n experiments for small perturbation in each of the n parameters.

The OAT method is also known as local SA method since it explores only a local space around the base point. However, the OAT method has been extensively used because it is easy to implement, computationally inexpensive, and useful to provide a glimpse at the model behavior (Saltelli 1999). By reviewing 33 SA-related papers published in *Science* between 1997 and 2003, Saltelli et al. (2006) found that the OAT method has been improperly applied by many researchers, even though this method is actually only justified for linear models.

2.1.2 Fractional Factorial Screening

The fractional factorial (FF) screening makes use of FF sampling (Box and Hunter 1961a, b) to design a small number of experiments for estimating parameter sensitivity. Assume that a model has n parameters with each of them having p levels. A full factorial design would require p^n experiments, while a FF design needs only $1/p^k$ fraction of the experiments of the full factorial design (i.e., p^{n-k} experiments), where k is the number of generators. For example, a 2^{5-2} FF design is 1/4 of a two-level five-parameter full factorial design (Table 2). If we denote the two levels of each parameter as “–” (the low level) and “+” (the high level), the 2^{5-2} FF design

Table 1 Features of different sensitivity analysis methods

Category	Method	Property	Model independent effect	Elementary effect	Main effect	Second-order interaction effect	Higher-order interaction effect	Total effect
Gradient-based	One-at-a-time (OAT)	Local	No (linearity and monotonicity)	✓	×	×	×	×
	Fractional factorial screening (FF)	Qualitative global	No (linearity or monotonicity)	×	✓	×	×	×
	Plackett-Burman screening (PB)	Qualitative global	No (linearity and additivity)	×	✓	×	×	×
	Morris one-at-a-time (MOAT)	Qualitative global	Yes	✓	×	(a)	✓	✓
	Analysis of variance (ANOVA)	Qualitative global	Yes	×	✓	✓	✓	×
	Fourier amplitude sensitivity test (FAST)	Quantitative global	Yes	×	✓	×	×	×
Variance-based	Extended FAST (EFAST)	Quantitative global	Yes	×	✓	×	×	✓
	McKay correlation ratios	Quantitative global	Yes	×	✓	✓	×	×
	Sobol' sensitivity indices	Quantitative global	Yes	×	✓	✓	✓	✓
	Linear regression (LR)	Qualitative global	No (linearity or monotonicity)	×	✓	×	×	×
	Multivariate adaptive regression splines (MARS)	Qualitative/quantitative global	Yes	(b)	(b)	(b)	(b)	(b)
	Delta test (DT)	Qualitative/quantitative global	Yes	(b)	(b)	(b)	(b)	(b)
Regression-based	Sum-of-tree (SOT)	Qualitative/quantitative global	Yes	(b)	(b)	(b)	(b)	(b)
	Gaussian process (GP)	Qualitative/quantitative global	Yes	(b)	(b)	(b)	(b)	(b)

(a) Specific order interactions cannot be detected, but all order interactions are measured together; (b) qualitative evaluation, parameter total effects; quantitative evaluation, dependents on the quantitative SA method that the surrogate model combined with

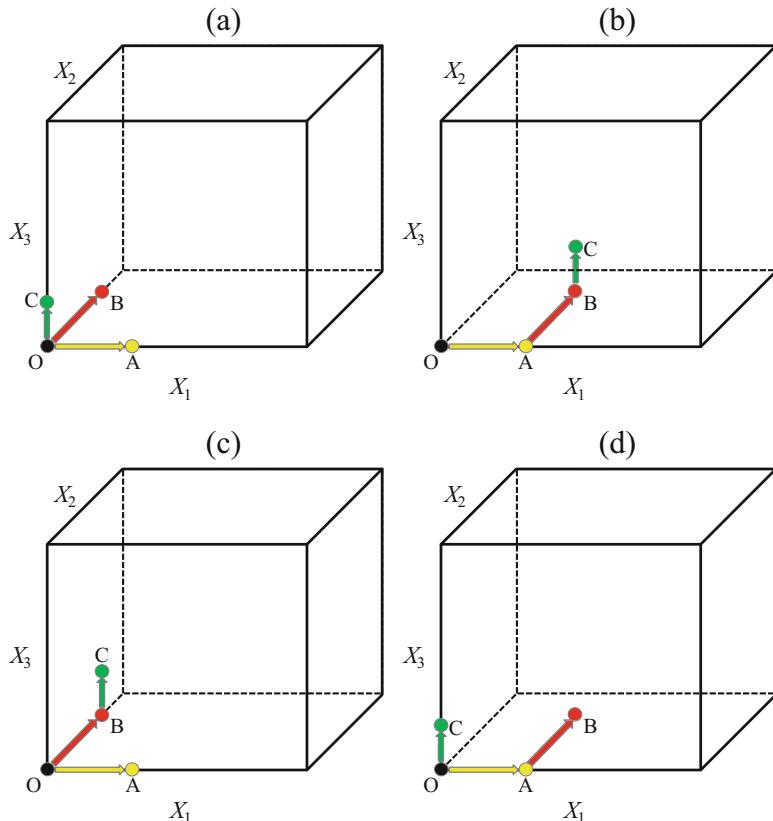


Fig. 1 Illustration of three-parameter one-at-a-time (OAT) design

Table 2 A 2^{5-2} fractional factorial design

Treatment combination	Factor effect					
	I	A	B	C	$D = AB$	$E = AC$
de	+	-	-	-	+	+
a	+	+	-	-	-	-
be	+	-	+	-	-	+
abd	+	+	+	-	+	-
cd	+	-	-	+	+	-
ace	+	+	-	+	-	+
bc	+	-	+	+	-	-
abcde	+	+	+	+	+	+

can be generated from a three-parameter (say A , B , and C) full factorial design and then choosing to confound the two remaining parameters D and E with interactions generated by $D = A \times B$ and $E = A \times C$. These two expressions are the generators of the 2^{5-2} FF design, with $D = A \times B$ means that the main effect of D is confounded

with the interactions involving A and B , while $E = A \times C$ means that the main effect of E is confounded with the interactions involving A and C .

The FF design shares the balance property of corresponding full factorial design, meaning that every level of a parameter appears the same number of times. However, some parameter effects of the FF design would inevitably be aliased with others since it uses only a fraction of the experiments of the full factorial design. A critical consideration when selecting a proper FF design is that the effects of primary interest are aliased only with higher-order interactions that are negligible.

For parameter screening purpose, two-level FF experiments are often designed to investigate parameter main effect. Suppose that parameter X_i ($i = 1, 2, \dots, n$) has two levels as “high” (denoted as $X_i^{(+)}$) and “low” (denoted as $X_i^{(-)}$), the main effect of X_i can be obtained by

$$S_i = \frac{\bar{Y}_i^{(+)} - \bar{Y}_i^{(-)}}{X_i^{(+)} - X_i^{(-)}} \quad (2)$$

where $\bar{Y}_i^{(+)}$ and $\bar{Y}_i^{(-)}$ are mean response values when X_i equals to high and low levels, respectively. FF screening is effective only if the parameter-response relationship is linear or monotonic.

Henderson-Sellers (1993) designed three sets of 32-run two-level FF experiments for assessing the relative importance of 23 ecotype parameters of the Biosphere-Atmosphere Transfer Scheme (BATS) (Dickinson et al. 1986) under three different climatic regimes. A detailed description of the FF screening method for parameter sensitivity analyses of environmental models was given by Henderson-Sellers and Henderson-Sellers (1996).

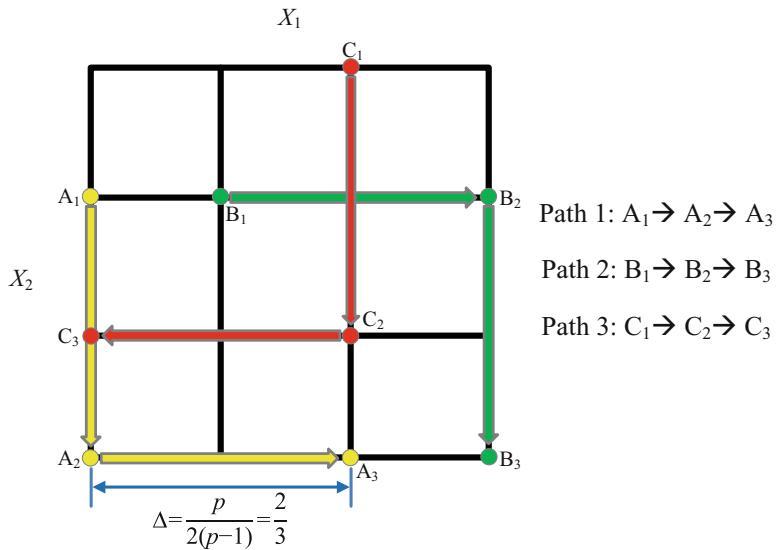
2.1.3 Plackett-Burman Screening

Take a two-level ten-parameter problem, for example, a 1/4 FF design would still require 256 ($=2^8$) experiments. The cost of FF design is sometimes prohibitive for time-consuming high-dimensional problems. Plackett-Burman (PB) design (Plackett and Burman 1946) provides an alternative when the FF design is impractical to implement. With a N -run PB design, one can run a screening experiment for up to $N - 1$ parameters, where N is a multiple of four. The design matrices for two-level n -parameter problem with a sample size up to 100 except 92 are given in Plackett and Burman (1946). Briefly, PB design can be generated by taking a one-dimensional matrix with “+” and “-” signs as the first column (or row), shifting it cyclically one place $N - 2$ times, and adding a final row of “-” signs to complete the design. A 12-run two-level PB design is shown in Table 3, and it can be used for screening experiment containing up to 11 parameters.

The sensitivity measure for PB screening is the same with FF screening. Main effects of PB screening are clear of each other but aliased with two-way interactions. Therefore, it is applicable when the two-way interactions are negligible. Besides, the parameter-response relationship should be linear and additive. Beres and Hawkins (2001) summarized the virtues of PB screening and gave a guide for performing

Table 3 A 12-run two-level Plackett-Burman design

Run	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}	X_{11}
1	+	+	-	+	+	+	-	-	-	+	-
2	-	+	+	-	+	+	+	-	-	-	+
3	+	-	+	+	-	+	+	+	-	-	-
4	-	+	-	+	+	-	+	+	+	-	-
5	-	-	+	-	+	+	-	+	+	+	-
6	-	-	-	+	-	+	+	-	+	+	+
7	+	-	-	-	+	-	+	+	-	+	+
8	+	+	-	-	-	+	-	+	+	-	+
9	+	+	+	-	-	-	+	-	+	+	-
10	-	+	+	+	-	-	-	+	-	+	+
11	+	-	+	+	+	-	-	-	+	-	+
12	-	-	-	-	-	-	-	-	-	-	-

**Fig. 2** Illustration of four-level two-parameter Morris one-at-a-time (MOAT) design, where A_1 , B_1 , and C_1 are random points and all other points are generated following OAT paths

it. Applications of PB screening for SA can also be found in Cryer and Havens (1999), Dion et al. (2011), and Grant et al. (2007).

2.1.4 Morris One-At-a-Time

The Morris one-at-a-time (MOAT) method (Morris 1991) was designed to overcome the deficiency of the OAT method, which is location dependent, by including multidimensional averaging of the local measures. The experimental plans consist of individually randomized OAT designs (Fig. 2). Theoretic basis of this method is

based on the elementary effect, which is representative of the change in a model response due to the change in a particular parameter.

Assume that we have an n -dimension p -level orthogonal parameter space, where each X_i may take on values from $\{0, 1/(p-1), 2/(p-1), \dots, 1\}$. The elementary effect of the i th parameter is defined as

$$d_i = \frac{Y(X_1, \dots, X_{i-1}, X_i \pm \Delta, X_{i+1}, \dots, X_n) - Y(\mathbf{X})}{\Delta} \quad (3)$$

where the increment Δ usually is set to $p/[2(p-1)]$ and p is an even number. Overall and interaction effects of each parameter can then be approximated, respectively, by the mean and standard deviation of the elementary effects from r OAT paths as

$$\mu_i = \sum_{j=1}^r d_i(j)/r \quad (4)$$

and

$$\sigma_i = \sqrt{\sum_{j=1}^r [d_i(j) - \mu_i]^2 / r} \quad (5)$$

The total number of experiments needed for a MOAT screening is $(n+1)r$.

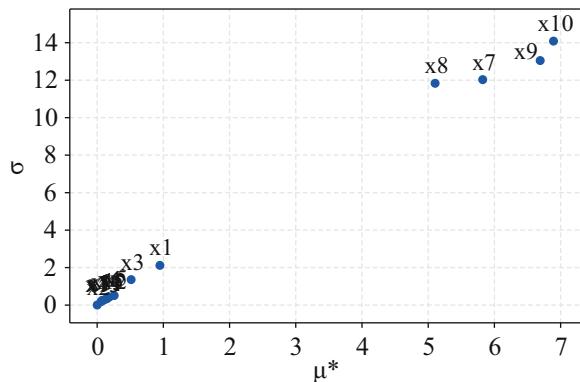
The MOAT screening method has been widely applied due to its efficiency for high-dimensional problems (Francos et al. 2003; Kleijnen 1997). On the other hand, improvements have also been made to this method. For example, van Griensven et al. (2006) replaced the Monte Carlo (MC) sampling with the Latin hypercube (LH) sampling for generating more uniform samples to improve the efficiency of the MOAT screening, which is known as LH-OAT screening. Campolongo et al. (2007) proposed a modified mean μ^* , which is an estimate of the mean of absolute elementary effects, to solve the problem of the compensating effect of opposite signs in elementary effects as

$$\mu_i^* = \sum_{j=1}^r |d_i(j)|/r \quad (6)$$

Example 2.1

Situation. Consider the g -function proposed by Sobol' (1993) as $f = \prod_{i=1}^n g_i(X_i)$, where $g_i(X_i) = (|4X_i - 2| + a_i)/(1 + a_i)$ depends on a nonnegative parameter a_i . Let $n = 15$, $a_1 = 9$, $a_2 = a_3 = 15$, $a_4 = a_5 = a_6 = 50$, $a_7 = a_8 = a_9 = a_{10} = 0$, $a_{11} = a_{12} = a_{13} = a_{14} = a_{15} = 70$, and $X_i \in [0, 1]$ with uniform distribution. Please find the influential parameters of this function.

Fig. 3 MOAT screening for the 15-parameter g -function



Solution. The MOAT screening method is adopted to analyze parameter sensitivity of this 15-parameter problem. The range of each parameter is evenly divided into four levels. We then design 320 ($= (15 + 1) \times 20$) experiments to screen out the insensitive parameters. The SA result is given in Fig. 3. It is easy to distinguish the insensitive parameters (i.e., $X_1, X_2, X_3, X_4, X_5, X_6, X_{11}, X_{12}, X_{13}, X_{14}$, and X_{15}) from the sensitive ones (X_7, X_8, X_9 , and X_{10}) from this figure.

2.2 Variance-Based Methods

Variance-based methods make quantitative decomposition of the variance of model response into the contributions from individual parameters and their interactions. They are model independent and accurate but computationally expensive. Sampling techniques such as LH (McKay et al. 1979), quasi-MC (QMC) (Sobol' 1990), orthogonal array (OA) (Owen 1992), and orthogonal array-based Latin hypercube (OALH) (Tang 1993) have been widely used to generate uniformly distributed samples for variance-based SA.

2.2.1 Analysis of Variance

Analysis of variance (ANOVA) requires no assumptions for the relationship between model parameters and responses, but the responses should be normally distributed with same variance. Suppose we consider a problem with two independent factors (i.e., parameters) A and B . Among them, factor A has a levels (or treatments) as A_1, A_2, \dots, A_a , and factor B has b levels as B_1, B_2, \dots, B_b . Each level combination of factors A and B is repeated n times ($n \geq 2$). Each model response can be recorded as Y_{ijk} , where $i = 1, 2, \dots, a; j = 1, 2, \dots, b; k = 1, 2, \dots, n$. Y_{ijk} is independent of each other, and $Y_{ijk} \sim N(\mu_{ij}, \sigma^2)$. The ANOVA model of this two-factor problem can be presented in terms of a linear statistical model as

$$\begin{cases} Y_{ijk} = \mu_{ij} + \varepsilon_{ijk} \\ \mu_{ij} = \mu + \alpha_i + \beta_j + \gamma_{ij} \end{cases} \quad (7)$$

where μ is the overall mean, α_i is the main effect of the i th level A_i , β_j is the main effect of the j th level B_j , γ_{ij} is the interaction effect of level combination (A_i, B_j) , μ_{ij} is the mean of level combination (A_i, B_j) , and $\varepsilon_{ijk} \sim N(0, \sigma^2)$ is the random error (or residual).

If we denote

$$\bar{Y} = \frac{Y_{...}}{abn} = \frac{1}{abn} \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n Y_{ijk} \quad (8)$$

$$\bar{Y}_{ij\cdot} = \frac{1}{n} Y_{ij\cdot} = \frac{1}{n} \sum_{k=1}^n Y_{ijk} \quad (9)$$

$$\bar{Y}_{i..} = \frac{1}{bn} Y_{i..} = \frac{1}{bn} \sum_{j=1}^b \sum_{k=1}^n Y_{ijk} \quad (10)$$

$$\bar{Y}_{.j} = \frac{1}{an} Y_{.j} = \frac{1}{an} \sum_{i=1}^a \sum_{k=1}^n Y_{ijk} \quad (11)$$

where abn is the total number of experiments and “dot” subscript notation represents the summation over the subscript that it replaced. Thus, the total sum of squares S_T can be expressed and then decomposed as

$$\begin{aligned} S_T &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y})^2 \\ &= \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n [(\bar{Y}_{i..} - \bar{Y}) + (\bar{Y}_{.j} - \bar{Y}) + (\bar{Y}_{ij\cdot} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y}) + (Y_{ijk} - \bar{Y}_{ij\cdot})]^2 \\ &= bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y})^2 + an \sum_{j=1}^b (\bar{Y}_{.j} - \bar{Y})^2 + n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij\cdot} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y})^2 \\ &\quad + \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij\cdot})^2 \end{aligned} \quad (12)$$

The ANOVA table for this two-way fixed effects model is given in Table 4. F -test can be used to determine whether there exists a significant difference among treatment means of one factor or interactions between two factors, at a significance level of α . The higher the F value, the more significant the main effect or interaction effect is to the factor.

Table 4 ANOVA table for the two-way fixed effects model

Source of variation	Sum of squares	Degree of freedom	Mean square	F statistic ^a
Factor A	$S_A = bn \sum_{i=1}^a (\bar{Y}_{i..} - \bar{Y})^2$	$a - 1$	$MS_A = \frac{S_A}{a-1}$	$F_A = \frac{MS_A}{MS_E}$
Factor B	$S_B = an \sum_{j=1}^b (\bar{Y}_{.j} - \bar{Y})^2$	$b - 1$	$MS_B = \frac{S_B}{b-1}$	$F_B = \frac{MS_B}{MS_E}$
Interaction	$S_{A \times B} = n \sum_{i=1}^a \sum_{j=1}^b (\bar{Y}_{ij.} - \bar{Y}_{i..} - \bar{Y}_{.j} + \bar{Y})^2$	$(a - 1)(b - 1)$	$MS_{A \times B} = \frac{S_{A \times B}}{(a-1)(b-1)}$	$F_{A \times B} = \frac{MS_{A \times B}}{MS_E}$
Error	$S_E = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y}_{ij.})^2$	$ab(n - 1)$	$MS_E = \frac{S_E}{ab(n-1)}$	
Total	$S_T = \sum_{i=1}^a \sum_{j=1}^b \sum_{k=1}^n (Y_{ijk} - \bar{Y})^2$	$abn - 1$		

^a $F_A \sim F[a - 1, ab(n - 1)]$, $F_B \sim F[b - 1, ab(n - 1)]$, $F_{A \times B} \sim F[(a - 1)(b - 1), ab(n - 1)]$

Frey and Patil (2002) gave a detailed description of the ANOVA method, including its advantages and disadvantages. Mokhtari and Frey (2005) showed that ANOVA is more reliable than correlation and regression methods by applying them for SA of a two-dimensional probabilistic risk assessment model. The reliability of ANOVA method was also illustrated by Tang et al. (2007).

2.2.2 Fourier Amplitude Sensitivity Test

Fourier amplitude sensitivity test (FAST) was presented by Cukier et al. (1973) for SA of multiparameter nonlinear model, in which conditional variances are represented by coefficients from the multiple Fourier series expansion of the response function and the ergodic theorem (Weyl 1938) is applied to transform the multi-dimensional integral into a one-dimensional integral in the evaluation of the Fourier coefficients. The FAST method is capable of computing the main effect of each parameter to the response variance.

Let $Y = f(\mathbf{X}) = f(X_1, X_2, \dots, X_n)$, where $X_i \in [0, 1]$ and $i = 1, 2, \dots, n$. Consider a set of transfer functions

$$X_i(s) = G_i[\sin(\omega_i s)] \quad (13)$$

where $\{\omega_i\}$ is a set of frequencies and $s \in (-\infty, \infty)$. The key idea of FAST is to apply the ergodic theorem to transform the n -dimensional integral $\int_0^1 \int_0^1 \dots \int_0^1 f(\mathbf{X}) dX_1 dX_2 \dots dX_n$ into a one-dimensional integral $\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(s) ds$. Since the numerical computation of this integral is impossible for an incommensurate set of frequencies, an approximate numerical integration can be made by using a set of positive integer frequencies, which makes the search curve s not space-filling but periodic with a 2π period. By considering $f(s)$ within the finite interval

$(-\pi, \pi)$, the expectation and variance of Y can then be approximated, respectively, by

$$E(Y) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s)ds \quad (14)$$

and

$$V(Y) \approx \frac{1}{2\pi} \int_{-\pi}^{\pi} f^2(s)ds - E^2(Y) \quad (15)$$

Following Parseval's theorem, we have

$$\frac{1}{\pi} \int_{-\pi}^{\pi} f^2(s)ds = \frac{1}{2} A_0^2 + \sum_{p=1}^{\infty} (A_p^2 + B_p^2) \quad (16)$$

where $A_p = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \cos(ps)ds$ and $B_p = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(s) \sin(ps)ds$ are the Fourier coefficients. By applying the above equality to the formulas of expectation and variance, we can get

$$V(Y) \approx 2 \sum_{p=1}^{\infty} (A_p^2 + B_p^2) \quad (17)$$

The first-order sensitivity index can be defined as

$$S_i = \frac{V_i}{V(Y)} = \frac{2 \sum_{q=1}^{\infty} (A_{q\omega_i}^2 + B_{q\omega_i}^2)}{2 \sum_{p=1}^{\infty} (A_p^2 + B_p^2)} \approx \frac{\sum_{q=1}^M (A_{q\omega_i}^2 + B_{q\omega_i}^2)}{\sum_{i=1}^n \sum_{q=1}^M (A_{q\omega_i}^2 + B_{q\omega_i}^2)} \quad (18)$$

where V_i is the estimated conditional variance of the i th parameter and M is the maximum harmonic usually taken to be 4 or higher. A large index means a significant first-order effect.

A lot of transfer functions have been proposed to provide uniformly distributed samples in the n -dimensional unit hypercube. Saltelli et al. (1999) suggested a popular periodic transfer function

$$X_i(s) = \frac{1}{2} + \frac{1}{\pi} \arcsin(\sin \omega_i s + \varphi_i) \quad (19)$$

where φ_i is a random phase shift chosen uniformly in $[0, 2\pi]$. The advantage of this function is that the starting point of the curve can be anywhere within the unit hypercube. By selecting N_r sets $\{\varphi_1, \varphi_2, \dots, \varphi_n\}$, N_r search curves can then be generated, and this procedure was named “resampling” by Saltelli et al. (1999). The sample size of FAST is therefore

$$N = N_r(2M\omega_{\max} + 1) \quad (20)$$

where ω_{\max} is the maximum frequency. The minimum sample size is $2M\omega_{\max} + 1$ when there is only a single search curve.

Schaibly and Shuler (1973) applied FAST to two chemical reaction systems involving sets of coupled nonlinear rate equations and verified its effectiveness in determining the parameter sensitivities of nonlinear complex systems. FAST was adopted by Collins and Avissar (1994) and Rodríguez-Camino and Avissar (1998) to estimate the relative importance of land surface model (LSM) parameters to the variability of surface heat fluxes.

2.2.3 Extended Fourier Amplitude Sensitivity Test

Saltelli et al. (1999) proposed an extension of the FAST to calculate parameter total effect, which is known as extended FAST (EFAST). Assign a frequency ω_i for the i th ($i = 1, 2, \dots, n$) parameter and a different frequency ω_{i-} for all the remaining parameters, where $i-$ means all parameters but the i th one. By evaluating the spectrum at the frequency ω_{i-} and higher harmonics $q \cdot \omega_{i-}$, the total sensitivity index of the i th parameter can be estimated by

$$S_{Ti} = 1 - \frac{V_{i-}}{V(Y)} = 1 - \frac{2 \sum_{q=1}^{\infty} (A_{q \cdot \omega_{i-}}^2 + B_{q \cdot \omega_{i-}}^2)}{2 \sum_{p=1}^{\infty} (A_p^2 + B_p^2)} \approx 1 - \frac{\sum_{q=1}^M (A_{q \cdot \omega_{i-}}^2 + B_{q \cdot \omega_{i-}}^2)}{\sum_{i=1}^n \sum_{q=1}^M (A_{q \cdot \omega_i}^2 + B_{q \cdot \omega_i}^2)} \quad (21)$$

where V_{i-} is the estimated conditional variance except for the i th parameter. A large index means a significant total effect. EFAST needs to choose two frequencies ω_i and ω_{i-} for each parameter, and usually a higher value is assigned to ω_i . Unlike FAST method that all indices can be calculated from a single curve, EFAST requires n curves for calculating all $n S_{Ti}$. Therefore, the sample size needed by EFAST is

$$N = nN_r(2M\omega_{\max} + 1) \quad (22)$$

where $\omega_{\max} = \max \{\omega_i, \omega_{i-}\} \equiv \omega_i$ and N_r is the number of resampling times as in FAST. The minimum sample size for EFAST is $n(2M\omega_{\max} + 1)$.

Wang et al. (2013) adopted the EFAST method to analyze the parameter sensitivity of the World Food Studies (WFOST) crop growth model. Other applications of the EFAST method can be found in Confalonieri et al. (2010) and Reusser et al. (2011).

2.2.4 McKay Correlation Ratios

McKay (1995) makes ANOVA-like decomposition of response variances for calculating correlation ratio, which is a ratio of the variance of expectation conditioned on one parameter and the total variances and is the representation of parameter main

effect. Tong (2005) extended the idea for main effect analysis to two-way interaction effect analysis for uncorrelated parameters.

Let $E(Y)$ and $V(Y)$ be the expectation and variance of the response Y , respectively, thus $V(Y)$ can be decomposed as

$$V(Y) = V[E(Y|X_i)] + E[V(Y|X_i)] = V[E(Y|X_i, X_j)] + E[V(Y|X_i, X_j)] \quad (23)$$

where X_i and X_j are the i th and j th parameter, respectively, $V[E(Y|X_i)]$ is the variance of the conditional expectation of Y conditioned on X_i , $E[V(Y|X_i)]$ is the residual term measuring the estimated variance of Y by fixing X_i , $V[E(Y|X_i, X_j)]$ is the variance of the conditional expectation of Y conditioned on X_i and X_j , and $E[V(Y|X_i, X_j)]$ is the residual term measuring the estimated variance of Y by fixing X_i and X_j . The correlation ratios of McKay main effect and two-way interaction effect are defined, respectively, as

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)} = \frac{V[E(Y|X_i)]}{V[E(Y|X_i)] + E[V(Y|X_i)]} \quad (24)$$

and

$$S_{ij} = \frac{V[E(Y|X_i, X_j)]}{V(Y)} = \frac{V[E(Y|X_i, X_j)]}{V[E(Y|X_i, X_j)] + E[V(Y|X_i, X_j)]} \quad (25)$$

The former measures the relative contribution of parameter X_i to the response variance, while the latter measures the relative contributions of parameters X_i and X_j together to the response variance. The higher the parameter correlation ratio is, the more significant the parameter effect is.

2.2.5 Sobol' Sensitivity Indices

The global method proposed by Sobol' (1993, 2001) is a milestone for global SA of nonlinear models, which makes ANOVA-like decomposition of response variances for calculating specific order sensitivity indices. This method has received much attention because it can provide accurate and robust sensitivity measures of any orders (Nossent et al. 2011; Wagener et al. 2009).

Let the function $Y = f(\mathbf{X}) = f(X_1, X_2, \dots, X_n)$, where $X_i \in [0, 1]$ and $i = 1, 2, \dots, n$. Assume that the model response can be decomposed into 2^n summands of increasing dimensions as

$$\begin{aligned} Y &= f(\mathbf{X}) \\ &= f_0 + \sum_{i=1}^n f_i(X_i) + \sum_{i=1}^{n-1} \sum_{j=i+1}^n f_{i,j}(X_i, X_j) + \dots \\ &\quad + f_{1, 2, \dots, n}(X_1, X_2, \dots, X_n) \end{aligned} \quad (26)$$

where f_0 is a constant, $f_i(X_i)$ are the functions of one parameter, and $f_{i,j}(X_i, X_j)$ are the functions of two parameters, etc. The above formula is called ANOVA representation of $f(\mathbf{X})$ if the integral of every summand is zero

$$\int_0^1 f_{i_1, \dots, i_s}(X_{i_1}, \dots, X_{i_s}) dX_k = 0, \text{ for } k = i_1, \dots, i_s \quad (27)$$

where $1 \leq i_1 < \dots < i_s \leq n$.

Assume that $f(\mathbf{X})$ is square integrable. The total response variance can be written as

$$V(Y) = \int_0^1 \dots \int_0^1 f^2(\mathbf{X}) d\mathbf{X} - f_0^2 \quad (28)$$

While the contribution of a generic term f_{i_1, \dots, i_s} ($1 \leq i_1 < \dots < i_s \leq n$) to the total variance can be written as

$$V_{i_1, \dots, i_s} = \int_0^1 \dots \int_0^1 f_{i_1, \dots, i_s}^2(X_{i_1}, \dots, X_{i_s}) dX_{i_1} \dots dX_{i_s} \quad (29)$$

Thus the ANOVA-like decomposition of total variance can be expressed as

$$V(Y) = \sum_{s=1}^n \sum_{i_1 < \dots < i_s} V_{i_1, \dots, i_s} = \sum_{i=1}^n V_i + \sum_{i=1}^{n-1} \sum_{j=i+1}^n V_{i,j} + \dots + V_{1, \dots, n} \quad (30)$$

The Sobol' sensitivity indices are defined as

$$S_{i_1, \dots, i_s} = \frac{V_{i_1, \dots, i_s}}{V(Y)}, \quad 1 \leq i_1 < \dots < i_s \leq n \quad (31)$$

and the sum of all indices is $\sum_{s=1}^n \sum_{i_1 < \dots < i_s} S_{i_1, \dots, i_s} = 1$.

Theoretically, this global method can compute sensitivity index of any order. However, the computation for higher-order terms is impractical when the number of parameters is large. Homma and Saltelli (1996) provided a simple way for computing the total effect of each parameter as

$$S_{Ti} = S_i + S_{i, ci} = 1 - S_{ci} \quad (32)$$

where S_i and $S_{i, ci}$ are representations of first-order effect and higher-order effect, respectively, and S_{ci} is the sum of all the S_{i_1, \dots, i_s} terms that excludes the index i .

Example 2.2

Situation. Suppose that we are planning to join a bungee jumping club and would like to enjoy real excitement but stay alive by approaching the ground as close

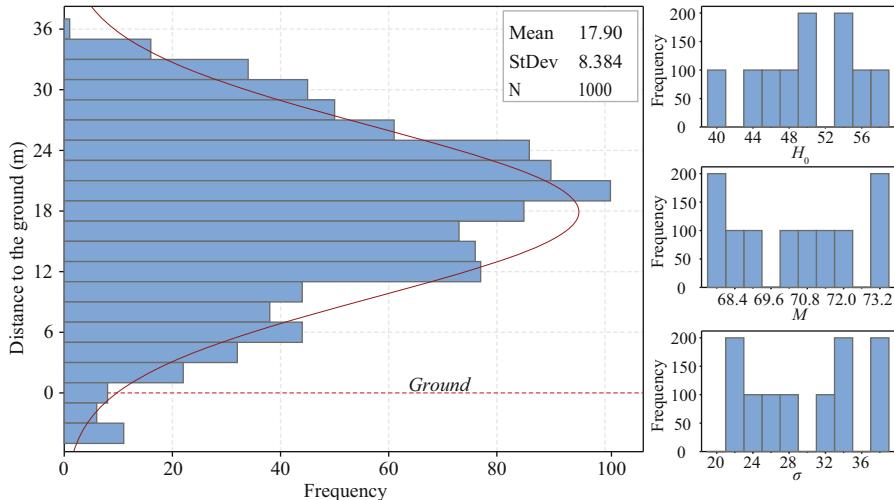


Fig. 4 Probability distributions of the model response and parameters for the bungee experiment

as possible. The minimum distance to the ground during the oscillation can be expressed as $h_{\min} = H_0 - 2Mg/(k\sigma)$, where $g = 9.8 \text{ m/s}^2$ is the acceleration of gravity and $k = 1.5 \text{ N/m}$ is the elastic constant of one strand. The uncertainties are from the height of the platform H_0 (40–60 m), the mass of our body M (67–74 kg), and the number of strands in the cord σ (20–40). In view of uncertainties, please evaluate the risk of safe jumps, and identify the main impact factors.

Solution. Assume that the uncertain factors H_0 , M , and σ are uniformly distributed in their ranges. One thousand LH samples are generated from the three-dimensional parameter space with each of the three factors having ten levels. Samples are then designed as separate experiments to run the model. The probability distributions of the model response and parameters are shown in Fig. 4. In 980 cases out of 1000, the jump is successful, that is, the risk of this bungee experiment is about 2%.

It is observable from the function that the model is linear on factors H_0 and M/σ , but not on M and σ separately. Therefore, we use the model-independent method to analyze parameter sensitivities. Fig. 5 shows the main effect and two-way interaction effect of different experimental factors using McKay's method. The analysis indicates that the number of strands (σ) is the most important factor, the height of the platform (H_0) is of secondary importance, and the influence of the mass of our body (M) is ignorable. The relative contribution of the first two factors (σ and H_0) account for more than 90% of the total variance. Hence, we should not waste much time on the accuracy of our weight but focus on the accuracy of the number of strands and the height of the platform.

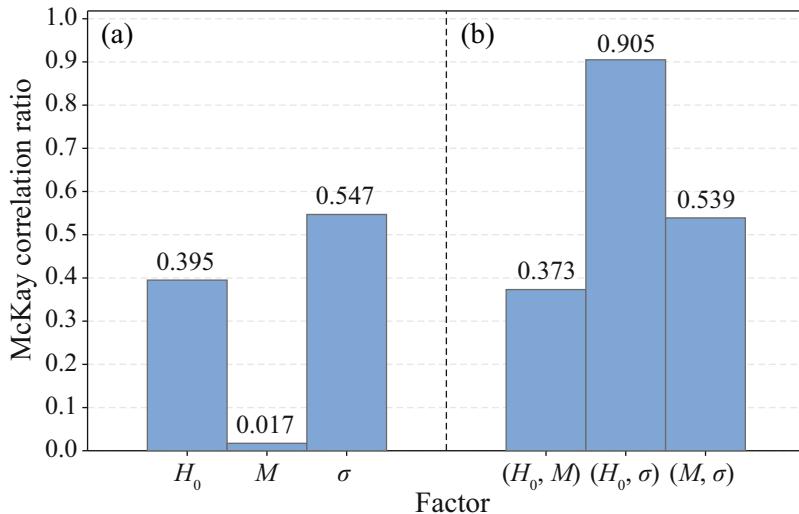


Fig. 5 McKay main effect and two-way interaction effect analysis for the bungee experiment

2.3 Regression-Based Methods

Since today's models are becoming more detailed and realistic and hence have many parameters, high computational costs are prohibitive to quantitative SA using variance-based methods. Therefore, regression-based methods, such as linear regression (LR) (Galton 1886), multivariate adaptive regression splines (MARS) (Friedman 1991), sum-of-trees (SOT) (Chipman et al. 2010), delta test (DT) (Pi and Peterson 1994), and Gaussian process (GP) (MacKay 1998) models, are often employed to screen out insensitive parameters by qualitatively evaluating parameter overall effects. On the other hand, many researchers have also investigated the possibility of replacing the original simulation models with computationally cheaper surrogate models (also called response surface models, metamodels, or emulators) that perform a similar function (Borgonovo et al. 2012; Shahsavani and Grimvall 2011). Regression-based methods are often used to construct surrogate models to improve overall computational efficiency (Wang and Shan 2007). Quantitative SA can then be applied to the surrogate model if it has been proved to be effective for approximating the simulation model. Shahsavani and Grimvall (2011) demonstrated the performance of variance-based SA using surrogate models. Detailed review of available surrogate models can be found in Storlie et al. (2009) and Razavi et al. (2012). A brief introduction of the LR, MARS, SOT, DT, and GP models is given as follows:

2.3.1 Linear Regression

The generalized form of a LR model relating model parameters and response is

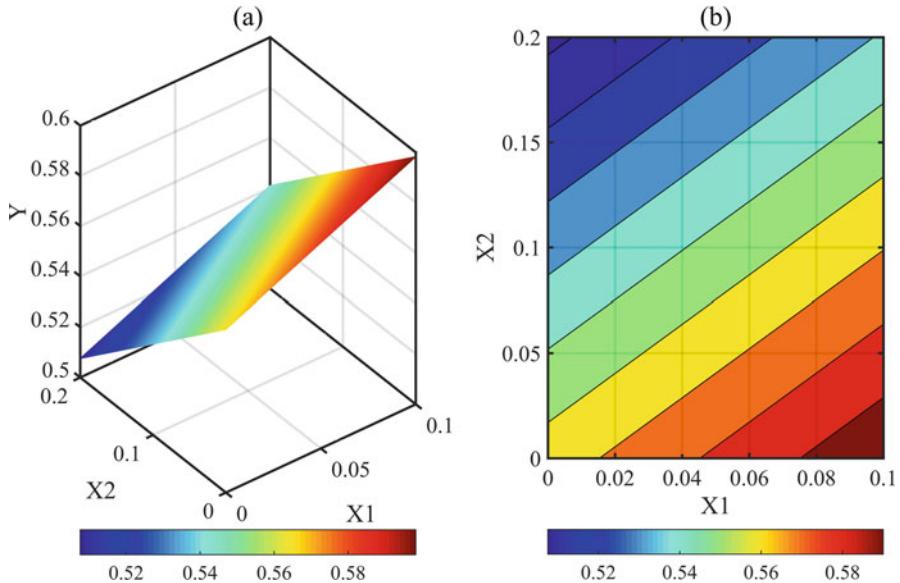


Fig. 6 (a) response surface and (b) contour plot for a two-parameter LR model

$$\hat{Y} = b_0 + \sum_{i=1}^n b_i X_i \quad (33)$$

While the actual simulation model response can be expressed as

$$Y = b_0 + \sum_{i=1}^n b_i X_i + \varepsilon \quad (34)$$

where b_0 is the intercept, b_i is the regression coefficient of the i th parameter X_i , and $\varepsilon \sim N(0, \sigma^2)$ is the error term between the simulation model response and the regression model response. Under the assumption of Gaussian errors, the regression coefficients can be obtained using the least squares approach. An example showing the parameter-response relationship of the LR model is given in Fig. 6.

Utilizing the means and standard deviations of the parameter and response, the LR model is usually normalized to

$$\frac{\hat{Y}^k - \bar{Y}}{\hat{s}} = \sum_{i=1}^n \frac{b_i \hat{s}_i}{\hat{s}} \frac{X_i^k - \bar{X}_i}{\hat{s}_i} \quad (35)$$

where k represents the k th sample and $b_i \hat{s}_i / \hat{s}$ is defined as the standardized regression coefficient (SRC), with

$$\hat{s} = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (Y^k - \bar{Y})^2} \quad (36)$$

and

$$\hat{s}_i = \sqrt{\frac{1}{N-1} \sum_{k=1}^N (X_i^k - \bar{X}_i)^2} \quad (37)$$

as the standard deviations of Y and X_i , respectively. A positive value of SRC indicates that X_i and Y tend to move in the same direction, otherwise in the opposite direction. The larger the absolute value of SRC, the more sensitive is the parameter X_i .

SRC is a sensitivity measure based on the linear parameter-response relationship, and it cannot provide reliable indication of parameter sensitivity when the underlying relationship is nonlinear. However, the transformation of raw data into ranks has been proven to work quite well when the parameter-response relationship is monotonic (Iman and Conover 1979). Therefore, standardized rank regression coefficient (SRRC) can be used as parameter sensitivity measure for nonlinear but monotonic problems.

2.3.2 Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) is an extension of LR models, which makes use of the LR, the mathematical construction of splines, the binary recursive partitioning, and brute search intelligent algorithms (Friedman 1991). The general form of MARS can be represented as

$$Y = f(\mathbf{X}) = a_0 + \sum_{m=1}^M a_m \prod_{k=1}^{K_m} [s_{km}(X_{v(k, m)} - t_{km})]_+^q \quad (38)$$

where a_0 is a constant, a_m are fitting coefficients, M is the number of basis functions, K_m is the number of factors in the m th basis function, s_{km} takes on values of either 1 or -1 and indicates the right or left sense of the associated step function, $v(k, m)$ is the label of the independent parameter and $1 \leq v(k, m) \leq n$, t_{km} indicates the knot location, and the exponent q is the order of the spline approximation. The subscript “+” means the function is a truncated power function

$$[s_{km}(X_{v(k, m)} - t_{km})]_+^q = \begin{cases} [s_{km}(X_{v(k, m)} - t_{km})]^q & s_{km}(X_{v(k, m)} - t_{km}) > 0 \\ 0 & s_{km}(X_{v(k, m)} - t_{km}) \leq 0 \end{cases} \quad (39)$$

MARS builds a model in two phases: the forward pass and the backward pass, which is the same as that used by recursive partitioning trees. The forward pass

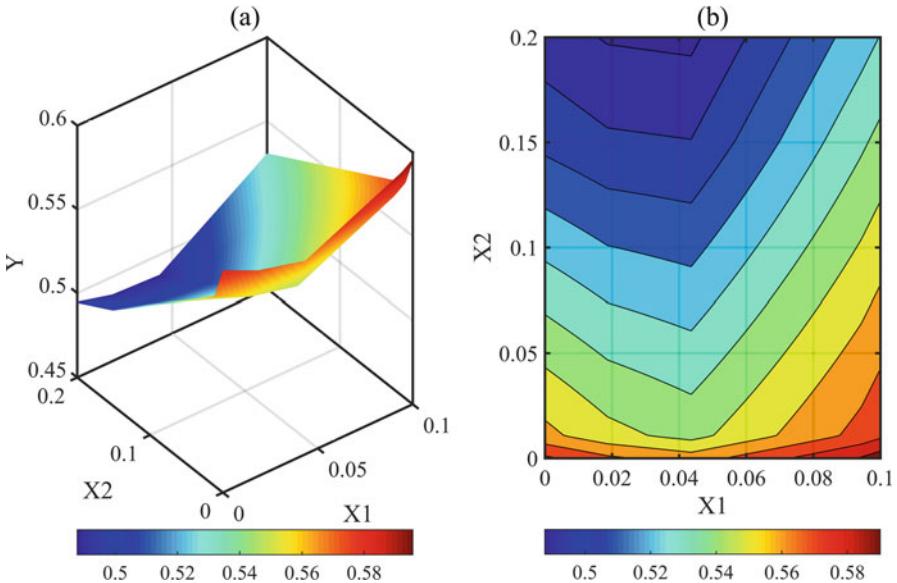


Fig. 7 (a) response surface and (b) contour plot for a two-parameter MARS model

builds an overfit model using all parameters, while the backward pass prunes the overfit model by removing one parameter from the model at a time. The lack-of-fit criterion called generalized cross-validation (GCV) criterion is then computed for both the overfit model and the pruned model

$$GCV(M) = \frac{1}{N} \frac{\sum_{l=1}^N [Y_l - \hat{f}_M(\mathbf{X}_l)]^2}{[1 - C(M)/N]^2} \quad (40)$$

with

$$C(M) = 1 + c(M)d \quad (41)$$

where N is the number of observations in the data set, M is the number of non-constant basis functions in the model $\hat{f}_M(\mathbf{X})$, d is the effective degrees of freedom, and $c(M)$ is a penalty for adding a basis function. An example showing the parameter-response relationship of the MARS model is given in Fig. 7.

The increase in GCV values between the pruned model and the overfitted model can be considered as the importance measure of the removed parameter (Steinberg et al. 1999). The most important parameter is the one that, when omitted, degrades the model fit the most. The score of the i th ($i = 1, 2, \dots, n$) parameter is given by

$$S_i = \frac{\Delta g(i)}{\max\{\Delta g(1), \Delta g(2), \dots, \Delta g(n)\}} \times 100 \quad (42)$$

where $\Delta g(i)$ is the increase in GCV when i th parameter is removed. The larger the GCV increase, the more important is the removed parameter.

2.3.3 Sum-of-Trees

Sum-of-trees (SOT) model is fundamentally a classification (or Bayesian) additive regression tree model with multivariate components (Chipman et al. 2010). Let T denotes a binary tree consisting of a set of interior node decision rules and a set of terminal nodes, and let $M = \{\mu_1, \mu_2, \dots, \mu_b\}$ denotes a set of values associated with each of the b terminal nodes of T . Thus the SOT model can be represented as

$$Y = \sum_{j=1}^m g(\mathbf{X}; T_j, M_j) + \varepsilon \quad (43)$$

where for each binary regression tree T_j and its associated terminal node values M_j , $g(\mathbf{X}; T_j, M_j)$ is the function which assigns $\mu_{ij} \in M_j$ to parameter set \mathbf{X} ; m is the total number of trees, and $\varepsilon \sim (0, \sigma^2)$. An example showing the parameter-response relationship of the SOT model is given in Fig. 8.

The residual sum of squares is used as the criteria for node splitting. A parameter that has the maximum decrease of residual sum of squares will be chosen to split the

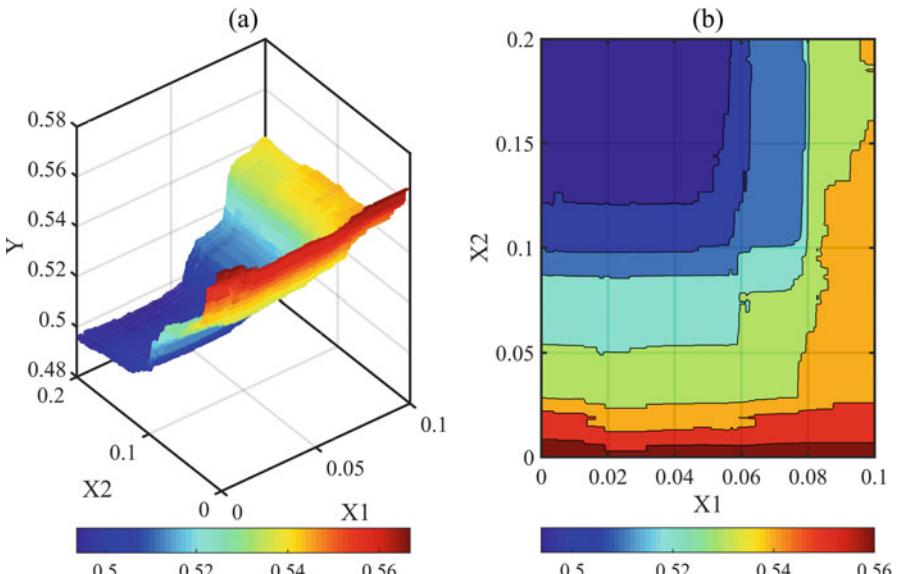


Fig. 8 (a) response surface and (b) contour plot for a two-parameter SOT model

node. The splitting process will not be stopped until per terminal node has minimum number of data points. The total number of splits for each parameter is then taken as the scoring criterion of sensitivity. The score for i th parameter is expressed as

$$S_i = \frac{p(i)}{\max\{p(1), p(2), \dots, p(n)\}} \times 100 \quad (44)$$

where $p(i)$ is the number of splits for i th parameter. The more splits the parameter has, the more sensitive is the parameter.

An illustration of the SOT model is given in Fig. 9. As can be seen from this figure, the two-dimensional space is split into seven subspaces (i.e., subtrees) by six splitting nodes, and the number of splits for X_1 and X_2 is four and two, respectively. Therefore, the sensitivity scores for them are 100 and 50, respectively.

2.3.4 Delta Test

Delta test (DT) is based on the nearest neighbor method for estimating the variance of the residuals (Pi and Peterson 1994). It is founded on the hypothesis of the

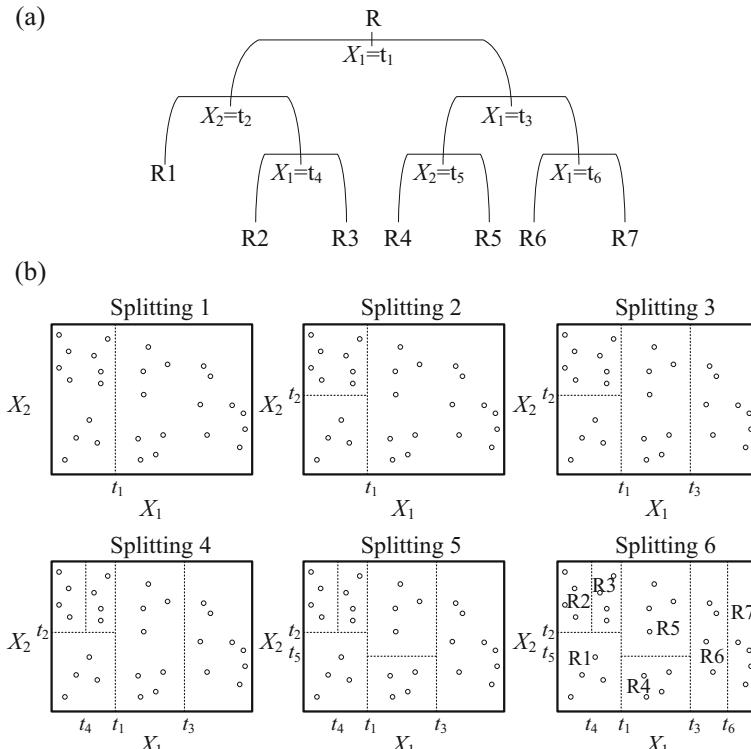


Fig. 9 Illustration of the SOT method for a two-parameter problem

continuity of the regression function, i.e., if two sample points are close in the parameter space, the responses of these two points will be close enough in the response space. Or else, it can be explained by the influence of noise. Assume that we have n parameters and sample points $\mathbf{X}_k \in [0, 1]^n$ for $1 \leq k \leq N$. Let $Y_k = f(\mathbf{X}_k) + \varepsilon_k$, where f is a continuous function with bounded first and second partial derivatives and the residuals $\varepsilon_k \sim (0, \sigma^2)$. Then the points $(\mathbf{X}_k, Y_k)_{k=1}^N$ comprise imitation data set. Let the DT metric that is restricted to the parameter subset space S be

$$\delta_s = \frac{1}{N} \sum_{k=1}^N (Y_k - Y_{N_s(k)})^2 \approx \text{Var}(\varepsilon) \quad (45)$$

where the nearest neighbor of k th sample is

$$N_s(k) = \arg \min_{l \neq k} \|\mathbf{X}_k - \mathbf{X}_l\|_S^2 \quad (46)$$

and the semi-norm

$$\|\mathbf{X}_k - \mathbf{X}_l\|_S^2 = \sum_{p \in S} (\mathbf{X}_k^{(p)} - \mathbf{X}_l^{(p)})^2 \quad (47)$$

Thus the DT metrics for all $2^n - 1$ non-empty parameter subsets can be calculated. Fig. 10 presents an illustration of the nearest neighbors of a point in different subset spaces. An example showing the parameter-response relationship of the nearest neighbor model is given in Fig. 11.

DT was proposed for parameter selection by Eirola et al. (2008). It takes the subset of parameters that minimize the noise variance from all the parameter combinations as sensitive ones. However, this procedure needs an efficient search algorithm to find this subset of parameter combinations. This search process can be too time-consuming, and usually it is impossible to do an exhaustive search of all combinations. DT assesses the final choice using forward sweep and uses genetic algorithm to speed up the search. The first 50 subsets which have the lowest value of DT metrics are taken for sensitivity scoring. The score of the i th ($i = 1, 2, \dots, n$) parameter is given by

$$S_i = \frac{\sum_{m=1}^{50} \delta_S^{(m)} \times I_i^{(m)}}{\sum_{m=1}^{50} \delta_S^{(m)}} \times 100 \quad (48)$$

where $\delta_S^{(m)}$ is the DT metric of the m th subset and $I_i^{(m)} = 1$ if the i th parameter is included in the m th subset, or else $I_i^{(m)} = 0$. A higher score means a more sensitive parameter.

Fig. 10 Illustration of the nearest neighbor of DT method for a two-parameter problem, where points A, B, and C are nearest neighbors of point O in subset spaces $\{X_1\}$, $\{X_2\}$, and $\{X_1, X_2\}$, respectively

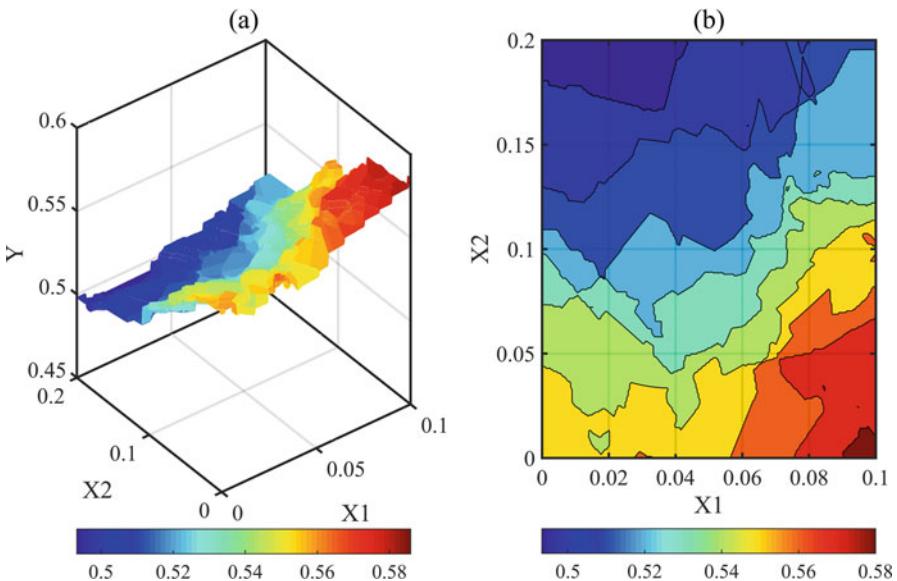
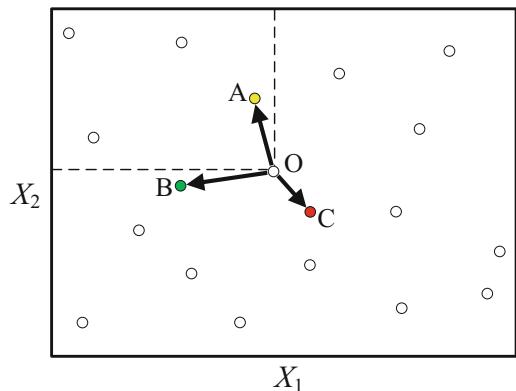


Fig. 11 (a) response surface and (b) contour plot for a two-parameter nearest neighbor model

2.3.5 Gaussian Process

Gaussian process (GP) method characterizes simulation responses over the parameter space as a multivariate Gaussian distribution (MacKay 1998). Let the training data set consist of parameter vectors $\{\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_N\}$ and the corresponding set of response values $\{Y_1, Y_2, \dots, Y_N\}$, where N is the sample size. A GP is a collection of variables $\mathbf{Y} = \{Y_1, Y_2, \dots, Y_N\}$ which have a joint probability distribution

$$P(\mathbf{Y}|\mu(\mathbf{X}), \mathbf{C}) = \frac{1}{Z} \exp \left\{ -\frac{1}{2} [\mathbf{Y} - \mu(\mathbf{X})]^T \mathbf{C}^{-1} [\mathbf{Y} - \mu(\mathbf{X})] \right\} \quad (49)$$

where $\mathbf{C} = \{C(\mathbf{X}_k, \mathbf{X}_l; \Theta)\}_{k,l=1}^N$ is a parameterized covariance function with hyperparameters Θ , $\mu(\mathbf{X})$ is the mean function of the distribution, and Z is the normalization factor. That is, random function \mathbf{Y} can be specified by its mean function $\mu(\mathbf{X})$ and covariance function $C(\mathbf{X}, \mathbf{X}')$. An example showing the parameter-response relationship of the GP model is given in Fig. 12.

Different kinds of mean and covariance functions lead to different GPs. Gibbs and MacKay (1997) presented a software package called “Tpros” for regression problem using GP. The form of covariance function given by them is

$$C(\mathbf{X}_k, \mathbf{X}_l; \Theta) = \theta_1 \exp \left[-\frac{1}{2} \sum_{i=1}^n \frac{(\mathbf{X}_k^{(i)} - \mathbf{X}_l^{(i)})^2}{r_i^2} \right] + \theta_2 + \varepsilon_{kl}(\mathbf{X}_k, \mathbf{X}_l) \quad (50)$$

where θ_1 is the hyperparameter that gives the overall vertical scale, θ_2 is the hyperparameter that gives the vertical uncertainty, $\varepsilon_{kl}(\mathbf{X}_k, \mathbf{X}_l)$ is the noise model, $\mathbf{X}_k^{(i)}$ and $\mathbf{X}_l^{(i)}$ are the i th components of sample points \mathbf{X}_k and \mathbf{X}_l , respectively, and r_i is the length scale that characterizes the distance in the direction of i th parameter over which \mathbf{Y} is expected to vary significantly.

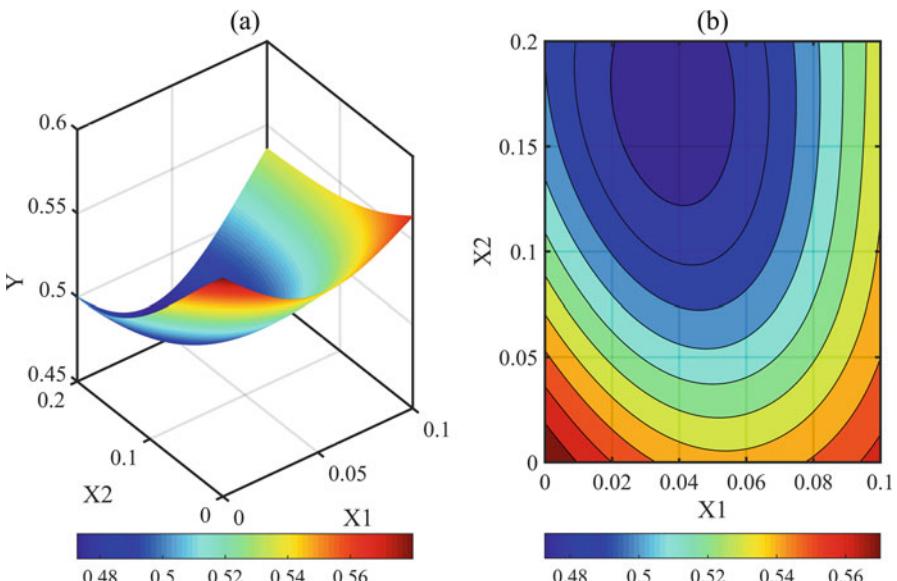


Fig. 12 (a) response surface and (b) contour plot for a two-parameter GP model

As can be seen from the covariance function that when two sample points are close (with respect to their length scales) in parameter space, the exponent is small, and thus the covariance is large, which means their corresponding response values are highly correlated. That is to say, points that are close in parameter space give rise to similar response values. On the contrary, a smaller length scale of a parameter leads to larger difference of response values for two close sample points, which means a more significant influence of this parameter on model response. Therefore, the length scales can be taken as the scoring criteria for parameter screening. The score for i th parameter is expressed as

$$S_i = \frac{1/r_i}{\max\{1/r_1, 1/r_2, \dots, 1/r_n\}} \times 100 \quad (51)$$

Example 2.3

Situation. Consider the artificial computational model proposed by Morris (1991) which contains 20 parameters and has the form as $Y = \beta_0 + \sum_{i=1}^{20} \beta_i w_i + \sum_{i < j}^{20} \beta_{i,j} w_i w_j$ $+ \sum_{i < j < l}^{20} \beta_{i,j,l} w_i w_j w_l + \sum_{i < j < l < s}^{20} \beta_{i,j,l,s} w_i w_j w_l w_s$, where $w_i = 2(X_i - 1/2)$ except for $i = 3, 5$, and 7 , where $w_i = 2[1.1X_i/(X_i + 0.1) - 1/2]$. Each parameter X_i is supposed to be uniformly distributed in $[0, 1]$. Coefficients with relatively large values are as follows: $\beta_i = 20$, with $i = 1, \dots, 10$; $\beta_{i,j} = -15$, with $i, j = 1, \dots, 6$; $\beta_{i,j,l} = -10$, with $i, j, l = 1, \dots, 5$; and $\beta_{i,j,l,s} = 5$, with $i, j, l, s = 1, \dots, 4$. The remainders of the first- and second-order coefficients are independently generated from a standard normal distribution. The remainders of the third- and fourth-order coefficients are set to zero. Assess parameter sensitivity, and reduce parameter dimensionality to a reasonable number.

Solution. We design 1000 LH experiments with each parameter having 1000 levels to run the model. Surrogate-based methods as MARS, SOT, DT, and GP are then adopted to qualitatively evaluate parameter sensitivity scores and quantitatively evaluate Sobol' total sensitivity indices (Sobol'-t). Analysis results for the 20-parameter test function are shown in Fig. 13. As it can be seen from the figure, parameter sensitivity rankings vary across SA methods, but parameter categories of different methods are consistent, that is, parameters 11 to 20 have lower sensitivities than the other 10 parameters. They can therefore be regarded as insensitive parameters and set to any fixed values over their ranges.

3 Which SA Methods to Use?

Many researchers have reviewed some popular SA methods in specific scientific fields, such as chemical reactions (Saltelli et al. 2005; Turányi 1990), ecological modeling (Cariboni et al. 2007), environmental modeling (Hamby 1994; Helton 1993), risk

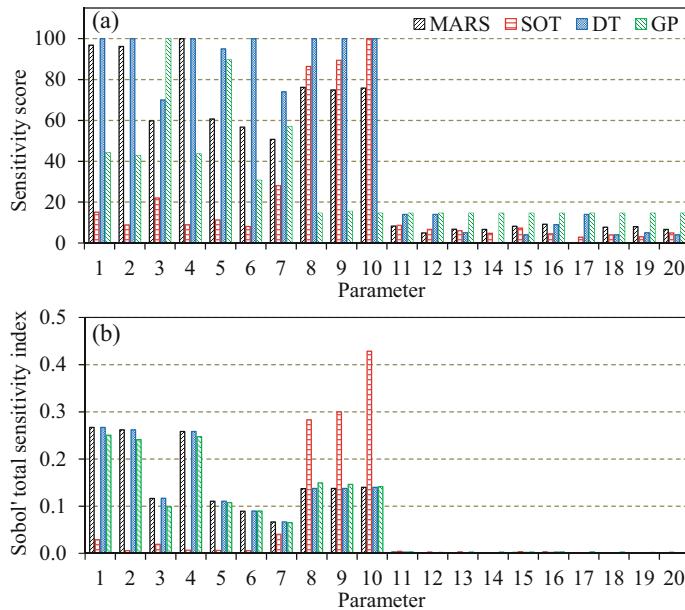


Fig. 13 Surrogate-based (a) parameter sensitivity scores and (b) Sobol' total sensitivity indices for the Morris 20-parameter test function

assessment (Frey et al. 2003; Frey and Patil 2002), linear programming (Ward and Wendell 1990), surface hydrology, and water quality modeling (Reusser et al. 2011), among many other disciplines. Here we provide an extensive discussion on suitability of some commonly used SA methods for different applications. Given the myriad of SA methods, there is a need for further investigations regarding which methods should be used for specific problems.

Various attempts have been made over the years to answer the above question. Campolongo and Saltelli (1997) compared the performances of MOAT screening, SRC, and Sobol' sensitivity indices for SA of an environmental model. Their comparison showed that MOAT screening is the most efficient method, while Sobol' sensitivity indices are the most robust method. Saltelli and Bolado (1998) investigated the relationship between FAST method and Sobol' sensitivity indices and showed that FAST method is equivalent to the first-order Sobol' sensitivity indices but is computationally more efficient than Sobol's method. Saltelli et al. (1999) showed that EFAST method is more efficient than Sobol's method in computing the total-effect indices. By contrasting two variance-based global methods – EFAST and Sobol' sensitivity indices – with the most widely used local method OAT, Saltelli et al. (2000) showed that the variance-based global SA methods are robust, model independent, and computationally convenient. Furthermore, EFAST is numerically more efficient than Sobol's method. By defining different ranges of variation, Lenhart et al. (2002) compared two

forms of a partial derivative-based local method using the hydrologic model SWAT (Arnold et al. 1998). Results indicate that both approaches provide similar results and hence can be considered as equivalent. Marino et al. (2008) reviewed and compared partial rank correlation coefficient (PRCC) with the EFAST method. Their results show that PRCC relies on the monotonicity assumption between parameter and response, whereas EFAST is computationally more expensive. Reusser et al. (2011) compared FAST, EFAST, and Sobol' sensitivity indices by applying them to the hydrologic model TOPMODEL (Beven 1997) in a small mountainous catchment. Their comparison shows that the three methods give comparable results, while FAST is computationally more efficient. Tang et al. (2007) compared four SA methods, including the local analysis using parameter estimation software (PEST), regional SA (RSA), ANOVA, and Sobol's method, for the application to the lumped Sacramento soil moisture accounting model (SAC-SMA) (Burnash et al. 1973) coupled with SNOW-17 (Anderson 1973). Their conclusion is that ANOVA and Sobol's method are overall superior to RSA and PEST, and ANOVA is more efficient but less robust than Sobol's method. Confalonieri et al. (2010) performed SA on a crop model using the MOAT screening; regression-based methods with LH, MC, and QMC sampling; and two variance-based methods: EFAST and Sobol' sensitivity indices. Their experiments demonstrate that the simplest method MOAT screening produced results comparable to those obtained by methods more computationally expensive. Sun et al. (2012) employed the OAT, MOAT, and RSA methods to assess parameter sensitivities of a water quality model. They concluded that the three methods are complementary, but the use of OAT method for interpreting parameter behaviors should be avoided unless the model uncertainty is small. A comparison of the interpretation and computational cost of the local SA method, MOAT screening, and Sobol's method was made by Wainwright et al. (2014) in the application to a pressure propagation problem. The three SA methods were shown to give similar interpretations and importance rankings of model parameters. Although Sobol's method is illustrated to be computationally less efficient than MOAT screening, it is nonsubstitutable because of its capability of interpreting the contribution of each parameter to the response uncertainty.

Generally, global, quantitative, and model-independent SA methods are advocated for all problem settings where finite parameter variations are involved (Saltelli 1999). Considering the large computational cost of global SA, Foglia et al. (2009) even argued that local SA is sufficient to identify insensitive parameters in preliminary model evaluation. Wainwright et al. (2014) think the reason for this argument is because the value of global methods has not been fully appreciated. Global methods are often limited to parameter importance ranking, even though they can provide additional information for systematic understanding of model behavior. The strengths and limitations of several qualitative and quantitative global SA methods were discussed by Gan et al. (2014). Overall, qualitative SA methods are more efficient but less accurate and robust than quantitative ones. The stepwise SA framework proposed by Gan et al. (2015), using qualitative SA method for preliminary parameter screening and then

quantitative SA method for assessing each parameter's contribution to the variance of model response, is an effective and efficient solution for understanding and simplifying complex system models.

4 Summary

The importance of SA for computer-based models is universally recognized. We reviewed a number of commonly used SA methods as gradient-based, variance-based, and regression-based methods. Features and applicability of those methods were described and illustrated with a few examples. Merits and limitations were also given by reviewing the literature on those different SA methods.

The choice of an appropriate SA method depends on (1) the number of considered parameters, (2) the computational costs of the model and the SA method, (3) the ability of the SA method to account for nonlinear and non-monotonic parameter-response relationship, and (4) the ability of the SA method to account for parameter interactions. We therefore emphasize several recommendations for SA based on the selection criteria and characteristics of different methods: (1) local SA methods are effective only for linear and monotonic problem, (2) qualitative global SA methods should be adopted when a single model run takes a significant amount of time and/or the model has a large number of uncertain parameters, (3) quantitative global SA methods can be used to evaluate parameter main effect, interaction effect, and total effect when the parameter dimensionality is low and the model is computationally efficient, and (4) surrogate models are computationally cheaper than time-consuming simulation models and can be used to obtain approximate results for quantitative global SA.

Parameters are often ranked according to the values of specific SA measure, which allows the analyst to focus research efforts on the most sensitive parameters (factor prioritization) and simplify the model by fixing the least sensitive parameters (factor fixing) (Saltelli et al. 2008). Parameter rankings may vary between different SA measures under the same settings, but the sensitivity categories of the parameters should be the same. Therefore, it is important that two types of errors should be avoided in categorizing model parameters, that is, insensitive parameters are classified as sensitive ones (Type I error) and, conversely, sensitive parameters are taken as insensitive ones (Type II error).

Acknowledgment This study was supported by the National Natural Science Foundation of China (41505092) and National Key Research and Development Program of China (2017YFC1404000).

References

- E.A. Anderson, *National Weather Service River Forecast System – Snow Accumulation and Ablation Model* (NOAA, Silver Spring, 1973)

- J.G. Arnold, R. Srinivasan, R.S. Muttiah, J.R. Williams, Large area hydrologic modeling and assessment. Part I: model development. *J. Am. Water Resour. Assoc.* **34**, 73–89 (1998). <https://doi.org/10.1111/j.1752-1688.1998.tb05961.x>
- D.L. Beres, D.M. Hawkins, Plackett – Burman technique for sensitivity analysis of many-parametered models. *Ecol. Model.* **141**, 171–183 (2001). [https://doi.org/10.1016/S0304-3800\(01\)00271-X](https://doi.org/10.1016/S0304-3800(01)00271-X)
- K. Beven, TOPMODEL: a critique. *Hydrol. Process.* **11**, 1069–1085 (1997). [https://doi.org/10.1002/\(SICI\)1099-1085\(199707\)11:9<1069::AID-HYP545>3.0.CO;2-O](https://doi.org/10.1002/(SICI)1099-1085(199707)11:9<1069::AID-HYP545>3.0.CO;2-O)
- E. Borgonovo, W. Castaings, S. Tarantola, Model emulation and moment-independent sensitivity analysis: an application to environmental modelling. *Environ. Model. Softw.* **34**, 105–115 (2012). <https://doi.org/10.1016/j.envsoft.2011.06.006>
- G.E.P. Box, J.S. Hunter, The 2^{k-p} fractional factorial designs: part I. *Technometrics* **3**, 311–351 (1961a). <https://doi.org/10.1080/00401706.1961.10489951>
- G.E.P. Box, J.S. Hunter, The 2^{k-p} fractional factorial designs: part II. *Technometrics* **3**, 449–458 (1961b). <https://doi.org/10.1080/00401706.1961.10489967>
- R.J.C. Burnash, R.L. Ferral, R.A. McGuire, *A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers* (US Department of Commerce, National Weather Service, Sacramento, 1973)
- F. Campolongo, A. Saltelli, Sensitivity analysis of an environmental model: an application of different analysis methods. *Reliab. Eng. Syst. Saf.* **57**, 49–69 (1997). [https://doi.org/10.1016/S0951-8320\(97\)00021-5](https://doi.org/10.1016/S0951-8320(97)00021-5)
- F. Campolongo, J. Cariboni, A. Saltelli, An effective screening design for sensitivity analysis of large models. *Environ. Model. Softw.* **22**, 1509–1518 (2007). <https://doi.org/10.1016/j.envsoft.2006.10.004>
- F. Campolongo, A. Saltelli, J. Cariboni, From screening to quantitative sensitivity analysis. A unified approach. *Comput. Phys. Commun.* **182**, 978–988 (2011). <https://doi.org/10.1016/j.cpc.2010.12.039>
- J. Cariboni, D. Gatelli, R. Liska, A. Saltelli, The role of sensitivity analysis in ecological modelling. *Ecol. Model.* **203**, 167–182 (2007). <https://doi.org/10.1016/j.ecolmodel.2005.10.045>
- H.A. Chipman, E.I. George, R.E. McCulloch, BART: Bayesian additive regression trees. *Ann. Appl. Stat.* **4**, 266–298 (2010). <https://doi.org/10.1214/09-AOAS285>
- D.C. Collins, R. Avissar, An evaluation with the Fourier amplitude sensitivity test (FAST) of which land-surface parameters are of greatest importance in atmospheric modeling. *J. Clim.* **7**, 681–703 (1994). [https://doi.org/10.1175/1520-0442\(1994\)007<0681:AETWTA>2.0.CO;2](https://doi.org/10.1175/1520-0442(1994)007<0681:AETWTA>2.0.CO;2)
- R. Confalonieri, G. Bellocchi, S. Bregaglio, M. Donatelli, M. Acutis, Comparison of sensitivity analysis techniques: a case study with the rice model WARM. *Ecol. Model.* **221**, 1897–1906 (2010). <https://doi.org/10.1016/j.ecolmodel.2010.04.021>
- S.A. Cryer, P.L. Havens, Regional sensitivity analysis using a fractional factorial method for the USDA model GLEAMS. *Environ. Model. Softw.* **14**, 613–624 (1999). [https://doi.org/10.1016/S1364-8152\(99\)00003-1](https://doi.org/10.1016/S1364-8152(99)00003-1)
- R.I. Cukier, C.M. Fortuin, K.E. Shuler, A.G. Petschek, J.H. Schaibly, Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I. theory. *J. Chem. Phys.* **59**, 3873–3878 (1973). <https://doi.org/10.1063/1.1680571>
- C. Daniel, On varying one factor at a time. *Biometrics* **14**, 430–431 (1958)
- R.E. Dickinson, A. Henderson-Sellers, P.J. Kennedy, M.F. Wilson, *Biosphere-Atmosphere Transfer Scheme (BATS) for the NCAR Community Climate Model* (NCAR, Boulder, 1986). <https://doi.org/10.5065/D6668B58>
- E. Dion, L. VanSchalkwyk, E.F. Lambin, The landscape epidemiology of foot-and-mouth disease in South Africa: a spatially explicit multi-agent simulation. *Ecol. Model.* **222**, 2059–2072 (2011). <https://doi.org/10.1016/j.ecolmodel.2011.03.026>
- Q. Duan, J. Schaake, V. Andreassian, S. Franks, G. Goteti, H.V. Gupta, Y.M. Gusev, F. Habets, A. Hall, L. Hay, T. Hogue, M. Huang, G. Leavesley, X. Liang, O.N. Nasonova, J. Noilhan,

- L. Oudin, S. Sorooshian, T. Wagener, E.F. Wood, Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. *J. Hydrol.* **320**, 3–17 (2006). <https://doi.org/10.1016/j.jhydrol.2005.07.031>
- E. Eirola, E. Liitiäinen, A. Lendasse, F. Corona, M. Verleysen, Using the Delta test for variable selection, in ESANN 2008 Proceedings, European Symposium on Artificial Neural Networks, Bruges, Belgium (2008)
- L. Foglia, M.C. Hill, S.W. Mehl, P. Burlando, Sensitivity analysis, calibration, and testing of a distributed hydrological model using error-based weighting and one objective function. *Water Resour. Res.* **45**, W6427 (2009). <https://doi.org/10.1029/2008WR007255>
- A. Francos, F.J. Elorza, F. Bouraoui, G. Bidoglio, L. Galbiati, Sensitivity analysis of distributed environmental simulation models: understanding the model behaviour in hydrological studies at the catchment scale. *Reliab. Eng. Syst. Saf.* **79**, 205–218 (2003). [https://doi.org/10.1016/S0951-8320\(02\)00231-4](https://doi.org/10.1016/S0951-8320(02)00231-4)
- H.C. Frey, S.R. Patil, Identification and review of sensitivity analysis methods. *Risk Anal.* **22**, 553–578 (2002). <https://doi.org/10.1111/0272-4332.00039>
- H.C. Frey, A. Mokhtari, T. Danish, Evaluation of selected sensitivity analysis methods based upon applications to two food safety process risk models. Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC (2003)
- J.H. Friedman, Multivariate adaptive regression splines. *Ann. Stat.* **19**, 1–67 (1991). <https://doi.org/10.1214/aos/1176347963>
- F. Galton, Regression towards mediocrity in hereditary stature. *J. Anthropol. Inst. G. B. Irel.* **15**, 246–263 (1886)
- Y. Gan, Q. Duan, W. Gong, C. Tong, Y. Sun, W. Chu, A. Ye, C. Miao, Z. Di, A comprehensive evaluation of various sensitivity analysis methods: a case study with a hydrological model. *Environ. Model. Softw.* **51**, 269–285 (2014). <https://doi.org/10.1016/j.envsoft.2013.09.031>
- Y. Gan, X.-Z. Liang, Q. Duan, H.I. Choi, Y. Dai, H. Wu, Stepwise sensitivity analysis from qualitative to quantitative: application to the terrestrial hydrological modeling of a conjunctive surface-subsurface process (CSSP) land surface model. *J. Adv. Model. Earth Syst.* **7**, 648–669 (2015). <https://doi.org/10.1002/2014MS000406>
- M. Gibbs, D.J.C. MacKay, *Efficient implementation of Gaussian processes*. Unpublished manuscript (1997)
- J. Grant, K.J. Curran, T.L. Guyondet, G. Tita, C. Bacher, V. Koutitonsky, M. Dowd, A box model of carrying capacity for suspended mussel aquaculture in Lagune de la Grande-Entrée, Iles-de-la-Madeleine, Québec. *Ecol. Model.* **200**, 193–206 (2007). <https://doi.org/10.1016/j.ecolmodel.2006.07.026>
- A. van Griensven, T. Meixner, S. Grunwald, T. Bishop, M. Diluzio, R. Srinivasan, A global sensitivity analysis tool for the parameters of multi-variable catchment models. *J. Hydrol.* **324**, 10–23 (2006). <https://doi.org/10.1016/j.jhydrol.2005.09.008>
- D.M. Hamby, A review of techniques for parameter sensitivity analysis of environmental models. *Environ. Monit. Assess.* **32**, 135–154 (1994). <https://doi.org/10.1007/BF00547132>
- J.C. Helton, Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal. *Reliab. Eng. Syst. Saf.* **42**, 327–367 (1993). [https://doi.org/10.1016/0951-8320\(93\)90097-I](https://doi.org/10.1016/0951-8320(93)90097-I)
- A. Henderson-Sellers, A factorial assessment of the sensitivity of the BATS land-surface parameterization scheme. *J. Clim.* **6**, 227–247 (1993). [https://doi.org/10.1175/1520-0442\(1993\)006<0227:AFAOTS>2.0.CO;2](https://doi.org/10.1175/1520-0442(1993)006<0227:AFAOTS>2.0.CO;2)
- B. Henderson-Sellers, A. Henderson-Sellers, Sensitivity evaluation of environmental models using fractional factorial experimentation. *Ecol. Model.* **86**, 291–295 (1996). [https://doi.org/10.1016/0304-3800\(95\)00066-6](https://doi.org/10.1016/0304-3800(95)00066-6)
- T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Saf.* **52**, 1–17 (1996). [https://doi.org/10.1016/0951-8320\(96\)00002-6](https://doi.org/10.1016/0951-8320(96)00002-6)

- R.L. Iman, W.J. Conover, The use of the rank transform in regression. *Technometrics* **21**, 499–509 (1979). <https://doi.org/10.1080/00401706.1979.10489820>
- A.J. Jakeman, R.A. Letcher, J.P. Norton, Ten iterative steps in development and evaluation of environmental models. *Environ. Model. Softw.* **21**, 602–614 (2006). <https://doi.org/10.1016/j.envsoft.2006.01.004>
- J.P.C. Kleijnen, Sensitivity analysis and related analyses: a review of some statistical techniques. *J. Stat. Comput. Simul.* **57**, 111–142 (1997). <https://doi.org/10.1080/00949659708811805>
- T. Lenhart, K. Eckhardt, N. Fohrer, H.G. Frede, Comparison of two different approaches of sensitivity analysis. *Phys. Chem. Earth* **27**, 645–654 (2002). [https://doi.org/10.1016/S1474-7065\(02\)00049-9](https://doi.org/10.1016/S1474-7065(02)00049-9)
- D.J.C. MacKay, Introduction to Gaussian processes, in *Neural Networks and Machine Learning*, ed. by C.M. Bishop, (Springer, Berlin, 1998). pp 133–165
- S. Marino, I.B. Hogue, C.J. Ray, D.E. Kirschner, A methodology for performing global uncertainty and sensitivity analysis in systems biology. *J. Theor. Biol.* **254**, 178–196 (2008). <https://doi.org/10.1016/j.jtbi.2008.04.011>
- L.S. Matott, J.E. Babendreier, S.T. Purucker, Evaluating uncertainty in integrated environmental models: a review of concepts and tools. *Water Resour. Res.* **45**, W6421 (2009). <https://doi.org/10.1029/2008WR007301>
- M.D. McKay, *Evaluating Prediction Uncertainty* (Los Alamos National Laboratory (LANL), Los Alamos, 1995)
- M.D. McKay, R.J. Beckman, W.J. Conover, A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**, 239–245 (1979). <https://doi.org/10.1080/00401706.2000.10485979>
- A. Mokhtari, H.C. Frey, Sensitivity analysis of a two-dimensional probabilistic risk assessment model using analysis of variance. *Risk Anal.* **25**, 1511–1529 (2005). <https://doi.org/10.1111/j.1539-6924.2005.00679.x>
- M.D. Morris, Factorial sampling plans for preliminary computational experiments. *Technometrics* **33**, 161–174 (1991). <https://doi.org/10.1080/00401706.1991.10484804>
- J. Nossent, P. Elsen, W. Bauwens, Sobol' sensitivity analysis of a complex environmental model. *Environ. Model. Softw.* **26**, 1515–1525 (2011). <https://doi.org/10.1016/j.envsoft.2011.08.010>
- A.B.. Owen, Orthogonal arrays for computer experiments, integration and visualization. *Stat. Sin.* **2**, 439–452 (1992)
- H. Pi, C. Peterson, Finding the embedding dimension and variable dependencies in time series. *Neural Comput.* **6**, 509–520 (1994). <https://doi.org/10.1162/neco.1994.6.3.509>
- R.L. Plackett, J.P. Burman, The design of optimum multifactorial experiments. *Biometrika* **33**, 305–325 (1946). <https://doi.org/10.1093/biomet/33.4.305>
- O. Rakovec, M.C. Hill, M.P. Clark, A.H. Weerts, A.J. Teuling, R. Uijlenhoet, Distributed evaluation of local sensitivity analysis (DELSA), with application to hydrologic models. *Water Resour. Res.* **50**, 409–426 (2014). <https://doi.org/10.1002/2013WR014063>
- S. Razavi, B.A. Tolson, D.H. Burn, Review of surrogate modeling in water resources. *Water Resour. Res.* **48**, W7401 (2012). <https://doi.org/10.1029/2011WR011527>
- B. Renard, D. Kavetski, G. Kuczera, M. Thyre, S.W. Franks, Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resour. Res.* **46**, W5521 (2010). <https://doi.org/10.1029/2009WR008328>
- D.E. Reusser, W. Buytaert, E. Zehe, Temporal dynamics of model parameter sensitivity for computationally expensive models with the Fourier amplitude sensitivity test. *Water Resour. Res.* **47**, W7551 (2011). <https://doi.org/10.1029/2010WR009947>
- E. Rodríguez-Camino, R. Avissar, Comparison of three land-surface schemes with the Fourier amplitude sensitivity test (FAST). *Tellus A* **50**, 313–332 (1998). <https://doi.org/10.3402/tellusa.v50i3.14529>
- A. Saltelli, Sensitivity analysis: could better methods be used? *J. Geophys. Res.* **104**, 3789–3793 (1999). <https://doi.org/10.1029/1998JD100042>

- A. Saltelli, R. Bolado, An alternative way to compute Fourier amplitude sensitivity test (FAST). *Comput. Stat. Data Anal.* **26**, 445–460 (1998). [https://doi.org/10.1016/S0167-9473\(97\)00043-1](https://doi.org/10.1016/S0167-9473(97)00043-1)
- A. Saltelli, S. Tarantola, K.P.S. Chan, A quantitative model-independent method for global sensitivity analysis of model output. *Technometrics* **41**, 39–56 (1999). <https://doi.org/10.2307/1270993>
- A. Saltelli, S. Tarantola, F. Campolongo, Sensitivity analysis as an ingredient of modeling. *Stat. Sci.* **15**, 377–395 (2000). <https://doi.org/10.1214/ss/1009213004>
- A. Saltelli, M. Ratto, S. Tarantola, F. Campolongo, Sensitivity analysis for chemical models. *Chem. Rev.* **105**, 2811–2827 (2005). <https://doi.org/10.1021/cr040659d>
- A. Saltelli, M. Ratto, S. Tarantola, F. Campolongo, Sensitivity analysis practices: strategies for model-based inference. *Reliab. Eng. Syst. Saf.* **91**, 1109–1125 (2006). <https://doi.org/10.1016/j.ress.2005.11.014>
- A. Saltelli, M. Ratto, T. Andres, F. Campolongo, *Global Sensitivity Analysis: The Primer* (Wiley, Chichester, 2008)
- J.H. Schaibly, K.E. Shuler, Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. II applications. *J. Chem. Phys.* **59**, 3879–3888 (1973). <https://doi.org/10.1063/1.1680572>
- D. Shahsavani, A. Grimvall, Variance-based sensitivity analysis of model outputs using surrogate models. *Environ. Model. Softw.* **26**, 723–730 (2011). <https://doi.org/10.1016/j.envsoft.2011.01.002>
- A. Sieber, S. Uhlenbrook, Sensitivity analyses of a distributed catchment model to verify the model structure. *J. Hydrol.* **310**, 216–235 (2005). <https://doi.org/10.1016/j.jhydrol.2005.01.004>
- I.M. Sobol', Quasi-Monte Carlo methods. *Prog. Nucl. Energy* **24**, 55–61 (1990). [https://doi.org/10.1016/0149-1970\(90\)90022-W](https://doi.org/10.1016/0149-1970(90)90022-W)
- I.M. Sobol', Sensitivity analysis for nonlinear mathematical models. *Math. Model. Comput. Exp.* **1**, 407–414 (1993)
- I.M. Sobol', Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul.* **55**, 271–280 (2001). [https://doi.org/10.1016/S0378-4754\(00\)00270-6](https://doi.org/10.1016/S0378-4754(00)00270-6)
- D. Steinberg, P.L. Colla, K. Martin, *MARS User Guide* (Salford Systems, San Diego, 1999)
- C.B. Storlie, L.P. Swiler, J.C. Helton, C.J. Sallaberry, Implementation and evaluation of nonparametric regression procedures for sensitivity analysis of computationally demanding models. *Reliab. Eng. Syst. Saf.* **94**, 1735–1763 (2009). <https://doi.org/10.1016/j.ress.2009.05.007>
- X.Y. Sun, L. Newham, B. Croke, J.P. Norton, Three complementary methods for sensitivity analysis of a water quality model. *Environ. Model. Softw.* **37**, 19–29 (2012). <https://doi.org/10.1016/j.envsoft.2012.04.010>
- B. Tang, Orthogonal array-based Latin hypercubes. *J. Am. Stat. Assoc.* **88**, 1392–1397 (1993). <https://doi.org/10.2307/2291282>
- Y. Tang, P. Reed, T. Wagener, K. van Werkhoven, Comparing sensitivity analysis methods to advance lumped watershed model identification and evaluation. *Hydrol. Earth Syst. Sci.* **11**, 793–817 (2007). <https://doi.org/10.5194/hess-11-793-2007>
- C. Tong, *PSUADE User's Manual* (Lawrence Livermore National Laboratory (LLNL), Livermore, 2005)
- T. Turányi, Sensitivity analysis of complex kinetic systems. Tools and applications. *J. Math. Chem.* **5**, 203–248 (1990). <https://doi.org/10.1007/BF01166355>
- L. Uusitalo, A. Lehikoinen, I. Helle, K. Myrberg, An overview of methods to evaluate uncertainty of deterministic models in decision support. *Environ. Model. Softw.* **63**, 24–31 (2015). <https://doi.org/10.1016/j.envsoft.2014.09.017>
- T. Wagener, K. van Werkhoven, P. Reed, Y. Tang, Multiobjective sensitivity analysis to understand the information content in streamflow observations for distributed watershed modeling. *Water Resour. Res.* **45**, W2501 (2009). <https://doi.org/10.1029/2008WR007347>

- H.M. Wainwright, S. Finsterle, Y. Jung, Q. Zhou, J.T. Birkholzer, Making sense of global sensitivity analyses. *Comput. Geosci. UK* **65**, 84–94 (2014). <https://doi.org/10.1016/j.cageo.2013.06.006>
- W.E. Walker, P. Harremoës, J. Rotmans, J.P. van der Sluijs, M.B. van Asselt, P. Janssen, M.P. Krämer Von Krauss, Defining uncertainty: a conceptual basis for uncertainty management in model-based decision support. *Integr. Assess.* **4**, 5–17 (2003). <https://doi.org/10.1076/iaij.4.1.5.16466>
- G.G. Wang, S. Shan, Review of metamodeling techniques in support of engineering design optimization. *J. Mech. Des.* **129**, 370–380 (2007). <https://doi.org/10.1115/1.2429697>
- J. Wang, X. Li, L. Lu, F. Fang, Parameter sensitivity analysis of crop growth models based on the extended Fourier amplitude sensitivity test method. *Environ. Model. Softw.* **48**, 171–182 (2013). <https://doi.org/10.1016/j.envsoft.2013.06.007>
- C. Wang, Q. Duan, C. Tong, W. Gong, A GUI platform for uncertainty quantification of complex dynamical models. *Environ. Model. Softw.* **76**, 1–12 (2016). <https://doi.org/10.1016/j.envsoft.2015.11.004>
- J.E. Ward, R.E. Wendell, Approaches to sensitivity analysis in linear programming. *Ann. Oper. Res.* **27**, 3–38 (1990). <https://doi.org/10.1007/BF02055188>
- H. Weyl, Mean motion. *Am. J. Math.* **60**, 889–896 (1938)

Part VI

Observation and Data Assimilation



Fundamentals of Data Assimilation and Theoretical Advances

Hamid Moradkhani, Grey S. Nearing, Peyman Abbaszadeh, and Sahani Pathiraja

Contents

1	Introduction	676
1.1	Purpose of Data Assimilation	676
1.2	State-Space Models	677
1.3	Types of Data Assimilation	678
2	Error Characterization	679
2.1	Uncertainty Quantification	679
3	Data Assimilation Methods	681
3.1	Linear Data Assimilation	681
3.2	Partially and Fully Nonlinear Deterministic Data Assimilation	683
3.3	Ensemble Data Assimilation	686
4	Applications	693
4.1	Variational	693
4.2	Kalman-Based Filters	693
4.3	Particle Filters	693
4.4	Parameter Inference and Model Structures	694
5	Conclusion	695
	References	695

H. Moradkhani (✉) · P. Abbaszadeh

Department of Civil, Construction and Environmental Engineering, The University of Alabama,
Tuscaloosa, AL, USA

e-mail: hmoradkhani@ua.edu; pabbaszadeh@crimson.ua.edu

G. S. Nearing

Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA
e-mail: gsnearing@ua.edu

S. Pathiraja

Institute for Mathematics, University of Potsdam, Potsdam, Germany
e-mail: pathiraja@uni-potsdam.de

Abstract

Hydrometeorological predictions are not perfect as models often suffer either from inadequate conceptualization of underlying physics or non-uniqueness of model parameters or inaccurate initialization. During the past two decades, Data Assimilation (DA) has received increased prominence among researchers and practitioners as an effective and reliable method to integrate the hydrometeorological observations from in situ measure and remotely-sensed sensors into predictive models for enhancing the forecast skills while taking into account all sources of uncertainties. The successful application of DA in different disciplines has resulted in an ever-increasing publications. This chapter provides a progressive essay covering fundamental and theoretical underpinnings of DA techniques and their applications in a variety of scientific fields. More detailed examples of applications are presented in following chapters in this section.

Keywords

Hydrometeorological predictions · Uncertainty · Data Assimilation (DA)

1 Introduction

1.1 Purpose of Data Assimilation

Forecasting in hydrometeorology is challenging due to the complex, heterogeneous, nonstationary, and nonlinear interactions between water and the environment. Such complexities make precise modeling of hydrometeorological processes infeasible, leading to persistent uncertainty in forecasting systems. Uncertainties are present in all aspects of hydrometeorological modeling and forecasting, due to errors or imprecision in observations of pertinent states and fluxes, gaps in knowledge of the physical science, and spatiotemporal heterogeneities that complicate the highly dynamic nature of water movement through the land and atmosphere. Such prevalence of uncertainty reduces a forecaster's ability to determine the magnitude and timing of catastrophic events (i.e., floods, droughts) and quantify variables of interest (i.e., water supply, soil moisture). Due to these uncertainties, it is advantageous to utilize the full extent of information about the state of the environment in a single unified forecast. This unification of information from models and observations for reduced uncertainty is the premise of data assimilation (DA).

Most generally, DA is defined as the application of Bayes' theorem to probabilistically condition the states of a dynamical model on observations. There are many different computational techniques for implementing DA, and each technique relies on a different set of tractability approximations. Because no tractable DA technique is perfect, it is important to understand the nature of the particular simulation model and observation data set that are being used for a particular forecast problem.

Most commonly, DA results in improvements to model forecasts by improving initial states (DeChant and Moradkhani 2011a); however it is also possible to use DA

to improve model parameters (e.g., Moradkhani et al. 2005a, b; Vrugt et al. 2005; Montzka et al. 2011; Pathiraja et al. 2016a, 2017; Abbaszadeh et al. 2018) and model structures (e.g., Bulygina and Gupta 2011; Nearing and Gupta 2015). DA methods that update initial states can typically be implemented sequentially in time; this is called *filtering*, and filtering approximations make DA particularly applicable to real-time forecasting problems. Forecasting agencies often want to make forecasts at regular time intervals, and a method that sequentially improves initial states is especially applicable to this type of situation.

In addition, many popular DA techniques use ensembles, which allow for probabilistic representation of complex systems and quantification of uncertainties in time-evolving simulations. Ensemble-based DA produces an ensemble of initial states, which allows for initializing an ensemble forecast with a range of possible state values. Ensemble-based DA techniques are therefore especially useful when accurately accounting for forecast uncertainty is important.

1.2 State-Space Models

Since the primary purpose of DA is to improve state estimates within the model, it is important to understand models from a state-space perspective. Dynamical models solve systems of partial differential equations (PDEs), and in hydrology, this is typically done in discrete time. We can therefore write a generalized state-transfer function to represent our discrete-time PDE solution:

$$x_t = f(x_{t-1}, u_t, \theta^f) + \omega_t \quad (1)$$

In Eq. (1), $f(\cdot)$ is a function that governs the evolution of the model state vector x in discrete time. Note that there are generally many model states, so x_t is generally a vector. Since Eq. (1) is an approximate PDE solution, the model requires some boundary conditions, or time-dependent forcing data (u_t), and also a (typically time-independent) parameter vector (θ^f). Since it is unavoidable that the model will have an error, this is accounted for by an additive error term ω_t which is drawn from a distribution. This is referred to as *model error*. The general DA problem is aimed at reducing model error and therefore improving the accuracy of state estimates.

In addition to the forward model operator, $f(\cdot)$, an observation operator, $h(\cdot)$, is necessary to relate observations with system states:

$$y_t = h(x_t, \theta^h) + \epsilon_t \quad (2)$$

In Eq. (2), $h(\cdot)$ is the *observation operator*, which relies on the current states (x_t) and a parameter vector (θ^h) to translate states into observation space. Similar to the forward model operator, the observation operator will have some error, which is typically accounted for using an additive error term like ϵ_t . Although ϵ_t could be specified to retrieve the “true” prediction, it is specified as the correction to reach the observation, which will simplify the explanation of the DA process.

1.3 Types of Data Assimilation

As mentioned above, DA is aimed at improving modeled states by conditioning on observation data. The most general expression for the DA problem is:

$$p(x_{1:t}|y_{1:t}, u_{1:t}, \theta^f, \theta^h) \propto p_h(y_{1:t}|x_{1:t}, \theta^h)p_f(x_{1:t}|u_{1:t}, \theta^f). \quad (3)$$

Equation (3) is Bayes' theorem applied to the problem of estimating model states conditional on observations. p_f and p_h are the probability density functions (PDF) implied by Eqs. (1) and (2), respectively.

Analytical solutions to Eq. (3) are infeasible for almost any real-world problem. In addition, $x_{1:t}$ is very high-dimensional (the dimension of the state vector, x_t , multiplied by the number of time steps), which makes it infeasible to estimate the posterior (i.e., the PDF on the right-hand side of Eq. (3)) by sampling. Thus, to implement Eq. (3) for real-world problems, we almost always require some tractability approximations.

1.3.1 Smoothers Versus Filters

The most common tractability approximation is to restrict information from assimilated observations from moving backward in time. That is, an observation from time t will not affect the state values at times $t-s$, where $s > 0$. When this approximation is used in conjunction with a Markovian model, like Eq. (1), the result is DA *filtering*; in contrast, the full DA problem in Eq. (3) is called *smoothing*. The general filtering problem is as follows:

$$p(x_t|y_{1:t}, u_{1:t}, \theta^f, \theta^h) \propto p_h(y_t|x_t, \theta^h)p_f(x_t|x_{t-1}, u_t, \theta^f, y_{1:t-1}). \quad (4)$$

Notice that the dimension of the posterior is greatly reduced – by a multiplicative factor in the number of time steps. This makes it (sometimes) feasible to sample the posterior effectively at each time step. The Markovian property of the dynamical systems model is explicit in the prior (p_f) in Eq. (4), and it is important to recognize that the prior – i.e., $p_f(x_t|x_{t-1}, u_t, \theta^f, y_{1:t-1})$ – is conditional on past observations.

1.3.2 Linear Versus Nonlinear

One important category of tractability approximations involves treating all or part of the system as linear. The distinction between linear and nonlinear DA methods is historically important because the original method for DA (i.e., the Kalman filter; Kalman 1960) was based on strong linearity assumptions. This is limiting, as most models in hydrometeorology are highly nonlinear. This has led to an increasing focus on nonlinear methods in recent decades. If a system is truly linear, then Gaussian uncertainties are preserved through the forward model operator, $f(\cdot)$, and observation operator, $h(\cdot)$. In this case, the DA process is significantly simplified, and there is a known analytical solution to Eq. (4). For dynamic systems that are truly linear, and where uncertainty is truly Gaussian, the Kalman filter is an optimal solution to the filtering problem.

Alternatively, in the presence of a partially nonlinear model/problem, a range of techniques are available. These include Kalman filter-based methods (extended Kalman filter, unscented Kalman filter, ensemble Kalman filter) and more

generalized solutions (variational, particle filter). Although a number of techniques are available, none may be considered optimal. Each nonlinear DA technique requires some limiting assumptions, which may complicate the choice of technique when applying DA. The main challenge in nonlinear DA is that there is no perfect inverse model. Therefore, the relationship between the states and the observation must be approximated prior to adjusting the states. Overcoming this challenge has been one of the primary focuses of DA scientists in recent decades.

1.3.3 Deterministic Versus Ensemble

Another distinction is between deterministic and ensemble DA methods. Deterministic methods perform updates on a single model realization, and the result is some metric (typically a mean field or maximum likelihood estimate) from the full posterior of either Eq. (3) or (4). Examples of deterministic DA methods are the Kalman filter, extended Kalman filter, and variational filters and smoothers. Deterministic methods are often more computationally efficient than ensemble methods, but have limitations when applied to complex models. Deterministic techniques require stricter assumptions about the forms of model error distributions and sometimes require model derivatives (e.g., for maximum likelihood estimation over nonlinear models). As a consequence, deterministic methods typically assume the state space at any particular time step is represented by a multivariate Gaussian distribution, which may be questionable in highly nonlinear models.

Alternatively, ensemble DA methods utilize multiple stochastic realizations of the model to represent uncertainties. Examples of ensemble DA methods include the ensemble Kalman filter (EnKF), the particle filter (PF), and the maximum likelihood ensemble filter (MLEF). Ensemble methods have the benefit of estimating the full PDF over model error, as it manifests in the state and/or prediction variables. In many cases, it is easier to apply ensemble methods, as opposed to deterministic methods, to nonlinear models. The primary drawback of ensemble methods is the increased computational demand of simulating the forward model and observational operators multiple times. Although these simulations are easily parallelizable, which reduces the computational demand, ensemble-based techniques still generally require increased simulation run time.

2 Error Characterization

2.1 Uncertainty Quantification

Uncertainty quantification is a key component of many DA systems. Since the intent of any DA system is to reduce uncertainty with respect to some pertinent model value, it is essential to understand the prior probabilistic characteristics of that uncertainty. This generally requires understanding the different sources of uncertainty in the full modeling and DA system. Generally, uncertainties will break down into three categories: boundary conditions, parameters, and model structure (i.e., process uncertainty). Boundary conditions include the initial state (x_0) and the model forcing data (u_t). Boundary conditions are required for solving any PDE system, and as mentioned

previously, estimating improved initial conditions is one of the primary motivations for state estimation with DA. Uncertainty in model forcing data (e.g., precipitation, temperature, radiation) is typically estimated a priori. It is theoretically possible to use DA, or something like Eqs. (3) and/or (4), to condition PDFs over model forcing data, but this is not done regularly in hydrology.

Model parameters (e.g., hydraulic conductivity, streambed roughness) also inevitably contribute some uncertainty to the forecast system. DA can also be used to help reduce parameter uncertainty (see references above), but these methods are not yet common in operational hydrology forecasting.

Finally, process uncertainty is the uncertainty due to incomplete knowledge of the underlying processes within the model. This manifests as errors in the forward model operator $f(\cdot)$ and observation operator $h(\cdot)$. Due to the requirement to discretize processes both spatially and temporally, the model cannot perfectly simulate reality, and therefore the model itself will have uncertainty. There are methods for using DA to infer or condition model structural uncertainty distributions (e.g., Ghahramani and Roweis 1999) – some of which have been applied to river forecasting models (references above); however this is a relatively immature area of DA research and will not be discussed further in this essay.

2.1.1 Probabilistic Simulations

Applying any approximation of either Eq. (3) or (4) requires estimating all relevant uncertainties in the simulation system. Thus, DA inherently requires some type of probabilistic simulation to quantify that uncertainty. Depending on the complexity of this problem, that may be difficult or computationally expensive. Early DA methods targeted linear systems, with the assumption that errors were Gaussian. Probabilistic simulations for linear-Gaussian systems may be performed by propagating the expected value and covariance structures of the modeled state estimates forward in time. More generally, deterministic methods require partial derivatives of $f(\cdot)$ for locating extremum of the posterior state PDF in nonlinear systems.

2.1.2 Ensemble Simulations

Ensemble simulations allow sampling of complex uncertainty distributions. This is beneficial when working with strongly nonlinear models and/or non-Gaussian uncertainties. In an ensemble DA framework, the forecast PDF is represented by multiple stochastic realizations of a model, where each ensemble member is a sample from the forecast density. All of the input uncertainties (parameters, structure, boundary conditions) to the model are sampled, and each sample is propagated through the model, generating an ensemble forecast at each time step. This ensemble forecast is made up of N ensemble members, each with a weight that may be nonuniform, which is dependent on the DA technique applied. We can notate this situation as follows:

$$p(x_t) = \sum_{i=1}^N w_{t,i} \delta(x_t, x_{t,i}). \quad (5)$$

where $x_{t,i}$ is the state from ensemble member i of N , $\delta(\cdot)$ is the Dirac delta function, and $w_{t,i}$ is the weight of ensemble member i . To simulate the ensemble of states, a model is run for N ensemble members, according to Eq. (6):

$$x_{t,i} = f\left(x_{t-1,i}, u_{t,i}, \theta_i^f\right) + \omega_{t,i} \quad (6)$$

In Eq. (6), $x_{t-1,i}$ is the state vector from the previous time step, $u_{t,i}$ is the current forcing sample, and $\omega_{t,i}$ is the current model error sample, each for ensemble member i . Each ensemble member represents a specific point within the state probability distribution. From this ensemble of model states, an ensemble of model-predicted observations may be generated:

$$y_{t,i} = h(x_{t,i}, \theta_i^h) + \epsilon_{t,i}. \quad (7)$$

In Eq. (7), $y_{t,i}$ and $\epsilon_{t,i}$ are the model prediction in observation space and error sample, respectively, for ensemble member i . The quantity $\epsilon_{t,i}$ captures deficiencies in the observation operator h and possibly also parameters θ_i^h (in some cases, these are considered separately with the parameters treated as random variables). Evaluation of Eqs. (6) and (7) for a large ensemble size allows for propagation from uncertainty distributions over parameters, forcing data, and model structure to uncertainty in model states and model-simulated observations. The uncertainty from the various aforementioned sources can be treated individually or lumped together as a “total uncertainty” term quantified by the additive errors ω_t and ϵ_t . The total uncertainty approach can be useful whenever quantifying the uncertainty in the individual sources is challenging. Often the additive errors are assumed to be zero mean Gaussian, although this assumption is seldom appropriate for hydrologic applications. Pathiraja et al. (2018a) presented a data-driven approach to estimate ω_t and ϵ_t from a total uncertainty perspective using only partial observations of the system and without relying on distributional assumptions on the errors. The approach is particularly suited to cases where model error characteristics are dependent on the system states and when the model-observed variables are of principal interest. It works by first generating a sample of additive errors on the latent states and observed variables using a sequential optimization approach. The probability density of these errors is then estimated via nonparametric kernel conditional density estimation, thereby allowing for the characterization of complex error densities.

3 Data Assimilation Methods

3.1 Linear Data Assimilation

3.1.1 Kalman Filter

The Kalman filter (Kalman 1960) was the first true DA technique. Although the Kalman filter is rarely applied in hydrometeorology, due to its specific applicability to linear filtering problems, it is the basis of many generalized filters, making it

a useful starting point for understanding many DA techniques. Kalman's solution to the filtering problem assumes the Gaussian distribution of errors, which greatly simplifies the state-updating process. Since the model is linear, and the form of uncertainty is known, the inversion of the model to estimate the optimal state value is analytical. The Kalman filter is applicable to models with linear state transition functions, of the form:

$$x_t^- = Ax_{t-1}^+ + Bu_t + \omega_t \quad (8)$$

where the model error is drawn from a Gaussian (normal) distribution of known covariance, Σ_m :

$$\omega_t \sim \mathcal{N}(0, \Sigma_m). \quad (9)$$

Within this linear model, A is a state transition matrix and B is an input transition matrix. Equation (8) allows direct propagation of the mean field of the state uncertainty distribution. When uncertainty in the initial states is also Gaussian, such that covariance of x_{t-1}^+ is denoted P_{t-1}^+ , then we can also directly propagate the state uncertainty variance:

$$P_t^- = AP_{t-1}^+A^T + \Sigma_o \quad (10)$$

In addition, the Kalman filter requires a linear observational operator:

$$y_t = Hx_t^- + \varepsilon_t \quad (11)$$

$$\varepsilon_t \sim \mathcal{N}(0, \Sigma_o) \quad (12)$$

In Eqs. (11) and (12), y_t is the observation; H is the observational operator, which is a function only of the modeled states; and ε_t is Gaussian observation error with covariance Σ_o . The Kalman filter only accounts for uncertainties due to model error and observation error. Based on these approximations, the states are linearly correlated with the observations and themselves have normally distributed uncertainty.

If all of these conditions are met, then we can solve the filtering problem (i.e., Eq. (4)) exactly:

$$x_t^+ = x_t^- + K_t(y_t - Hx_t^-) \quad (13)$$

$$K_t = P_t H^T (HP_t H^T + \Sigma_m)^{-1} \quad (14)$$

K_t is called the *Kalman gain* and x_t^+ are the updated model states. The updated state covariance is:

$$P_t^+ = (I - K_t H)P_{t-1}^+ \quad (15)$$

where I is the identity matrix.

3.2 Partially and Fully Nonlinear Deterministic Data Assimilation

3.2.1 Kalman Filter Extensions

Extended Kalman Filter

The extended Kalman filter (EKF) is a method developed in an effort to apply the Kalman filter to nonlinear dynamical systems models. Within the EKF, updates are performed on linearized approximations of the nonlinear model and work with nonadditive errors. Thus we will generalize Eqs. (1) and (2):

$$x_t = f(x_{t-1}, u_t, \theta^f, \omega_t) \quad (16)$$

$$y_t = h(x_t, \theta^h, \varepsilon_t) \quad (17)$$

As in the Kalman filter, the EKF estimates the prior states by progressing the model forward deterministically:

$$\hat{x}_t^- = f(\hat{x}_{t-1}^+, u_t, \theta^f) \quad (18)$$

where \hat{x}_{t-1}^+ and \hat{x}_t^- are the updated and forecast states at time $t-1$ and t , respectively. The fact that the model is nonlinear makes estimating the Kalman gain significantly more difficult. Application to this system requires linearization of the model, allowing the Kalman update equation to effectively estimate the gradient of the state-observation relationship. To update the states, four partial derivatives are required. These partial derivatives will be taken from each model with respect to the states, as shown in Eqs. (19) and (20), and with respect to the model and observational errors, as shown in Eqs. (21) and (22):

$$A = \frac{df(\cdot)}{dx} \quad (19)$$

$$H = \frac{dh(\cdot)}{dx} \quad (20)$$

$$W = \frac{df(\cdot)}{d\omega} \quad (21)$$

$$V = \frac{dh(\cdot)}{d\varepsilon} \quad (22)$$

The partial derivatives in Eqs. (19), (20), (21), and (22) linearize the model, which allows for application of the linear updating scheme of the Kalman filter. The error covariance is estimated as:

$$P_t = AP_{t-1}A^T + W\Sigma_mW^T, \quad (23)$$

where Σ_m is the state error covariance and based on that the Kalman gain may be estimated similarly to the standard Kalman filter, with a correction to the observation variance:

$$K_t = P_t H^T (H P_t H^T + V \Sigma_o V^T)^{-1} \quad (24)$$

where Σ_o is the observation error covariance. With this Kalman gain, the states may be updated in a way analogous to Eq. (13):

$$\hat{x}_t^+ = \hat{x}_t^- + K_t (y_t - h(\hat{x}_t^-, \theta^h, \varepsilon_t)) \quad (25)$$

The difference between the EKF update and the standard Kalman filter update is that the innovation, $y_t - h(\hat{x}_t^-, \theta^h, \varepsilon_t)$, is calculated from the nonlinear observational operator.

Unscented Kalman Filter

The unscented Kalman filter (UKF) is similar to the EKF, but linearizes around a set of state samples, instead of only one state estimate. The UKF can be thought of as a hybrid between deterministic and ensemble DA techniques. Since the method uses a sampling procedure to propagate uncertainty forward, error characterization is similar to ensemble techniques, but it retains a strictly Gaussian assumption by only updating the state expected value. This means that the posterior is a single deterministic value, representing the mean of the distribution, with a corresponding state error covariance. Due to the deterministic representation of the posterior, it cannot be considered a purely ensemble-based technique.

To apply the sampling strategy, one will estimate multiple sigma points with the model, allowing for calculation of the error covariance from a sample. Each of these sigma points will be used for initialization of the model, as is shown in Eq. (26):

$$\hat{x}_{t,i}^- = f(\hat{x}_{t-1,i}^+, u_t, \theta^f) \quad (26)$$

In Eq. (26), $\hat{x}_{t-1,i}$ is the i th sigma point estimate of the initial states, which is described in Eqs. (27) and (28):

$$\hat{x}_{t-1,1}^- = \hat{x}_{t-1}^- \quad (27)$$

$$\hat{x}_{t-1,i}^- = \begin{cases} \hat{x}_{t-1}^- + \sqrt{n + \lambda} \sqrt{P_{t-1}^-} & \text{if } 1 < i \leq n \\ \hat{x}_{t-1}^- - \sqrt{n + \lambda} \sqrt{P_{t-1}^-} & \text{if } n < i \end{cases} \quad (28)$$

Sigma points are generated to capture the mean and covariance of the state estimates. In Eq. (28), n is the length of the state vector, $\sqrt{P_{t-1}^-}$ is the i th column of the Cholesky decomposition of P_{t-1}^- , and λ is a scaling factor. After initialization of the model with each sigma point, the sigma point for the prior model states at the current time step is available. At this point, the state expected value is estimated according to

Eq. (29), the expected value of the observation forecast is estimated in Eq. (30), and the covariances are calculated according to Eqs. (31) and (32):

$$\hat{x}_t^- = \frac{1}{n+\lambda} \left(\lambda \hat{x}_{t,1}^- + \frac{1}{2} \sum_{i=2}^{2n+1} \hat{x}_{t,i}^- \right) \quad (29)$$

$$\hat{y}_t^- = \frac{1}{n+\lambda} \left(\lambda h(\hat{x}_{t,1}^-, \theta^h) + \frac{1}{2} \sum_{i=2}^{2n+1} h(\hat{x}_{t,i}^-, \theta^h) \right) \quad (30)$$

$$C_{XX} = \frac{1}{n+\lambda} \left(\lambda (\hat{x}_{t,1}^- - \hat{x}_t^-) (\hat{x}_{t,1}^- - \hat{x}_t^-)^T + \frac{1}{2} \sum_{i=2}^{2n+1} (\hat{x}_{t,i}^- - \hat{x}_t^-) (\hat{x}_{t,i}^- - \hat{x}_t^-)^T \right) \quad (31)$$

$$C_{XY} = \frac{1}{n+\lambda} \left(\lambda (\hat{x}_{t,1}^- - \hat{x}_t^-) \left(h(\hat{x}_{t,1}^-, \theta^h) - \hat{y}_t^- \right)^T + \frac{1}{2} \sum_{i=2}^{2n+1} \left(h(\hat{x}_{t,i}^-, \theta^h) - \hat{y}_t^- \right) \left(h(\hat{x}_{t,i}^-, \theta^h) - \hat{y}_t^- \right)^T \right) \quad (32)$$

From the above equations, the covariance of the states ($C_{XX} \approx P_t^- H^T$) and the covariance between the states and observations ($C_{XY} \approx HP_t^- H^T$) are estimated, allowing approximation of the optimal linear update. Following the standard Kalman filter, the Kalman gain is estimated from the covariances, as shown in Eq. (33), and the updated state vector is estimated from Eq. (34):

$$K_t = C_{XY} (C_{YY} + \Sigma_o)^{-1} \quad (33)$$

$$\hat{x}_t = \hat{x}_t^- + K_t (y_t - \hat{y}_t^-) \quad (34)$$

Similar to the EKF, the UKF estimates the proper linear update for the model states, allowing for approximation of the posterior state value of the nonlinear model.

3.2.2 Variational Data Assimilation

The premise of variational DA surrounds the idea of a cost function. A cost function represents errors in the system, which we seek to minimize. Rather than requiring the linearization of nonlinear models, as is performed in the Kalman filter extensions, variational methods rely on optimization tools to find the optimal state values with respect to a predefined cost function. The general form of the cost function, for Gaussian error structures, is shown in Eq. (35):

$$C = (\hat{x}_t - \hat{x}_t^-) \sum_m^{-1} (\hat{x}_t - \hat{x}_t^-) + (y_t - h(\hat{x}_t, \theta^h)) \sum_o^{-1} (y_t - h(\hat{x}_t, \theta^h)_o^-) \quad (35)$$

In Eq. (35), C is the value of the cost function, and all other variables were defined in earlier sections. In this form, the cost function compares the state error and the forecast error, which may be minimized to find the optimal solution to the filtering

problem. Since a solution to the cost function may not be derived analytically, inverse modeling must be performed.

One method for solving the cost function is through iterative optimization techniques. These methods will search the state space for the state values that optimize (minimize) the cost function. This optimal value is considered to be the expected value of the states and therefore the best estimate available for the true states. Although this strategy is effective, it requires multiple evaluations of the model itself, increasing the computational burden. Since this is a deterministic DA method, it is advantageous to avoid multiple model evaluations. In order to achieve this goal, the derivative of the cost function is required:

$$\nabla C = \sum_m^{-1} (\hat{x}_t - \hat{x}_t^-) + J(\hat{x}_t) \sum_o^{-1} (y_t - h(\hat{x}_t, \theta^h)) \quad (36)$$

$J(\hat{x}_t)$ is the Jacobian of the model, also referred to as the adjoint model. This requires finding the partial derivatives of the model with respect to each state. Once the adjoint model is available, Eq. (36) may be used to find the minimum of the cost function by finding the \hat{x}_t vector that satisfies $\nabla C = 0$. Therefore, the primary challenge is developing the adjoint model. This is a separate topic of study, and the reader is referred to Errico (1997). There are also software tools for developing adjoint models, including the Tangent linear and Adjoint Model Compiler (TAMC) (Giering 1997).

Four-dimensional variational DA (4D-Var) is a generalization of the variational filter from Eq. (35) where the time dimension of the observations is taken into account. This creates a smoothing methodology to account for more observations in the cost function. By examining multiple observations simultaneously, more information is available to reduce the state uncertainty, that is, information from observations can be projected backward in time. Through this reduction in uncertainty, more accurate and precise estimates of the model states are expected. The general form of the 4D-Var cost function is:

$$C = (\hat{x}_0 - \hat{x}_0^-) B^{-1} (\hat{x}_0 - \hat{x}_0^-) + \sum_{t=1}^T (y_t - h(\hat{x}_t, \theta^h)) \sum_o^{-1} (y_t - h(\hat{x}_t, \theta^h)) \quad (37)$$

This cost function is applied to all observations over some time period of length of T . Due to the increased information available to the technique, initial state estimates generally become more accurate, and therefore 4D-Var is often preferred to 3D-Var.

3.3 Ensemble Data Assimilation

3.3.1 Ensemble Filters

Ensemble Kalman Filter

Application of the ensemble Kalman filter (EnKF) has become highly popular within the hydrometeorology forecast community. This popularity is due to several factors

including simplicity of application, efficiency of the method, and the explicit treatment of complex and interacting uncertainties in the form of an ensemble. The EnKF is relatively simple to apply, compared with other nonlinear DA techniques, resulting from the use of an ensemble to quantify the covariances required for the Kalman update equation (Evensen 2003). This removes the need to take model derivatives, which is very challenging due to the complexity of hydrologic models. With respect to model efficiency, the assumption of Gaussian error structure has been shown to be reasonable in some applications, which leads to efficient updates of model states. Finally, the ensemble nature of the EnKF explicitly quantifies the uncertainty with the ensemble, where each ensemble member is equally weighted.

Application of the EnKF begins with an ensemble simulation, as described in Eqs. (6) and (7). After performing these simulations, the covariances are estimated directly from the ensembles:

$$E[\hat{x}_t^-] = \frac{1}{N} \sum_{i=1}^N \hat{x}_{t,i}^- \quad (38)$$

$$E[\hat{y}_t] = \frac{1}{N} \sum_{i=1}^N \hat{y}_{t,i} \quad (39)$$

$$\begin{aligned} C_{XY} &= E\left[(\hat{x}_t^- - E[\hat{x}_t^-])(\hat{y}_t - E[\hat{y}_t])^T \right] \\ &= \frac{1}{N} \sum_{i=1}^N \left((\hat{x}_{t,i}^- - E[\hat{x}_t^-])(\hat{y}_{t,i} - E[\hat{y}_t])^T \right) \end{aligned} \quad (40)$$

$$C_{YY} = E\left[(\hat{y}_t - E[\hat{y}_t])(\hat{y}_t - E[\hat{y}_t])^T \right] = \frac{1}{N} \sum_{i=1}^N ((\hat{y}_t - E[\hat{y}_t])(\hat{y}_t - E[\hat{y}_t])) \quad (41)$$

With these covariances, the Kalman gain may be estimated according to:

$$K_t = C_{XY}(C_{YY} + \Sigma_o)^{-1} \quad (42)$$

This formulation of the Kalman gain follows the original Kalman filter, except that all model covariances are approximated with the ensemble. By applying this approximation, there is no need to apply linearization of the model, greatly simplifying the update process. Once the Kalman gain is available, each ensemble member is updated as:

$$\hat{x}_{t,i} = \hat{x}_{t,i}^- + K \left(y_{t,i} - \hat{y}_{t,i} \right) \quad (43)$$

In Eq. (43), $y_{t,i}$ is the i th sample of the observation, which is estimated as follows:

$$y_{t,i} = y_t + \varepsilon_{t,i} \quad \varepsilon_{t,i} \sim N(0, \Sigma_o) \quad (44)$$

The additional error sampling in Eq. (44) is required to account for uncertainty in the observations.

Ensemble Square Root Filter

The ensemble square root filter (EnSRF) was developed to remove the need to perturb the observations in the updates of the EnKF (Whitaker and Hamill 2002). By removing the need to perturb the observation, the necessary ensemble size is reduced, as no sampling of the observation uncertainty is performed. This is achieved by formulating the observation error into the Kalman gain. When explicitly accounting for the observation uncertainty into Eq. (42), the Kalman gain formulation becomes:

$$K_t = C_{XY} \left[\sqrt{C_{YY} + \sum_o}^{-1} \right]^T \left[\sqrt{C_{YY} + \sum_o} + \sqrt{\sum_o} \right] \quad (45)$$

With this formulation of the Kalman gain, each ensemble member may be updated following Eq. (43).

3.3.2 Particle Filters

PFs were developed to overcome the challenges associated with the assumptions required to apply the Kalman-based filters. Although Kalman-based filters have been shown to be effective in many applications, the imposition of a Gaussian error structure can become problematic in some hydrometeorological applications. This scenario motivates the use of an increasingly generalized filter, which can effectively manage skewed, or even multimodal, distributions.

PFs are rooted more firmly in Bayes' theorem than any of the methods we have discussed so far.

According to Eq. (1), the model is known to be Markovian, and therefore the Chapman-Kolmogorov equation may be used to expand the prior probability:

$$p(x_t | y_{1:t-1}) = \int p(x_t | x_{t-1}) p(x_{t-1} | y_{1:t-1}) dx_{t-1} \quad (46)$$

This makes the prior distribution the integration of the transition probability ($p(x_t | x_{t-1})$) and the posterior at the previous time step. Beyond the initial time step, $p(x_{t-1} | y_{1:t-1})$ will be available, and the transition probability may be approximated through the sequential Monte Carlo algorithms (described in following sections). The implied proportionality constant in Eq. (4) comes from the observation probability, which may be expanded as:

$$p(y_t | y_{1:t-1}) = \int p(y_t | x_t) p(x_t | y_{1:t-1}) dx_t \quad (47)$$

Therefore, one may solve the observation probability through the integration of the numerator of Eq. (4). This leads to Eq. (48), where only the likelihood, transition probability, and posterior at the previous time step are required to solve sequential Bayes' theorem:

$$p(x_t|y_{1:t}) = p(x_t|y_t, y_{1:t-1}) = \frac{p(y_t|x_t) \int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}}{\int p(y_t|x_t) [\int p(x_t|x_{t-1})p(x_{t-1}|y_{1:t-1})dx_{t-1}]dx_t} \quad (48)$$

Sequential Importance Sampling

In general, it is often not possible to derive an analytical expression for the Bayesian posterior from Eq. (48), but it is possible to utilize importance sampling to estimate the probabilities sequentially. This is referred to as sequential importance sampling (SIS), which is the most general solution available for the filtering problem (Gordon et al. 1993). SIS relies on a weighted sample of “particles” to estimate the posterior distribution. Similar to Eq. (5), the posterior sample may be represented by Eq. (49):

$$p(x_t|y_t) \approx \sum_{i=1}^N w_{t,i} \delta(x_t - \hat{x}_{t,i}) \quad (49)$$

Sampling directly from the posterior is often not possible, which necessitates importance sampling. This is performed by sampling from a known distribution and weighting based on our knowledge of the system. Posterior importance weights may be estimated according to Eq. (50), where $q(x_{t,i}|x_{t-1,i}, y_t)$ is the importance density:

$$w_{t,i} \propto \frac{p(y_t|\hat{x}_{t,i})p(\hat{x}_{t,i}|\hat{x}_{t-1,i})}{q(\hat{x}_{t,i}|\hat{x}_{t-1,i}, y_t)} \quad (50)$$

The most common and convenient method for developing the importance density is through Eq. (51), where the importance density is set equal to the transition probability:

$$q(\hat{x}_{t,i}|\hat{x}_{t-1,i}, y_t) = p(\hat{x}_{t,i}|\hat{x}_{t-1,i}) \quad (51)$$

Through this choice of importance density, the posterior weights may be estimated as the normalized product of the prior weights and the likelihood, as is shown in Eq. (52).

$$w_{t,i} = \frac{p(y_t|\hat{x}_{t,i})w_{t-1,i}}{\sum_{i=1}^N p(y_t|\hat{x}_{t,i})w_{t-1,i}} \quad (52)$$

This provides a weighted sample at each model time step, representing the posterior distribution of the model states.

Sampling Importance Resampling

Although SIS is the most generalized solution available to solving sequential Bayes' law, it is subject to failures over long simulations. Over a large number of time steps, it is common for many particles to drift from the observations, leading to weight

degeneracy. Weight degeneracy refers to the scenario in which the weight of nearly all particles approach zero, with only a small portion of the sample having significant weight (Arulampalam et al. 2002). During this occurrence, the filter will fail, as it is unable to represent the posterior distribution. Weight degeneracy may be avoided with increasingly large sample size to represent all possible model trajectories, but this becomes limiting as the computational demand increases exponentially with dimensionality. In order to overcome this issue, the novel approach of resampling has been used. This is referred to as sampling importance resampling (SIR), where the sample of particles is resampled, based on the weights of the particles.

When applying SIR, it is common to estimate the effective sample size after each assimilation time step, to determine if resampling is required. The effective sample size may be estimated from Eq. (53) and compared to some predefined threshold:

$$N_{\text{eff}} = \sum_{i=1}^N \frac{1}{w_{t,i}^2} \quad (53)$$

where N_{eff} is the effective sample size. If this value drops below a specified value, it may be assumed that weight degeneracy is occurring, or about to occur, and resampling is required to avoid failure in the filter. At this stage, a representative posterior distribution may be developed by replicating higher probability particles and discarding low probability particles, which is a method referred to as resampling. The challenge when developing a resampling scheme is to create a representative sample of the posterior from current posterior sample. A common method is multinomial resampling.

Multinomial resampling is considered the simplest resampling technique for PFs (Douc and Cappe 2005). The first step in this resampling scheme is the development of the empirical cumulative density from the weights, as described in Eq. (54):

$$F_w(I) = \sum_{i=1}^I w_{t,i} \quad (54)$$

where I is the specified index of the current sample, which may be any integer on the range of $[0,1]$. With this cumulative density, it is possible to sample from the density with the uniformly sampled random variables (U), as described in Eq. (55):

$$I = F_w^{-1}(U) \quad U \sim U(0, 1) \quad (55)$$

where $F_w^{-1}()$ is the inverse of the cumulative density. By putting a uniformly distributed random variable into the inverse cumulative density of the weights, it is possible to extract a corresponding index that should be sampled. This index may then be sampled according to Eq. (56), to develop the corresponding resampled value for ensemble member i ($x_{t,i}^{\text{res}}$):

$$x_{t,i}^{\text{res}} = x_{t,I} \quad (56)$$

A resampled state vector for each of the N particles will be sampled, at which point all weights are set uniformly, shown in Eq. (57), as the density of the particles represents the posterior:

$$w_{t,i} = \frac{1}{N} \quad (57)$$

According to Eq. (1), given that state variables are influenced by the forcing data uncertainty and system noise, the parameters are more susceptible to sample impoverishment as they are not dynamic quantities. To circumvent this issue, Moradkhani et al. (2005b) proposed a method to avoid sample impoverishment by perturbing the resampled parameters:

$$\theta_{t+1,i}^- = \theta_{t,i}^+ + \gamma_{t,i} \quad \gamma_{t,i} \sim N\left[0, s\text{Var}\left(\theta_{t,i}^-\right)\right] \quad (58)$$

where $\theta_{t+1,i}^-$ is the parameter at time step $t + 1$, $\text{Var}\left(\theta_{t,i}^-\right)$ is the variance of the prior parameters, and s is the variance multiplier, which should be tuned.

Despite the successful application of PF-SIR in many hydrologic practices, the convergence of parameters is dependent on the choice of tuning parameter s in the perturbation process. Moradkhani et al. (2012) developed a method to reduce the potential of sample impoverishment in PF-SIR by combining the strength of PF and Markov chain Monte Carlo (MCMC).

Particle Filter-Markov Chain Monte Carlo (PF-MCMC)

The PF-MCMC uses the PF-SIR algorithm to resample the state variables and parameters. Then, a proposal distribution is created to generate parameters $\theta_{t,i}^p$ allowing for larger move steps:

$$\theta_{t,i}^p = \theta_{t,i}^+ + \gamma_{t,i} \quad \gamma_{t,i} \sim N\left[0, s\text{Var}\left(\theta_{t,i}^-\right)\right] \quad (59)$$

where s is the parameter variance tuning factor. To accept or reject the $\theta_{t,i}^{i,p}$ parameters, a metropolis acceptance ratio α is calculated:

$$\alpha = \min\left(1, \frac{p\left(x_{t,i}^p, \theta_{t,i}^p | y_{1:t}\right)}{p\left(x_{t,i}^+, \theta_{t,i}^+ | y_{1:t}\right)}\right) \quad (60)$$

where $p\left(x_{t,i}^p, \theta_{t,i}^p | y_{1:t}\right)$ is the proposed joint probability distribution:

$$p\left(x_{t,i}^p, \theta_{t,i}^p | y_{1:t}\right) \propto p\left(y_{1:t} | x_{t,i}^p, \theta_{t,i}^p\right) \cdot p\left(x_{t,i}^p | \theta_{t,i}^p, y_{1:t-1}\right) \cdot p\left(\theta_{t,i}^p | y_{1:t-1}\right) \quad (61)$$

$$x_{t,i}^p = f\left(x_{t-1,i}^+, u_{t,i}, \theta_{t,i}^p\right) \quad (62)$$

where $x_{t,i}^p$ is a sample from the proposal state distribution at time step t .

Since the optimal tuning factor s is unknown in a sequential framework, it is beneficial to treat the s as a time-varying parameter and estimate it automatically. Moradkhani et al. (2012) modified the variable variance multiplier (VVM) method proposed by Leisenring and Moradkhani (2012) to automatically obtain the most fitting tuning factor s in Eq. 59.

Evolutionary PF-MCMC (EPFM)

The EPFM was proposed by Abbaszadeh et al. (2018) to characterize a more accurate and reliable posterior distribution for state variables in data assimilation applications. What distinguishes the EPFM approach from the PF-MCMC is the utilization of hybrid genetic algorithm (GA) and MCMC (GA-MCMC) technique in the importance sampling step of the PF-MCMC model. In fact, the GA-MCMC expands the search space by implementing the crossover and mutation steps in the GA, and subsequently the search space is refined via the MCMC technique resulting in more desirable prior distribution. This approach significantly minimizes the particle degeneracy and sample impoverishment problems that have been the main concerns in using the particle filters. The main structures of EPFM approach are summarized below:

1. Particles are selected from the initial ensemble pool for the crossover operation. To do this, one can use roulette wheel selection method, and a fitness value for each ensemble member is assigned. The value of weights, as an appropriate indication of ensemble member quality, can be directly used as the fitness value.
2. The arithmetic crossover is adopted for the crossover operation (Park et al. 2009; Yin and Zhu 2015). For this, a pair of new particles (offspring) is generated by a linear combination of a pair of selected particles in step 1:

$$x_{t-1}^{j'} = \xi \cdot x_t^i + (1 - \xi) \cdot x_{t-1}^j \quad (63)$$

$$x_{t-1}^{j'} = (1 - \xi) \cdot x_{t-1}^i + \xi \cdot x_{t-1}^j \quad (64)$$

where x_{t-1}^i and x_{t-1}^j are the parent particles, $x_{t-1}^{j'}$ and $x_{t-1}^{j''}$ are the pair of new offspring particle, and ξ is a uniform random value in the range of $[0, 1]$.

3. To further promote diversity of the particles, a mutation strategy is designed. It is realized by Eq. 65 that x_{t-1}^k and $x_{t-1}^{k'}$ are the particles before and after mutation process, respectively:

$$x_{t-1}^{k'} = x_{t-1}^k + \eta \quad x_{t-1}^k \in \{x_{t-1}^{j'}, x_{t-1}^{j''}\} \quad \eta \sim N(0, \psi \text{Var}(x_{t-1}^{k-})) \quad (65)$$

where η represents a random sample from a Gaussian distribution with mean zero and variance $\psi \text{Var}(x_{t-1}^{k-})$, where $\text{Var}(x_{t-1}^{k-})$ is the variance of the prior states at the time $t - 1$ and ψ is a small tuning parameter.

4. The MCMC algorithm is used to accept or reject the new particles generated by GA operators. This step is similar to the one used in the PF-MCMC model.

4 Applications

4.1 Variational

Variational methods have become popular for atmospheric DA, but are less popular in land surface applications. Although applicable to both, the atmospheric community has more readily developed the adjoint models necessary for variational methods, whereas the land surface community has generally relied on ensemble methods. Within the atmospheric DA community, several examples of variational DA applications are available (Barker et al. 2004; Dee et al. 2011; Hou et al. 2013; Županski and Mesinger 1995). Although uncommon, there are a few examples of variational DA within the hydrologic community (Reichle et al. 2001; Seo et al. 2003). A few recent examples of land surface DA with variational methods include Hoppe et al. (2014), which assimilated soil moisture and temperature measurements into the Community Land Model (CLM); Meng et al. (2009), which assimilated land surface temperature into the CLM; and Lee et al. (2012), which assimilated streamflow into the Sacramento Soil Moisture Accounting model.

4.2 Kalman-Based Filters

Kalman-based filters are the most commonly used DA methods in hydrometeorology. This is a result of the combination of ease of application and effectiveness of the technique. Of particular focus have been ensemble applications of the Kalman filter. In the atmospheric and land surface DA communities alike, the EnKF has been widely applied. Although variational methods would be advantageous from an efficiency perspective, the development of an adjoint model can be challenging for highly nonlinear models, making ensemble methods attractive. Alternatively, PFs are highly robust estimators of the posterior distribution, but are subject to failure in small ensemble sizes. This makes the EnKF a useful tool in large dimensional problems. For atmospheric DA, applications typically involve observations of wind speed, wind direction, temperature, and humidity from radiosondes and satellites (Annan et al. 2005; Houtekamer and Mitchell 1998; Lorenc 2003). For land surface and hydrologic DA, the observations and applications are much more varied. Applications include soil moisture (e.g., Kumar et al. 2014; Reichle et al. 2002; De Lannoy et al. 2007; De Rosnay et al. 2013), passive microwave brightness temperature (Crow and Wood 2003; DeChant and Moradkhani 2011b; Durand and Margulis 2008), snow cover fraction (Andreadis and Lettenmaier 2006; Slater and Clark 2005), snow water equivalent (De Lannoy et al. 2012; Leisenring and Moradkhani 2011; Liu et al. 2012), streamflow (Clark et al. 2008; Moradkhani et al. 2005a; Noh et al. 2011; Samuel et al. 2014), and consideration of nonstationarity in dynamic catchments (Pathiraja et al. 2016a, b).

4.3 Particle Filters

The DA community has been slow to adopt the PF methods, primarily due to the understanding that PFs are overly demanding computationally (Snyder et al. 2008).

The “Curse of Dimensionality” has been termed to designate the exponential scaling of necessary sample size for estimating the posterior with increasing degrees of freedom in the system (Bengtsson et al. 2008). Although this criticism has shown that certain PFs are subject to failure in large-scale systems, PFs are gaining popularity in many applications. Due to improvements in filter efficiency, and the identification of applicable problems, the PF has become a viable method for DA (Moradkhani et al. 2012). With the introduction of PF to hydrologic community, the PFs have gained a considerable attention in a variety of land surface applications by assimilating variables including streamflow (e.g., Moradkhani et al. 2005b, 2012; Weerts and El Serafy 2006; DeChant and Moradkhani 2011a; Yan and Moradkhani 2016; Abbaszadeh et al. 2018), soil moisture (Montzka et al. 2011, 2013; Guingla et al. 2012; Yan et al. 2015, 2017), snow water equivalent (Leisenring and Moradkhani 2011; DeChant and Moradkhani 2011b, 2012), sediment load (Leisenring and Moradkhani 2012), flood inundation (Matgen et al. 2011; Plaza et al. 2012), and multi-modeling (Parrish et al. 2012; DeChant and Moradkhani 2014a).

4.4 Parameter Inference and Model Structures

In addition to the methods described above, there are several DA methods in hydrology and from other branches of scientific literature that focus on more holistic treatments of the problem of reducing uncertainties in dynamical systems models by probabilistically conditioning model states on observations. Perhaps most notably, there have been several applications of various methods that simultaneously estimate model parameters and model states (e.g., Moradkhani et al. 2005a, b, 2012; DeChant and Moradkhani 2012; Smith et al. 2013; Ruiz and Pulido 2015; Gharamti et al. 2017; Abbaszadeh et al. 2018).

The sequential estimation of temporally varying model parameters through DA can also be used to improve hydrologic forecasting in systems with changing catchment properties (such as land use or land cover change) (Pathiraja et al. 2016a, b). Such a time-varying parameter framework can be useful whenever the catchment system is undergoing change in real time, that may be unknown to the modeler. Model parameters are sequentially updated in response to signals of change in observations, such that the model is improved as soon as an information about a change becomes available. This can be done through a joint state-parameter estimation DA framework with a careful choice on the parameter evolution model, i.e., the method for generating prior distributions of the parameters at each time (Pathiraja et al. 2016b). Additionally, the choice of the model structure itself is critical in ensuring that such a time-varying parameter framework can be useful under changing conditions. Specifically, the model structure must be sufficiently flexible so that it can represent the range of possible future changes to catchment conditions (Pathiraja et al. 2018b). In other words, the entire feasible parameter space and forcings must produce model states and outputs that capture all possible future outcomes.

Additionally, state-updating DA has been used to help understand complex model error distributions and to update model structures (Bulygina and Gupta 2009, 2010, 2011; Wilkinson et al. 2011; Nearing and Gupta 2015; Nearing et al. 2013). It is often difficult to use observation data to directly infer structural errors in complex systems models, because we often do not have observations related to all of the interacting biogeophysical processes in a watershed or other hydrologic system. Careful applications of DA can be used to update the internal states of the model given whatever partial observations are available, so long as the uncertainty due to the model structure is appropriately quantified (Pathiraja et al. 2018a).

5 Conclusion

DA techniques are valuable tools for estimating initial conditions for hydrometeorological forecasts. Due to the uncertainties in initial conditions (DeChant and Moradkhani 2011a, 2014a), it is necessary to quantify and reduce these uncertainties, and DA is widely seen as the forefront of the science in performing this task. With developments in DA science in the last two decades, several assimilation techniques are becoming standard tools for quantifying and reducing model uncertainty. These tools have seen wide ranging applications, particularly in simulating atmospheric and land surface processes.

Although DA is becoming a standard set of tools, the variety of techniques requires significant thought in determining the proper technique for a given application. If the underlying system is linear, or nearly linear, the Kalman filter will likely be chosen as it will be an optimal filter. Alternatively, in highly nonlinear problems, which are the norm in hydrometeorology, the choice in technique becomes much more difficult. Choosing between deterministic and ensemble techniques is a challenge. Although generalized techniques (i.e., PFs) are preferred from a theoretical perspective, they require the ability to execute the model enough times to sample from the posterior. Assuming that a large number of simulations are possible, it is likely that the PF will be preferred. If the model is highly computationally demanding, this may be infeasible, and therefore it may be impossible to fully represent the posterior distribution. In this scenario, the EnKF, and similar methods, may be used to retrieve the expected value with smaller ensemble size. Another option is variational DA, which may also retrieve the expected value. Variational methods bring the challenge of requiring an adjoint model, but are very competitive with the EnKF. Overall, each DA technique will have benefits and drawbacks, which often makes the choice of technique situation specific.

References

- P. Abbaszadeh, H. Moradkhani, H. Yan, Enhancing hydrologic data assimilation by evolutionary Particle Filter and Markov Chain Monte Carlo method. *Adv. Water Resour.* **111**, 192–204 (2018). <https://doi.org/10.1016/j.advwatres.2017.11.011>

- K.M. Andreadis, D.P. Lettenmaier, Assimilating remotely sensed snow observations into a macro-scale hydrology model. *Adv. Water Resour.* **29**, 872–886 (2006)
- J.D. Annan, J.C. Hargreaves, N.R. Edwards, R. Marsh, Parameter estimation in an intermediate complexity Earth system model using an ensemble Kalman filter. *Ocean Model.* **8**(1), 135–154 (2005)
- M.S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking. *IEEE Trans. Signal Process.* **50**(2), 174–188 (2002)
- D.M. Barker, W. Huang, Y.-R. Guo, A.J. Bourgeois, Q.N. Xiao, A three-dimensional variational data assimilation system for MM5: implementation and initial results. *Mon. Weather Rev.* **132**(4), 897–914 (2004)
- T. Bengtsson, P. Bickel, B. Li, Curse of dimensionality revisited: the collapse of importance sampling in very large scale systems, in *IMS Collections: Probability and Statistics: Essays in Honor of David A. Freedman*, vol. 2, ed. by D. Nolan, T. Speed (Institute of Mathematical Statistics, Beachwood), pp. 316–334 (2008)
- N. Bulygina, H. Gupta, Estimating the uncertain mathematical structure of a water balance model via Bayesian data assimilation. *Water Resour. Res.* **45**(12), W00B13 (2009). <https://doi.org/10.1029/2007WR006749>
- N. Bulygina, H. Gupta, How Bayesian data assimilation can be used to estimate the mathematical structure of a model. *Stoch. Environ. Res. Risk Assess.* **24**(6), 925 (2010). <https://doi.org/10.1007/s00477-0010-00387-y>
- N. Bulygina, H. Gupta, Correcting the mathematical structure of a hydrological model via Bayesian data assimilation. *Water Resour. Res.* **47**(5), W05514 (2011). <https://doi.org/10.1029/2010WR009614>
- M.P. Clark, D.E. Rupp, R.A. Woods, X. Zheng, R.P. Ibbitt, A.G. Slater, J. Schmidt, M.J. Uddstrom, Hydrological data assimilation with the ensemble Kalman filter: use of streamflow observations to update states in a distributed hydrological model. *Adv. Water Resour.* **31**, 1309 (2008)
- W.T. Crow, E.F. Wood, The assimilation of remotely sensed soil brightness temperature imagery into a land surface model using ensemble Kalman filtering: a case study based on ESTAR measurements during SGP97. *Adv. Water Resour.* **26**(2), 137–149 (2003)
- G.J.M. De Lannoy, R.H. Reichle, P.R. Houser, V.R.N. Pauwels, N.E.C. Verhoest, Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter. *Water Resour. Res.* **43**, W09410 (2007). <https://doi.org/10.1029/2006WR00544>
- G.J.M. De Lannoy, R.H. Reichle, K.R. Arsenault, P.R. Houser, S. Kumar, N.E.C. Verhoest, V. Pauwels, Multiscale assimilation of advanced microwave scanning radiometer–EOS snow water equivalent and moderate resolution imaging spectroradiometer snow cover fraction observations in northern Colorado. *Water Resour. Res.* **48**, W01522 (2012). <https://doi.org/10.1029/2011WR010588>
- P. De Rosnay, M. Drusch, D. Vasiljevic, G. Balsamo, C. Albergel, L. Isaksen, A simplified Extended Kalman Filter for the global operational soil moisture analysis at ECMWF. *Q. J. R. Meteorol. Soc.* **139**(674), 1199–1213 (2013). <https://doi.org/10.1002/qj.2023>
- C. DeChant, H. Moradkhani, Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation. *Hydrol. Earth Syst. Sci.* **15**, 3399–3410 (2011a). <https://doi.org/10.5194/hess-15-3399>
- C. DeChant, H. Moradkhani, Radiance data assimilation for operational snow and streamflow forecasting. *Adv. Water Resour.* **34**(3), 351–364 (2011b)
- C.M. DeChant, H. Moradkhani, Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting. *Water Resour. Res.* **48**(4), W04518 (2012)
- C.M. DeChant, H. Moradkhani, Toward a reliable prediction of seasonal forecast uncertainty: addressing model and initial condition uncertainty with ensemble data assimilation and sequential Bayesian combination. *J. Hydrol.* **519**, 2967–2977 (2014a). <https://doi.org/10.1016/j.jhydrol.2014.05.045>. Special issue on Ensemble Forecasting and data assimilation

- C.M. DeChant, H. Moradkhani, Hydrologic prediction and uncertainty quantification, in *Handbook of Engineering Hydrology, Modeling, Climate Change and Variability* (CRC Press, Taylor & Francis Group, Boca Raton, 2014b), pp. 387–414
- D.P. Dee et al., The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137**(656), 553–597 (2011)
- R. Douc, O. Cappe, Comparison of resampling schemes for particle filtering, paper presented at image and signal processing and analysis, 2005. ISPA 2005, in *Proceedings of the 4th International Symposium on, 15–17 Sept 2005* (2005)
- M. Durand, S.A. Margulis, Effects of uncertainty magnitude and accuracy on assimilation of multiscale measurements for snowpack characterization. *J. Geophys. Res.* **113**(D2), D02105 (2008)
- R.M. Errico, What is an adjoint model? *Bull. Am. Meteorol. Soc.* **78**(11), 2577–2591 (1997)
- G. Evensen, The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**(4), 343–367 (2003)
- Z. Ghahramani, S.T. Roweis, Learning nonlinear dynamical systems using an EM algorithm. *Adv. Neural Inf. Process. Syst.* **11**, 431–437 (1999)
- M.E. Gharami, J. Tjiputra, I. Bethke, A. Samuelsen, I. Skjelvan, M. Bentsen, L. Bertino, Ensemble data assimilation for ocean biogeochemical state and parameter estimation at different sites. *Ocean Model.* **112**, 65–89 (2017)
- R. Giering, *Tangent Linear and Adjoint Model Compiler, Users Manual* (Center for Global Change Sciences, Department of Earth, Atmospheric, and Planetary Science. MIT, Cambridge, 1997)
- N. Gordon, D. Salmond, A. Smith, Novel approach to nonlinear/non-Gaussian Bayesian state estimation. *Proc. Inst. Elect. Eng. F.* **140**(2), 107–113 (1993)
- P. Guingla, D. Antonio, R. De Keyser, G. De Lannoy, L. Giustarini, P. Matgen, V. Pauwels, The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter. *Hydrol. Earth Syst. Sci.* **16**(2), 375–390 (2012)
- C.M. Hoppe, H. Elbern, J. Schwinger, A variational data assimilation system for soil–atmosphere flux estimates for the Community Land Model (CLM3. 5). *Geosci. Model Dev.* **7**(3), 1025–1036 (2014)
- T. Hou, F. Kong, X. Chen, H. Lei, Impact of 3DVAR data assimilation on the prediction of heavy rainfall over Southern China. *Adv. Meteorol.* **2013**, 1 (2013)
- P.L. Houtekamer, H.L. Mitchell, Data assimilation using an ensemble Kalman filter technique. *Mon. Weather Rev.* **126**(3), 796–811 (1998)
- R.E. Kalman, A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.* **82**(Series D), 35–45 (1960). <https://doi.org/10.1115/1111.3662552>
- S. Kumar, C. Peters-Lidard, D. Mocko, R. Reichle, Y. Liu, K. Arsenault, Y. Xia, M. Ek, G. Riggs, B. Livneh, M Cosh, Assimilation of remotely sensed soil moisture and snow depth retrievals for drought estimation. *J. Hydrometeorol.* **15**, 2446–2469 (2014). <https://doi.org/10.1175/JHM-D-13-0132.1>
- H. Lee, D.J. Seo, Y. Liu, V. Koren, P. McKee, R. Corby, Variational assimilation of streamflow into operational distributed hydrologic models: effect of spatiotemporal scale of adjustment. *Hydrol. Earth Syst. Sci.* **16**(7), 2233–2251 (2012)
- M. Leisenring, H. Moradkhani, Snow water equivalent prediction using Bayesian data assimilation methods. *Stoch. Environ. Res. Risk Assess.* **25**(2), 253–270 (2011)
- M. Leisenring, H. Moradkhani, Analyzing the uncertainty of suspended sediment load prediction using sequential Monte Carlo methods. *J. Hydrol.* **468–469**, 268–282 (2012). <https://doi.org/10.1016/j.jhydrol.2012.08.049>
- Y. Liu, A.H. Weerts, M. Clark, H.J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.J. Seo, D. Schwanenberg, P. Smith, A.I.J.M. van Dijk, N. van Velzen, M. He, H. Lee, S.J. Noh, O. Rakovec, P. Restrepo, Toward advancing data assimilation in operational hydrologic forecasting and water resources management: current status, challenges, and emerging opportunities. *Hydrol. Earth Syst. Sci.* **16**, 3863–3887 (2012)

- A.C. Lorenc, The potential of the ensemble Kalman filter for NWP – a comparison with 4D-Var. *Q. J. R. Meteorol. Soc.* **129**(595), 3183–3203 (2003)
- P. Matgen, R. Hostache, G. Schumann, L. Pfister, L. Hoffmann, H.H.G. Savenje, Towards an automated SAR-based flood monitoring system, Lessons learned from two case studies. *Phys. Chem. Earth.* **36**(7–8), 241–252 (2011). <https://doi.org/10.1016/j.pce.2010.12.009>
- C.L. Meng, Z.L. Li, X. Zhan, J.C. Shi, C. Y. Liu, Land surface temperature data assimilation and its impact on evapotranspiration estimates from the Common Land Model. *Water Resour. Res.* **45**, W02421 (2009). <https://doi.org/10.1029/2008WR006971>
- C. Montzka, H. Moradkhani, L. Weihermüller, H.J. Hendricks Franssen, M. Canty, H. Vereecken, Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter. *J. Hydrol.* **399**(3–4), 410–421 (2011). <https://doi.org/10.1016/j.jhydrol.2011.01.020>
- C. Montzka, J. Grant, H. Moradkhani, H.J. Hendricks Franssen, L. Weihermüller, M. Drusch, H. Vereecken, Estimation of radiative transfer parameters from L-Band passive microwave brightness temperatures using data assimilation. *Vadose Zone Hydrol. Special Issue of Remote Sensing.* (2013). <https://doi.org/10.2136/vzj2012.0040>
- H. Moradkhani, S. Sorooshian, H.V. Gupta, P.R. Houser, Dual state-parameter estimation of hydrological models using ensemble Kalman filter. *Adv. Water Resour.* **28**(2), 135–147 (2005a)
- H. Moradkhani, K.L. Hsu, H. Gupta, S. Sorooshian, Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter. *Water Resour. Res.* **41**, W05012 (2005b)
- H. Moradkhani, C.M. DeChant, S. Sorooshian, Evolution of ensemble data assimilation for uncertainty quantification using the Particle Filter-Markov Chain Monte Carlo method. *Water Resour. Res.* **48**, W12520 (2012). <https://doi.org/10.1029/2012WR012144>
- G.S. Nearing, H.V. Gupta, The quantity and quality of information in hydrologic models. *Water Resour. Res.* **51**(1), 524–538 (2015)
- G.S. Nearing, H.V. Gupta, W.T. Crow, Information loss in approximately bayesian estimation techniques: a comparison of generative and discriminative approaches to estimating agricultural productivity. *J. Hydrol.* **507**, 163–173 (2013)
- S.J. Noh, Y. Tachikawa, M. Shiiba, S. Kim, Applying sequential Monte Carlo methods into a distributed hydrologic model: lagged particle filtering approach with regularization. *Hydrol. Earth Syst. Sci.* **15**(10), 3237 (2011)
- S. Park, J.P. Hwang, E. Kim, H. Kang, A new evolutionary particle filter for the prevention of sample impoverishment. *IEEE Trans. Signal Process.* **13**(4), 801–809 (2009)
- M. Parrish, H. Moradkhani, C.M. DeChant, Towards reduction of model uncertainty: integration of Bayesian model averaging and data assimilation. *Water Resour. Res.* **48**, W03519 (2012). <https://doi.org/10.1029/2011WR011116>
- S. Pathiraja, L. Marshall, A. Sharma, H. Moradkhani, Detecting non-stationary hydrologic model parameters in a paired catchment system using data assimilation. *Adv. Water Resour.* **94**, 103–119 (2016a). <https://doi.org/10.1016/j.advwatres.2016.04.021>
- S. Pathiraja, L. Marshall, A. Sharma, H. Moradkhani, Hydrologic modeling in dynamic catchments: a data assimilation approach. *Water Resour. Res.* (2016b). <https://doi.org/10.1002/2015WR017192>
- S. Pathiraja, D. Anghileri, P. Burlando, A. Sharma, L. Marshall, H. Moradkhani, Time varying parameter models for catchments with land use change: the importance of model structure. *Hydrol. Earth Syst. Sci. Discuss.* (2017). <https://doi.org/10.5194/hess-2017-382>
- S. Pathiraja, H. Moradkhani, L. Marshall, A. Sharma, G. Geenens, Data driven model uncertainty estimation in data assimilation. *Water Resour. Res.* (2018a). <https://doi.org/10.1002/2018WR022627>
- S. Pathiraja, D. Anghileri, P. Burlando, A. Sharma, L. Marshall, H. Moradkhani, Insights on the impact of systematic model errors on data assimilation performance in changing catchments. *Adv. Water Resour.* (2018b). <https://doi.org/10.1016/j.advwatres.2017.12.006>

- D.A. Plaza, R. De Keyser, G.J.M. De Lannoy, L. Giustarini, P. Matgen, V.R.N. Pauwels, The importance of parameter resampling for soil moisture data assimilation into hydrologic models using the particle filter. *Hydrol. Earth Syst. Sci.* **16**(2), 375–390 (2012)
- R.H. Reichle, D. Entekhabi, D.B. McLaughlin, Downscaling of radio brightness measurements for soil moisture estimation: a four-dimensional variational data assimilation approach. *Water Resour. Res.* **37**(9), 2353–2364 (2001)
- R.H. Reichle, D.B. McLaughlin, D. Entekhabi, Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Weather Rev.* **130**(1), 103–114 (2002)
- J. Ruiz, M. Pulido, Parameter estimation using ensemble-based data assimilation in the presence of model error. *Mon. Weather Rev.* **143**(5), 1568–1582 (2015)
- P. Salamon, L. Feyen, Assessing parameter, precipitation, and predictive uncertainty in a distributed hydrological model using sequential data assimilation with the particle filter. *J. Hydrol.* **376**(3), 428–442 (2009)
- J. Samuel, P. Coulibaly, G. Dumedah, H. Moradkhani, Assessing model state variation in hydrologic data assimilation. *J. Hydrol.* **513**, 127–141 (2014). <https://doi.org/10.1016/j.jhydrol.2014.03.048>
- D.-J. Seo, V. Koren, N. Cajina, Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting. *J. Hydrometeorol.* **4**(3), 627–641 (2003)
- D.J. Seo, Y. Liu, H. Moradkhani, A. Weerts, Ensemble prediction and data assimilation for operational hydrology. *J. Hydrol.* **519**, 2661–2662 (2014). <https://doi.org/10.1016/j.jhydrol.2014.11.035>
- A.G. Slater, M.P. Clark, Snow data assimilation via an ensemble Kalman filter. *J. Hydrometeorol.* **7**, 478 (2005)
- P.J. Smith, G.D. Thornhill, S.L. Dance, A.S. Lawless, D.C. Mason, N.K. Nichols, Data assimilation for state and parameter estimation: application to morphodynamic modelling. *Q. J. R. Meteorol. Soc.* **139**(671), 314–327 (2013)
- C. Snyder, T. Bengtsson, P. Bickel, J. Anderson, Obstacles to high-dimensional particle filtering. *Mon. Weather Rev.* **136**(12), 4629 (2008)
- J.A. Vrugt, C.G.H. Diks, H.V. Gupta, W. Bouten, J.M. Verstraten, Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. *Water Resour. Res.* **41**(1), W01017 (2005). <https://doi.org/10.1029/2004WR003059>
- A.H. Weerts, G.Y.H. El Serafy, Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resour. Res.* **42**, W09403 (2006). <https://doi.org/10.1029/2005WR004093>
- J.S. Whitaker, T.M. Hamill, Ensemble data assimilation without perturbed observations. *Monthly Weather Rev.* **130**(7), 1913–1924 (2002). [https://doi.org/10.1175/1520-0493\(2002\)130<1913:EDAWPO>2.0.CO;2](https://doi.org/10.1175/1520-0493(2002)130<1913:EDAWPO>2.0.CO;2)
- R.D. Wilkinson, M. Vrettas, D. Cornford, J.E. Oakley, Quantifying simulator discrepancy in discrete-time dynamical simulators. *J. Agric. Biol. Environ. Stat.* **16**(4), 554–570 (2011)
- H. Yan, H. Moradkhani, Combined assimilation of streamflow and satellite soil moisture with the particle filter and geostatistical modeling. *Adv. Water Resour.* **94**, 364–378 (2016). <https://doi.org/10.1016/j.advwatres.2016.06.002>
- H. Yan, C.M. DeChant, H. Moradkhani, Improving soil moisture profile prediction with the Particle Filter-Markov Chain Monte Carlo method. *IEEE Trans. Geosci. Remote Sens.* (2015). <https://doi.org/10.1109/TGRS.2015.2432067>
- H. Yan, H. Moradkhani, M. Zarekarizi, A probabilistic drought forecasting framework: a combined dynamical and statistical approach. *J. Hydrol.* **548**, 291–304 (2017). <https://doi.org/10.1016/j.jhydrol.2017.03.004>
- S. Yin, X. Zhu, Intelligent particle filter and its application to fault detection of nonlinear systems. *IEEE Trans. Ind. Electron.* **62**(6), 3852–3861 (2015)
- D.a. Županski, F. Mesinger, Four-dimensional variational assimilation of precipitation data. *Mon. Weather Rev.* **123**(4), 1112–1127 (1995)



Soil Moisture Data Assimilation

Gabrielle Jacinthe Maria De Lannoy, Patricia de Rosnay, and
Rolf Helmut Reichle

Contents

1	Introduction	703
2	Components of a Soil Moisture Data Assimilation System	704
3	Observations Related to Soil Moisture	706
4	Land Surface Modeling	708
5	Observation Operator	709
5.1	Screen-Level Observation Predictions	709
5.2	Microwave Observation Predictions	710
5.3	Terrestrial Water Storage Predictions	712
6	Assimilation of Observations Related to Soil Moisture	712
6.1	Sequential State Updating	712
6.2	Smoothing	720
6.3	Joint State and Parameter Updating	721
7	Random Errors and Biases	722
7.1	Bias, Autocorrelated Error	722
7.2	Random Error	725
8	Evaluation of Soil Moisture Estimates from Data Assimilation	729
8.1	In Situ Soil Moisture	729

G. J. M. De Lannoy (✉)

NASA Goddard Space Flight Center, Code 610.1, Greenbelt, MD, USA

KU Leuven, Department of Earth and Environmental Sciences, Leuven, Belgium

e-mail: Gabrielle.DeLannoy@kuleuven.be

P. de Rosnay (✉)

Data Assimilation Section, European Center for Medium-Range Weather Forecasts, Reading,
Berkshire, UK

e-mail: patricia.rosnay@ecmwf.int

R. H. Reichle (✉)

NASA Goddard Space Flight Center, Code 610.1, Greenbelt, MD, USA

e-mail: rolf.reichle@nasa.gov

© This is a U.S. government work and not under copyright protection in the U.S.;
foreign copyright protection may apply 2019

701

Q. Duan et al. (eds.), *Handbook of Hydrometeorological Ensemble Forecasting*,
https://doi.org/10.1007/978-3-642-39925-1_32

8.2 Validation Metrics	730
8.3 Example	732
9 Toward Operational Soil Moisture Data Assimilation	733
9.1 ECMWF Soil Moisture Data Assimilation for NWP	734
9.2 NASA SMAP Surface and Root-Zone Soil Moisture Product	736
10 Conclusions	739
References	740

Abstract

Accurate knowledge of soil moisture at the continental scale is important for improving predictions of weather, agricultural productivity, and natural hazards, but observations of soil moisture at such scales are limited to indirect measurements, either obtained through satellite remote sensing or from meteorological networks. Land surface models simulate soil moisture processes, using observation-based meteorological forcing data, and auxiliary information about soil, terrain, and vegetation characteristics. Enhanced estimates of soil moisture and other land surface variables, along with their uncertainty, can be obtained by assimilating observations of soil moisture into land surface models. These assimilation results are of direct relevance for the initialization of hydrometeorological ensemble forecasting systems. The success of the assimilation depends on the choice of the assimilation technique, the nature of the model and the assimilated observations, and, most importantly, the characterization of model and observation error. Systematic differences between satellite-based microwave observations or satellite-retrieved soil moisture and their simulated counterparts require special attention. Other challenges include inferring root-zone soil moisture information from observations that pertain to a shallow surface soil layer, propagating information to unobserved areas and downscaling of coarse information to finer-scale soil moisture estimates. This chapter summarizes state-of-the-art solutions to these issues with conceptual data assimilation examples, using techniques ranging from simplified optimal interpolation to spatial ensemble Kalman filtering. In addition, operational soil moisture assimilation systems are discussed that support numerical weather prediction at ECMWF and provide value-added soil moisture products for the NASA Soil Moisture Active Passive mission.

Keywords

Soil moisture retrieval · Microwave brightness temperature · Radar backscatter · Terrestrial water storage · Analysis · Innovation · Increment · Kalman filter · Observation operator · Numerical weather prediction · Initialization · State update · Calibration · Radiative transfer model · Land surface model · Screen-level observations · ASCAT · AMSR2 · SMOS · SMAP · GRACE

1 Introduction

Soil moisture is the quantity of water contained in the upper layers of the soil, water that directly interacts with the atmosphere through evapotranspiration and partitions rainfall into infiltration and runoff. While the exact definition of soil moisture is slightly different for weather forecasters, farmers, water managers, construction engineers, and ecologists, in hydrological and Earth system applications, soil moisture usually includes the water in the upper ~1–2 m of the soil. Even though soil moisture represents only 0.0012% of all water available on Earth (~ 1.4 billion km³), and only 0.05% of all freshwater (~ 35,000 km³) (Gleick 1996), it is of primary importance because it links the water, energy, and carbon cycles and therefore has a significant impact on weather and global climate (Dirmeyer 2000; Koster et al. 2004), and it controls droughts and floods, agricultural yield, diseases, and other socioeconomic phenomena.

Soil moisture is being monitored with modeling and observing systems. Numerical models provide spatially and temporally continuous estimates of soil moisture at customized space and time resolutions, and at various soil depths down to the water table. However, models are always simplified and prone to errors. Observations are usually obtained from sparse in situ networks or satellite swaths, and they are therefore limited in their spatial and temporal coverage. One particularly important limitation of current remote-sensing observations is that they only provide soil moisture in a shallow surface layer. By using models and observations synergistically, observational gaps can be filled and superior estimates of soil moisture can be constructed. This process, known as data assimilation, merges the observations into soil moisture modeling systems, either to improve model simulations or to add value to observations.

The focus of this chapter is on updating the soil moisture state in dynamic land surface models for continental and global applications. At these scales, observations related to soil moisture are mainly provided by global surface observational networks or through remote sensing, either in the form of satellite retrievals, microwave radiances, or radar backscatter values. Data assimilation interpolates and extrapolates the observations in space and time and updates the entire simulated soil moisture column, while ensuring consistency with all other geophysical variables in the model. This process leads to improved initial conditions for subsequent hydrometeorological forecasts across a range of applications such as weather and drought forecasts. In addition, meaningful estimates of the uncertainty in soil moisture estimates can be obtained, provided the characteristics of observation and forecast error are well described in the assimilation system. These uncertainty estimates can be used to quantify the uncertainty in hydrometeorological ensemble forecasts (chapters ▶ “Ensemble Methods for Meteorological Predictions” and ▶ “Hydrological Ensemble Prediction Systems Around the Globe”).

A broader definition of “data assimilation” methods aimed at improving soil moisture estimates might include the construction of improved forcing information, the revision of modeling systems, or the estimation of model parameters (Part V,

“Model Parameter Estimation and Uncertainty Analysis”) using in situ or satellite observations. Parameter estimation is crucial to limit biases in the modeling part of the assimilation system. This mostly involves the calibration of static model parameters against historical datasets. The dynamic estimation of evolving parameters, possibly along with soil moisture updates, has also been explored, but not for large-scale soil moisture modeling systems. The use of observational information to construct superior forcing information and improve modeling systems is at the core of land surface reanalysis products generated by various operational groups (Sect. 9).

The objective of this chapter is to educate students and scientists new to the discipline about the current, well-established, practices in large-scale soil moisture data assimilation, without providing an in-depth literature review. The chapter first provides an overview of the components of a soil moisture data assimilation system (Sect. 2). Next, the assimilated observations (Sect. 3), the soil moisture modeling (Sect. 4) and the modeling of observation predictions (Sect. 5) are discussed. State-of-the-art assimilation techniques for soil moisture state updating are presented in Sect. 6, with attention to an optimal characterization of random and systematic errors (Sect. 7). The implementation of advanced ensemble Kalman filter systems is highlighted for its direct relevance to hydrometeorological ensemble forecasting. Section 8 discusses the evaluation of large-scale assimilation results with in situ soil moisture observations. Section 9 provides examples of cutting-edge, preoperational soil moisture assimilation systems. The chapter is concluded with a summary (Sect. 10).

2 Components of a Soil Moisture Data Assimilation System

The basic components of a data assimilation system include observations (Sect. 3), modeling (Sects. 4 and 5), and the analysis update (Sect. 6). The observations used for large-scale soil moisture data assimilation systems include screen-level observations or remote-sensing observations. The modeling comprises two parts: (i) the prognostic land surface model to dynamically propagate the state and (ii) diagnostic modeling to translate land surface variables (e.g., soil moisture) into observed quantities such as remotely sensed radiances. The analysis update optimally combines the information from the observations and the model.

The land surface system (Sect. 4) dynamically propagates the prognostic land surface variables in time. The prognostic variables are the key variables that are needed to initialize or restart a model simulation. Depending on the formulation of the land surface model, the prognostic variables could consist of soil moisture, soil temperature, and snow in all its layers, as well as vegetation variables. These model variables have a memory of the past and evolve in time governed by physical laws such as energy and mass conservation and gravity, and in response to external

meteorological forcings, such as precipitation and evapotranspiration. The land surface model $\mathbf{f}_{i,i-1}(\cdot)$ uses the estimate of the state $\hat{\mathbf{x}}_{i-1}^+$ at the previous time $i-1$ (which is an *a posteriori* or *analysis* estimate, if observations were used to update it, see Sect. 6) together with external forcing information \mathbf{u}_i to predict the state $\hat{\mathbf{x}}_i^-$ at the current time i , also called the *a priori* or *forecast* state estimate. For simplicity, the land surface parameters $\boldsymbol{\alpha}$ are assumed constant. The state trajectory can be written as

$$\hat{\mathbf{x}}_i^- = \mathbf{f}_{i,i-1}(\hat{\mathbf{x}}_{i-1}^+, \mathbf{u}_i, \boldsymbol{\alpha}) \quad (1)$$

which approximates the true system

$$\mathbf{x}_i = \mathbf{f}_{i,i-1}(\mathbf{x}_{i-1}, \mathbf{u}_i, \boldsymbol{\alpha}, \mathbf{w}_i) \quad (2)$$

where model errors \mathbf{w}_i are due to errors in the external forcings \mathbf{u}_i , in the model $\mathbf{f}_{i,i-1}(\cdot)$ structure and parameters $\boldsymbol{\alpha}$. Errors in the analysis state $\hat{\mathbf{x}}_{i-1}^+$ and model errors \mathbf{w}_i will introduce errors in the forecasted state $\hat{\mathbf{x}}_i^-$. The uncertainties in $\hat{\mathbf{x}}_{i-1}^+$, \mathbf{w}_i , and $\hat{\mathbf{x}}_i^-$ are described by the analysis error covariance \mathbf{P}_{i-1}^+ , the model error covariance \mathbf{Q}_i , and the forecast error covariance matrix \mathbf{P}_i^- , respectively. One way to estimate the uncertainties is to generate an ensemble of trajectories $\hat{\mathbf{x}}_{i,j}^-$ by perturbing forcings, state variables, or parameters (further discussed in Sect. 7.2.1).

The surface soil moisture that is simulated with a land surface model is often directly comparable to satellite-based soil moisture retrievals. If the land surface model output does not directly correspond to the assimilated observations, such as for brightness temperature or radar backscatter observations, a second modeling step is needed to transform the land model output into observation predictions $\hat{\mathbf{y}}_i^-$ (Sect. 5):

$$\hat{\mathbf{y}}_i^- = \mathbf{h}_i(\hat{\mathbf{x}}_i^-, \boldsymbol{\beta}) \quad (3)$$

This observation model, or *observation operator*, $\mathbf{h}_i(\cdot)$ maps state variables from state space to observation space. This step may also include any spatial or temporal aggregation of land surface variables to satellite-scale observation predictions. Here again, for simplicity, the parameters $\boldsymbol{\beta}$ of the observation operator are assumed constant, and uncertainties in the observation predictions could be estimated through ensemble methods.

The assimilated observations $\mathbf{y}_{\text{obs},i}$ (Sect. 3) can be written as function of the true state \mathbf{x}_i and the observation operator $\mathbf{h}_i(\cdot)$:

$$\mathbf{y}_{\text{obs},i} = \mathbf{h}_i(\mathbf{x}_i, \boldsymbol{\beta}, \mathbf{v}_i) \quad (4)$$

The observation error term \mathbf{v}_i includes measurement errors as well as representativeness errors and is assumed additive in most data assimilation systems. The corresponding observation error covariance matrix will be denoted as \mathbf{R}_i (discussed in Sect. 7.2.2).

The difference between the observations and observation predictions is used to update the state in the analysis step (Sect. 6), for example, as follows:

$$\hat{\mathbf{x}}_i^+ = \hat{\mathbf{x}}_i^- + \mathbf{K}_i \left[\mathbf{y}_{\text{obs}, i} - \hat{\mathbf{y}}_i^- \right] \quad (5)$$

where \mathbf{K}_i is a gain matrix. In most soil moisture data assimilation systems, the state $\hat{\mathbf{x}}_i^-$ used in the updating procedure is a limited subset of the land model prognostic variables that render the system “observable.” (For simplicity, the same notation $\hat{\mathbf{x}}_i^-$ is used for either the full set or a subset of prognostic variables.) A system is observable if the state $\hat{\mathbf{x}}_i^-$ is sufficiently connected to the assimilated observations \mathbf{y}_{obs} . In practice, the state in Eq. 5 contains “observed” state variables that directly contribute to observation predictions, plus “unobserved” variables that are correlated (in the errors) to the “observed” state variables and that will be updated along with the observed state variables during the data assimilation. For example, for the assimilation of surface soil moisture retrievals, the state vector will contain both (observed) surface and (unobserved) deeper-layer soil moisture (Sect. 6.1). Furthermore, the state vector can contain state variables for multiple fine-scale grid cells needed to generate coarser-scale observation predictions, “unobserved” neighboring grid cells (Sect. 6.1.4), and state variables at different time steps (Sect. 6.2).

3 Observations Related to Soil Moisture

Global-scale soil moisture can be inferred from global surface observational networks or from satellite-based observations. Near-surface meteorological observations of two-meter air temperature and relative humidity are measured routinely by the land surface synoptic report (SYNOP) operational network at the global scale and, under certain conditions, are related to surface and root-zone soil moisture. The coverage of these station observations varies greatly between dense measurements in Europe, North America, and parts of Asia and significantly sparser coverage elsewhere. Screen-level variables are assimilated operationally at major numerical weather prediction centers, because their assimilation improves medium-range forecasts of surface meteorological conditions, albeit with limited improvement to the soil moisture estimates themselves (Sect. 6.1.1).

Satellite remote sensing offers a way to observe the surface soil water content at continental to global scales. Most existing remote-sensing devices use the distinct physical properties of soil water in interaction with electromagnetic signals at specific wavelengths. Microwave radiometry (wavelength 1–20 cm, frequency 10–1.4 GHz) has been used for several decades to estimate the soil’s dielectric constant, and hence soil moisture. Examples of radiometers on board satellite platforms that passively measure microwave emission from the land surface are the Advanced Microwave Sounding Radiometer 2 (AMSR2), the Microwave Imaging Radiometer using Aperture Synthesis on board the Soil Moisture Ocean Salinity (SMOS) mission, the Aquarius radiometer, and the radiometer on board the Soil

Moisture Active Passive (SMAP) mission, among others. Active microwave sensors send and receive microwave signals. Examples include the radars on board the two European remote-sensing satellites (ERS-1 and ERS-2), the Advanced SCATterometer (ASCAT) on board the European Organisation for the Exploitation of Meteorological Satellites (METOP) series of satellites, and also the radar on board SMAP. All the above-mentioned satellites are polar orbiting, and each provides global coverage approximately every 2–3 days.

Microwave remote sensing is attractive because the signals at L-band wavelengths (1.4 GHz) used for SMOS and SMAP are most sensitive to soil moisture, less impacted by vegetation, and not impacted by clouds or light rain. There are, however, some obvious limitations:

- (i) Microwave instruments only sense the soil moisture in a thin (1–5 cm) surface layer, with the penetration depth depending on the microwave wavelength and soil moisture content.
- (ii) The spatial coverage is limited to swaths of $\sim 250 - 1000$ km, and the revisit time is once every couple of days.
- (iii) Passive microwave data have coarse spatial resolutions ($\sim 10 - 100$ km).
- (iv) Active microwave data are typically very noisy.

Through data assimilation, some of these limitations can be overcome, as will be illustrated in Sect. 6. Microwave observations can be assimilated directly as brightness temperatures or backscatter values or after inversion to soil moisture retrievals. Various methods exist to infer soil moisture retrievals from active or passive microwave signals. For example, AMSR2, SMOS ,and SMAP retrievals are obtained by explicitly inverting the relationship between soil moisture and passive microwave emission (e.g., Wigneron et al. 2007) and ASCAT retrievals employ a change detection technique (e.g., Bartalis et al. 2007). It is important to note that the climatologies of various retrieval products can be very different and require careful attention when comparing or merging various retrieval products (Dorigo et al. 2015; Reichle and Koster 2004). The soil moisture retrieval process is very sensitive to radiative transfer or backscatter model parameters (Sect. 5.2), and to auxiliary information (e.g., about soil temperature and vegetation) provided by land surface modeling systems. For example, the SMOS soil moisture retrievals use ECMWF's simulated surface soil temperature and moisture as prior information in the retrieval, and SMAP retrievals use prior and auxiliary information from the NASA Goddard Earth Observing System Model, version 5 (GEOS-5) (Sect. 4). In addition, retrievals can technically be calculated under any conditions, but the soil moisture estimates are only meaningful in areas with moderate topographic complexity, nonfrozen and snow-free conditions, sparse vegetation, and at times with limited precipitation.

An alternative to using microwave signals is to use gravity measurements to determine changes in the mass of water: the Gravity Recovery and Climate Experiment (GRACE) mission consists of two satellites whose relative distance and velocity can be related to anomalies in water amounts at and near the land surface,

including soil moisture, snow, and surface water. GRACE observations thus provide information on deeper soil moisture under any weather conditions, but their resolution is very coarse in space (~ 250 km) and time (monthly).

4 Land Surface Modeling

A wide variety of models has been used to simulate the dynamic evolution of soil moisture, ranging from simple solutions of equations that represent the movement of water in unsaturated soils to full land surface models (LSM, chapter on ► “[Land Surface Hydrological Models](#)”) that simulate the soil-vegetation-atmosphere interactions. The strength of models is in their reliance on physical laws known to operate in nature and in their ability to provide a consistent and balanced distribution of water and heat. Historically, some LSMs have been optimized to simulate select land surface variables for specific regions, whereas others are tuned to provide good boundary information in general circulation models, sometimes with little attention paid to the physical realism of the simulated soil moisture. Examples of LSMs that are part of operational integrated modeling systems are the Hydrology-Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL, Balsamo et al. 2009) in ECMWF’s operational Integrated Forecasting System (IFS) and the Catchment land surface model (Koster et al. 2000) in NASA’s GEOS-5 system.

The prognostic variables related to soil moisture (i.e., part of $\widehat{\mathbf{x}}_i^-$ in Eq. 1) are very different in these models. In HTESSEL the soil moisture is calculated in four layers of thicknesses of 0.07, 0.21, 0.72, and 1.89 m from top to bottom. The Catchment model defines three prognostic variables that describe the equilibrium soil moisture profile and deviations from the equilibrium across the entire watershed (or modeling unit). Specifically, the catchment deficit (catdef), root-zone excess (rzexc), and surface excess (srfecc) prognostic variables are used together to diagnose soil moisture in a surface layer (sfmc, 0–0.05 m), a root-zone layer (rzmc, 0–1 m), and the entire profile (prmc). The latter extends from the surface to the bedrock at a variable depth between 1.3 and 10 m. The temperature (tsoil) of the topmost soil layer (of thickness 0.1 m) is diagnosed from the corresponding ground heat content prognostic variable (ght).

The LSM structure (denoted $\mathbf{f}_{i,i-1}(\cdot)$ in Sect. 2), parameters ($\boldsymbol{\alpha}$), and forcing inputs (\mathbf{u}_i) determine the climatology of the simulated soil moisture, that is, its long-term average and seasonal variation. For example, the porosity parameter determines the maximum amount of water that can be contained in a soil layer, and the hydraulic conductivity determines how fast soil moisture moves across soil layers. Locally, optimal parameters for soil moisture modeling could be found through calibration against historic data records of observations (Part V, “Model Parameter Estimation and Uncertainty Analysis”). However, the calibration of LSMs is a nontrivial task, because of multiparameter interactions, equifinality, and the scale-dependency of the land model parameters. Global LSMs are usually not calibrated and solely rely on auxiliary static information about soil and vegetation properties. Soil physical

parameters, for example, are typically inferred from global soil texture maps using static lookup tables or pedotransfer functions. These parameters are not perfect, yet they determine the average (climatological) level of soil moisture and may thus be responsible for persistent biases. LSM parameters also impact random errors because they affect the nature of shorter-term soil moisture dynamics.

The most important forcing input to soil moisture simulations is precipitation, but other surface meteorological fields are also required, including radiation, air temperature and humidity, and wind speed. For small-scale applications, forcing data are usually collected from meteorological towers. Forcing data for large-scale applications rely on a merger of surface observations and atmospheric reanalysis fields. Examples include the recent MERRA-Land (Reichle et al. 2011) and ERA-Land (Balsamo et al. 2013) products. The underlying atmospheric reanalysis products assimilate a very large number of conventional and satellite-based observations of the atmosphere into a global atmospheric model and provide self-consistent meteorological fields with complete spatial and temporal coverage. However, these products have a relatively coarse resolution and are subject to errors in the reanalysis systems. Errors in the long-term average precipitation amounts or intensity result in biased simulations. At shorter time scales, missed or excessive precipitation events cause random errors in simulated soil moisture. Merging the reanalysis data with satellite and gauge-based precipitation data products mitigates some, but not all, of these errors.

Another determining factor for soil moisture simulations is the land model initialization. A so-called cold model start (using an arbitrary state initialization) will generally cause a drift in soil moisture until a steady soil moisture level is reached, which typically takes simulations across several years. It is therefore important to spin up the model until its prognostics are in agreement with the climatological boundaries as determined by the forcings and model parameters.

5 Observation Operator

If satellite-based soil moisture retrievals are assimilated, then the model output from land surface models can be directly compared to the assimilated observations. However, the assimilation of screen-level observations, microwave observations, and terrestrial water storage anomalies requires a diagnostic modeling step with an observation operator that facilitates the direct comparison between the land model output and the observed variables (Reichle et al. 2014).

5.1 Screen-Level Observation Predictions

Screen-level observations are *in situ* measurements of temperature and relative humidity ($T_{2m,obs}$, $RH_{2m,obs}$) at 2 m above the land surface. In global atmospheric models, estimates of \bar{T}_{2m} and \bar{RH}_{2m} are typically obtained by interpolating the model

variables from the surface (as computed by the land surface model component) to the atmospheric conditions at the height of the lowest atmospheric model level (Mahfouf et al. 2009), following the Monin-Obukhov similarity theory. The latter describes flow and turbulence properties in the lowest 10% of the atmospheric boundary layer. This assumes that the first atmospheric level is not at 2 m, but instead imposed higher up. Formally, this can be written as

$$\begin{bmatrix} \widehat{T}_{2m} \\ \text{RH}_{2m} \end{bmatrix}^- = \mathbf{h}_{2m}(\widehat{\mathbf{x}}^-, \mathbf{u}, \boldsymbol{\beta}_{2m}) \quad (6)$$

where $\widehat{\mathbf{x}}^- = [\widehat{\text{sfmc}}^-, \widehat{\text{rzmc}}^-, \widehat{\text{tsurf}}^-, \widehat{\text{tsoil}}^-]^T$ are the land surface state variables, including surface and root-zone soil moisture, and surface and soil temperature, respectively. The vector \mathbf{u} contains the imposed atmospheric variables, with air temperature, humidity, and wind speed at the lowest atmospheric level, and $\mathbf{h}_{2m}()$ represents the vertical interpolation with $\boldsymbol{\beta}_{2m}$ the interpolation parameters.

5.2 Microwave Observation Predictions

Passive microwave emission from the land surface is often referred to as brightness temperature (T_b) and measured for certain polarizations, wavelengths, and incidence angles, collectively denoted as the instrument configuration. The brightness temperature for a given configuration (c) can be simulated as a function of surface soil temperature, attenuated by soil and vegetation characteristics:

$$\widehat{T}_{bc}^- = \mathbf{h}_p(\widehat{\mathbf{x}}^-, \boldsymbol{\beta}_p, \mathbf{c}_p) \quad (7)$$

with $\mathbf{h}_p()$ a radiative transfer model for passive (p) microwave emission; $\widehat{\mathbf{x}}^- = [\widehat{\text{sfmc}}^-, \widehat{\text{tsurf}}^-, \widehat{\text{tsoil}}^-]^T$ a set of dynamic land surface variables such as surface soil moisture and surface and soil temperature; $\boldsymbol{\beta}_p$ a vector with land surface-specific parameters, including the microwave soil roughness length, scattering albedo, and vegetation parameters; and \mathbf{c}_p a set of radiometer configuration constants.

Active radar backscattering coefficients (σ^0) measured for a given sensor configuration c can be similarly related to land surface variables as follows:

$$\widehat{\sigma}_c^0 = \mathbf{h}_a(\widehat{\mathbf{x}}^-, \boldsymbol{\beta}_a, \mathbf{c}_a) \quad (8)$$

with $\mathbf{h}_a()$ the backscattering model for active (a) microwave signals; $\widehat{\mathbf{x}}^-$ again a set of dynamic land surface variables including soil moisture; $\boldsymbol{\beta}_a$ a vector with land surface specific parameters, such as the root-mean-square (rms) surface height and correlation length to quantify the roughness; and \mathbf{c}_a a set of radar properties.

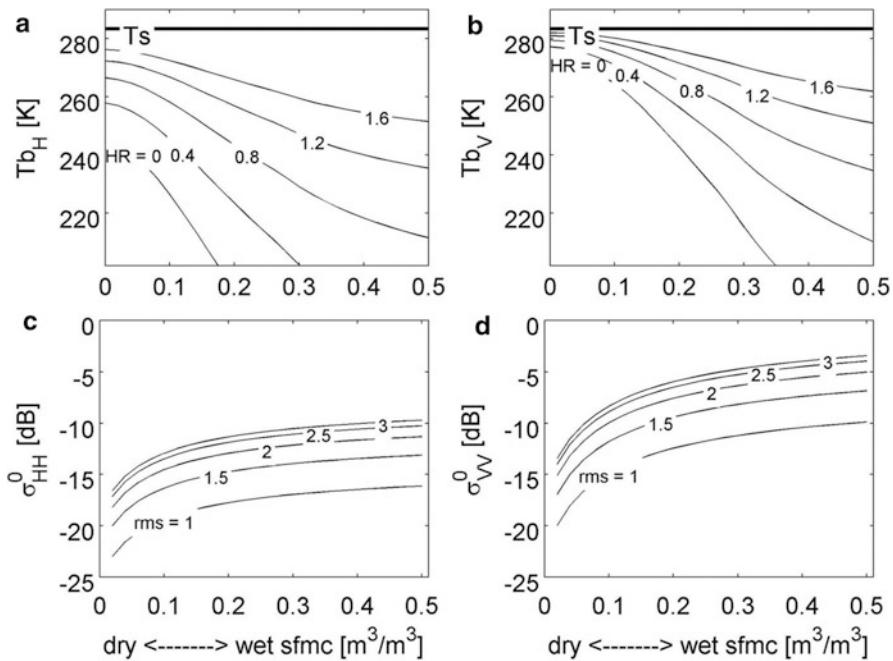


Fig. 1 (Top) Relationship between soil moisture (sfmc) and brightness temperature (T_b) at (a) horizontal and (b) vertical polarization. (Bottom) Same but for radar backscatter (σ_c^0) at (c) horizontal and (d) vertical co-polarization. The soil contains 43% sand and 23% clay and has a bulk density of 1.24 g/cm³. The weighted surface/soil temperature (T_s) is set to 283 K. The sensor-specific constants are following the SMAP instrument details (40° incidence angle, radar at 1.29 GHz, radiometer at 1.41 GHz)

Figure 1 illustrates the nonlinear relationship between the surface soil moisture content sfmc and T_b or σ_c^0 as simulated by physically based radiative transfer and backscattering models over bare soil. In these models, soil moisture is first converted to a soil dielectric constant using a dielectric mixing model. The dielectric constant then determines the surface reflectivity and thus the emission, reflection, and scattering of waves. Figures 1a–b use a zero-order radiative transfer model (Mo et al. 1982) as $\mathbf{h}_p(\cdot)$ for T_b simulation. Figures 1c–d use the Integral Equation Model (Fung et al. 1992) as $\mathbf{h}_a(\cdot)$ for σ_c^0 simulation. For the illustrations in the figure, the sensor configurations c are chosen according to the design of the radiometer and radar sensors on board SMAP, i.e., with an incidence angle of 40°, either horizontal (H) or vertical (V) polarization (co-polarization for the radar), and a frequency of 1.41 GHz for the radiometer and 1.26 GHz for the radar. Figures 1a–b illustrate that the brightness temperature is close to surface soil temperature (T_s) under dry conditions, and it decreases with soil moisture. As a rule of thumb, a 2–3 K increase in T_b is associated with a 0.01 m^3/m^3 decrease in soil moisture for incidence angles around 40° and for low vegetation regions (vegetation water content of less than about 5 kg/m²). Figures 1c–d show that a 1–5 dB decrease in σ_c^0 can be expected for a

0.1 m³/m³ decrease in soil moisture below 0.35 m³/m³, whereas little sensitivity is found for wetter soil moisture. The figure further highlights the sensitivity of the relationships to a change in only one select parameter ($\in \beta$), i.e., the microwave roughness HR [–] for Tb_c simulation, or the rms surface height [cm] for σ_c^0 simulation. A realistic range of HR (Fig. 1a–b) easily introduces differences of 50 K or more in Tb for a moderate level of soil moisture. Figures 1c–d show that a 1 cm increase in the rms roughness parameter could increase the backscattering by 5 dB. In reality, vegetation further complicates the picture. To summarize, the relationship between the soil moisture content and Tb_c or σ_c^0 is highly dependent on a set of very uncertain parameters.

5.3 Terrestrial Water Storage Predictions

Changes in terrestrial water storage (TWS) are among the dominant mass variations that can be detected in the GRACE signal (Rodell et al. 2007). The simulated monthly TWS represents the vertically integrated water amount as

$$\widehat{\text{TWS}}^- = \mathbf{h}_{\text{TWS}}(\widehat{\mathbf{x}}^-, \boldsymbol{\beta}_{\text{TWS}}) \quad (9)$$

with $\mathbf{h}_{\text{TWS}}(\cdot)$ the vertical integration of the individual components of $\widehat{\mathbf{x}}^- = [\widehat{\text{sfmc}}^-, \widehat{\text{rzmc}}^-, \widehat{\text{gwt}}^-, \widehat{\text{snow}}^-, \widehat{\text{ice}}^-, \dots]^T$, including surface soil moisture, root-zone soil moisture, depth to the groundwater table (gwt), snow, ice, and possibly water stored in or on vegetation, and $\boldsymbol{\beta}_{\text{TWS}}$ refers to any parameter needed to compute the TWS. The corresponding observations are mostly provided in terms of anomalies, i.e., as deviations from a long-term (multi-month, multiyear) average, which is not necessarily known. For assimilation into a model, the observed TWS anomalies are typically converted to actual TWS_{obs} by adding a long-term model estimate $\langle \widehat{\text{TWS}}^- \rangle$.

6 Assimilation of Observations Related to Soil Moisture

6.1 Sequential State Updating

Popular methods for the sequential assimilation of soil moisture observations at large scales include direct insertion, nudging and statistical correction (Dharssi et al. 2011), optimal interpolation (Mahfouf et al. 2009), extended Kalman filtering (Sabater et al. 2007), variational assimilation (Reichle et al. 2001; Hess et al. 2008), ensemble Kalman filtering (Reichle et al. 2002), and particle filtering (Pan et al. 2008). Details of these techniques are given in chapter on ► “Fundamentals of Data Assimilation and Theoretical Advances”. The ensemble Kalman filter and particle filtering variants are the methods that lend themselves most directly to integration in hydrometeorological ensemble forecast systems.

Table 1 Summary of (1) forecast and (2) update equations for optimal interpolation (OI), (Extended) Kalman filtering ((E)KF) and ensemble Kalman filtering (EnKF). The OI and (E)KF use a single state trajectory and a predefined or linearly evolving \mathbf{P}_i^- , respectively, whereas the EnKF uses N ensemble members, perturbed with model error $\mathbf{w}_{i,j}$ to diagnose \mathbf{P}_i^-

OI, (E)KF	EnKF
(1) A priori state and uncertainty	
$\widehat{\mathbf{x}}_i^- = \mathbf{f}_{i,i-1}(\widehat{\mathbf{x}}_{i-1}^+, \mathbf{u}_i, \boldsymbol{\alpha})$	$\widehat{\mathbf{x}}_{i,j}^- = \mathbf{f}_{i,i-1}\left(\widehat{\mathbf{x}}_{i-1,j}^+, \mathbf{u}_i, \boldsymbol{\alpha}, \mathbf{w}_{i,j}\right)$ with $j = 1, \dots, N$
$\mathbf{P}_i^- = \mathbf{B}$ (OI) $\mathbf{P}_i^- = \mathbf{F}_{i,i-1}\mathbf{P}_{i-1}^+\mathbf{F}_{i,i-1}^T + \mathbf{Q}_i$ ((E)KF)	$\mathbf{P}_i^- = \text{Cov}(\widehat{\mathbf{x}}_i^-, \widehat{\mathbf{x}}_i^-)$
Observation predictions	
$\widehat{\mathbf{y}}_i^- = \mathbf{h}_i(\widehat{\mathbf{x}}_i^-, \boldsymbol{\beta})$	$\widehat{\mathbf{y}}_{i,j}^- = \mathbf{h}_i(\widehat{\mathbf{x}}_{i,j}^-, \boldsymbol{\beta})$
(2) A posteriori state and uncertainty	
$\mathbf{K}_i = \mathbf{P}_i^- \mathbf{H}_i^T [\mathbf{H} \mathbf{P}_i^- \mathbf{H}^T + \mathbf{R}_i]^{-1}$	$\mathbf{K}_i = \text{Cov}(\widehat{\mathbf{x}}_i^-, \widehat{\mathbf{y}}_i^-) [\text{Cov}(\widehat{\mathbf{y}}_i^-, \widehat{\mathbf{y}}_i^-) + \mathbf{R}_i]^{-1}$
$\widehat{\mathbf{x}}_i^+ = \widehat{\mathbf{x}}_i^- + \mathbf{K}_i [\mathbf{y}_{\text{obs},i} - \widehat{\mathbf{y}}_i^-]$	$\widehat{\mathbf{x}}_{i,j}^+ = \widehat{\mathbf{x}}_{i,j}^- + \mathbf{K}_i [\mathbf{y}_{\text{obs},i,j} - \widehat{\mathbf{y}}_{i,j}^-]$ $\widehat{\mathbf{x}}_i^+ = \frac{1}{N} \sum_{j=1}^N \widehat{\mathbf{x}}_{i,j}^+$
$\mathbf{P}_i^+ = [\mathbf{I} - \mathbf{K}_i \mathbf{H}]_i \mathbf{P}_i^-$	$\mathbf{P}_i^+ = \text{Cov}(\widehat{\mathbf{x}}_i^+, \widehat{\mathbf{x}}_i^+)$

Sequential filtering involves cycling through two steps, as summarized in Table 1. In the first step, a *background*, *a priori*, or *forecast state* estimate is generated with a dynamic model $\mathbf{f}_{i,i-1}(\cdot)$. This forecast could either be deterministic, i.e., $\widehat{\mathbf{x}}_i^-$, or consist of an ensemble $\widehat{\mathbf{x}}_{i,j}^-$ where forecast perturbations ($\mathbf{w}_{i,j}$) are applied to generate each ensemble member j ($j = 1, \dots, N$). The second step generates an *a posteriori* or *analysis* state by correcting the state with observations, as in Eq. 5. Two update variants have been classic for the assimilation of soil moisture observations. The most commonly used variant adds an *increment* to the forecasted state. This increment is determined using the difference between observations and observation predictions $[\mathbf{y}_{\text{obs},i} - \widehat{\mathbf{y}}_i^-]$ and a blending matrix, or gain matrix \mathbf{K}_i . In “optimal” (minimum analysis error variance) assimilation schemes, the gain \mathbf{K}_i is found by weighting the uncertainty in the forecast state \mathbf{P}_i^- and in the observations \mathbf{R}_i . If the forecast error covariance \mathbf{P}_i^- is dynamically propagated in time, then \mathbf{K}_i is called a Kalman gain. If \mathbf{P}_i^- is diagnosed from ensemble forecasts, then the filter is called an ensemble Kalman filter. Another update variant preferentially weighs a set of possible forecasts (particles, ensemble members) so that the resulting observation prediction is closest to observations. This approach is used in particle filters, which have a great potential for soil moisture assimilation problems and subsequent ensemble forecasting, but have not yet been explored thoroughly for large-scale or multi-scale land surface data assimilation.

The following examples conceptually describe the assimilation of various soil moisture observations using filtering techniques with increasing complexity. Each example can be seen as a variant of the basic sequential update Eq. 5. It is important

to note that the combinations of the selected observation types and assimilation techniques are not exclusive and only chosen for illustrative purposes.

6.1.1 Screen-Level Data Assimilation Illustrated with Optimal Interpolation

Screen-level observations ($T_{2m,obs}$, $RH_{2m,obs}$) have been assimilated operationally for numerical weather prediction (NWP) (Giard and Bazile 2000; Bélair et al. 2003). To limit the computational effort, the state vector is often limited to a few prognostic variables. Consider a state consisting of surface moisture content (sfmc), root-zone soil moisture (rzmc), surface temperature (tsurf), and soil temperature (tsoil). The typical update equation can be written as:

$$\begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \\ \widehat{\text{tsurf}} \\ \widehat{\text{tsoil}} \end{bmatrix}_i^+ = \begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \\ \widehat{\text{tsurf}} \\ \widehat{\text{tsoil}} \end{bmatrix}_i^- + \mathbf{K}_i \left(\begin{bmatrix} T_{2m} \\ RH_{2m} \end{bmatrix}_{\text{obs}} - \begin{bmatrix} \widehat{T}_{2m} \\ \widehat{RH}_{2m} \end{bmatrix}_i^- \right)_i \quad (10)$$

The initial implementations of this analysis scheme at operational centers use a priori defined constants for the \mathbf{K}_i matrix (4x2). These constants are a priori calibrated, and do not use any explicit observation operator or any statistical information about background or observation errors. Theoretically this approach cannot be classified as “optimal interpolation,” but it has been commonly referred to as such in the literature.

Optimal interpolation uses explicit expressions for the a priori and observation error covariance matrices to determine the \mathbf{K}_i matrix. At Météo-France and at ECMWF, the blending matrix \mathbf{K}_i for the assimilation of screen-level observations uses statistical error information and is further advanced by the use of analytical Jacobians of the land surface model (Mahfouf et al. 2009; Drusch et al. 2009; de Rosnay et al. 2013), i.e.,

$$\mathbf{K}_i = \mathbf{B} \mathbf{H}_i^T [\mathbf{H}_i \mathbf{B} \mathbf{H}_i^T + \mathbf{R}]^{-1} \quad (11)$$

with \mathbf{B} (4×4) a time-invariant background error covariance matrix, \mathbf{R} (2×2) a time-invariant diagonal observation error covariance matrix, and $\mathbf{H}_i = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}|_{\tilde{\mathbf{x}}_i}$ the linearized observation operator of dimension (2×4). The observation operator $\mathbf{h}_i(.)$ is a physically based model, and the Jacobian elements are computed in finite differences, i.e., the change in \widehat{T}_{2m} and \widehat{RH}_{2m} is computed for a small perturbation in the individual state components (sfmc, rzmc, etc.). The \mathbf{H}_i matrix maps differences between simulated and observed screen-level variables (*innovations*) to updates (*increments*) in soil moisture and temperature. In general, the limited sensitivity of \widehat{T}_{2m} or \widehat{RH}_{2m} to root-zone soil moisture leads to small updates (Drusch et al. 2009).

The above approach of using a dynamic Jacobian for the observation operator in the blending matrix \mathbf{K}_i has been referred to as “simplified extended Kalman filtering”

by the land surface community involved in NWP, because of its close ties with the extended Kalman filter. However, the use of a fixed background error covariance matrix \mathbf{B} by definition means that no Kalman filtering is involved, and the assimilation technique is theoretically an “optimal interpolation.”

6.1.2 Soil Moisture (Retrieval) Data Assimilation Illustrated with (Extended) Kalman Filtering

The (extended) Kalman filter ((E)KF) is similar to the optimal interpolation method in its incremental update equation. The difference is that the Kalman filter dynamically propagates the a priori error covariance matrix, using a linear dynamic model. The (E)KF or its close variants (Sabater et al. 2007; Hess et al. 2008) have not been widely used for operational soil moisture data assimilation, because land surface models typically require a rather complex linearization. However, the (E)KF provides the fundamentals to all other Kalman filter variants.

Assume that observations (e.g., retrievals) of soil moisture content (sfmc_{obs}) are assimilated into a model, with surface and root-zone soil moisture as the state variables. The update equation is

$$\begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \end{bmatrix}_i^+ = \begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \end{bmatrix}_i^- + \mathbf{K}_i \left[\text{sfmc}_{\text{obs}} - \widehat{\text{sfmc}} \right]_i, \text{ with} \quad (12)$$

$$\mathbf{K}_i = \mathbf{P}_i^- \mathbf{H}_i^T [\mathbf{H} \mathbf{P}^- \mathbf{H}^T + \mathbf{R}]_i^{-1} = \left[\rho \cdot \sigma_{\text{sfmc}_i}^2 \cdot \sigma_{\text{rzmc}_i}^2 \right]_i \left[\sigma_{\text{sfmc}_i}^2 + \sigma_{\text{sfmc}, \text{obs}}^2 \right]_i^{-1} \quad (13)$$

$$\mathbf{P}_i^+ = [\mathbf{I} - \mathbf{K}_i] \mathbf{P}_i^- \quad (14)$$

where $\mathbf{H} = [1 \ 0]$ in this case, but the linearized observation operator could be written more generally as $\mathbf{H}_i = \frac{\partial \mathbf{h}}{\partial \mathbf{x}}|_{\mathbf{x}_i}$. The observation error variance is $\sigma_{\text{sfmc}, \text{obs}, i}^2$, the observation prediction error variance is $\sigma_{\text{sfmc}_i}^2$, and \mathbf{P}_i^- (2x2) contains a time-variable a priori surface and root-zone error variances ($\sigma_{\text{sfmc}_i}^2, \sigma_{\text{rzmc}_i}^2$) on the diagonal, and covariances $\rho_i \sigma_{\text{sfmc}_i}^2, \sigma_{\text{rzmc}_i}^2$ as off-diagonal elements:

$$\mathbf{P}_i^- = \begin{bmatrix} \sigma_{\text{sfmc}_i}^2 & \rho \cdot \sigma_{\text{sfmc}_i}^2 \cdot \sigma_{\text{rzmc}_i}^2 \\ \rho \cdot \sigma_{\text{sfmc}_i}^2 \cdot \sigma_{\text{rzmc}_i}^2 & \sigma_{\text{rzmc}_i}^2 \end{bmatrix}_i \quad (15)$$

It is through the error correlations (ρ_i) in \mathbf{P}_i^- that surface soil moisture *innovations* $\left[\text{sfmc}_{\text{obs}} - \widehat{\text{sfmc}} \right]_i$ are propagated to both surface and root-zone *increments* $\mathbf{K}_i \left[\text{sfmc}_{\text{obs}} - \widehat{\text{sfmc}} \right]_i$. The a priori \mathbf{P}_i^- is reduced to \mathbf{P}_i^+ after each assimilation update (Eq. 14).

The a priori \mathbf{P}_i^- is determined dynamically as function of the modeling system. For additive Gaussian model error \mathbf{w}_i with an error covariance matrix of \mathbf{Q}_i (Eq. 2), the forecast error covariance \mathbf{P}_i^- can be approximated by

$$\mathbf{P}_i^- = \mathbf{F}_{i,i-1} \mathbf{P}_{i-1}^+ \mathbf{F}_{i,i-1}^T + \mathbf{Q}_i \quad (16)$$

where $\mathbf{F}_{i,i-1}$ (2×2) is a linearized version of $\mathbf{f}_{i,i-1}$ (.) and $\mathbf{F}_{i,i-1} \mathbf{P}_{i-1}^+ \mathbf{F}_{i,i-1}^T$ is the propagated analysis error covariance. In using a tangent linear $\mathbf{F}_{i,i-1}$ operator, the method is referred to as “extended” Kalman filter. If $\mathbf{F}_{i,i-1}$ is a linearized model version, then Eq. 16 is known to suffer from unlimited error variance growth, because the third and higher order moments of the Taylor expansion are discarded in Eq. 16 (closure problem). It is possible to avoid these problems with other Kalman filter variants, as discussed in the next sections.

6.1.3 Soil Moisture (Retrieval) Data Assimilation Illustrated with 1D Ensemble Kalman Filtering

The ensemble Kalman filter (EnKF, Reichle et al. 2002) circumvents the need for a linear(ized) state propagation model and observation operator by diagnosing error covariance matrices from ensemble information. Examples of one-dimensional (“1D”) EnKF studies using soil moisture retrievals from various microwave sensors include Liu et al. (2011) and Draper et al. (2012). A “1D” EnKF updates the state at the locations that coincide with the assimilated observations, and it is assumed that the observations and the model have the same spatial resolution. In the next Sect. 6.1.4, a spatially distributed or three-dimensional (“3D”) expansion of the EnKF will be presented, with inclusion of horizontal information propagation and with the ability to possibly deal with multiple scales.

Assume again that satellite-based surface soil moisture retrievals sfmc_{obs} are assimilated and that the model operates at the same spatial resolution as the observations. An ensemble of states $\widehat{\mathbf{x}}_{i,j}$ ($j = 1, \dots, N$) is generated by perturbing the model simulations (discussed in Sect. 7.2.1). The observations are also perturbed to ensure consistency in the EnKF formulation used here. Note though that some variants of the EnKF exist that avoid such perturbations. The observation predictions are given by $\widehat{\mathbf{y}}_{i,j} = \widehat{\text{sfmc}}_{i,j} = h(\widehat{\mathbf{x}}_{i,j})$, i.e., accounting for a possibly nonlinear mapping between the observed $\text{sfmc}_{\text{obs},i,j}$ and the state variables, even though in this example the observation operator is linear $\mathbf{H} = [1 \ 0]$. The update equation for each state member can be written as:

$$\begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \end{bmatrix}_{i,j}^+ = \begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \end{bmatrix}_{i,j}^- + \mathbf{K}_i \left[\text{sfmc}_{\text{obs}} - \widehat{\text{sfmc}} \right]_{i,j}^-, \text{ with} \quad (17)$$

$$\begin{aligned}\mathbf{K}_i &= \text{Cov}(\widehat{\mathbf{x}}_i^-, \widehat{\mathbf{y}}_i^-) [\text{Cov}(\widehat{\mathbf{y}}_i^-, \widehat{\mathbf{y}}_i^-) + \mathbf{R}_i]^{-1} \\ &= \text{Cov}(\widehat{\mathbf{x}}_i^-, \widehat{\mathbf{y}}_i^-) \left[\sigma_{\text{sfmc}}^{-2} + \sigma_{\text{sfmc, obs}}^2 \right]_i^{-1}\end{aligned}\quad (18)$$

Unlike Eq. 16, the error covariances used in the Kalman gain are now dynamically diagnosed from the ensemble dispersion in the forecasts and observation predictions. Specifically, $\text{Cov}(\widehat{\mathbf{x}}_i^-, \widehat{\mathbf{y}}_i^-)$ is found by correlating the ensemble departures in the state variables with those in the observation predictions, and $\text{Cov}(\widehat{\mathbf{y}}_i^-, \widehat{\mathbf{y}}_i^-)$ is the error covariance of the observation predictions. Note that $\mathbf{P}_i^- = \text{Cov}(\widehat{\mathbf{x}}_i^-, \widehat{\mathbf{x}}_i^-)$, but the computation of this matrix is not required for the Kalman gain (Eq. 18). The gain factor \mathbf{K}_i maps the surface soil moisture *innovation* $[\text{sfmc}_{\text{obs}} - \widehat{\text{sfmc}}]_i$ to *increments* $\mathbf{K}_i [\text{sfmc}_{\text{obs}} - \widehat{\text{sfmc}}]_i$ in all prognostic state variables, using the diagnosed error covariances between these variables.

The final a posteriori state estimate $\widehat{\mathbf{x}}_i^+$ and its uncertainty are given by

$$\begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \end{bmatrix}_i^+ = \frac{1}{N} \sum_{j=1}^N \begin{bmatrix} \widehat{\text{sfmc}} \\ \widehat{\text{rzmc}} \end{bmatrix}_{i,j}^+ \quad \text{with } \mathbf{P}_i^+ = \text{Cov}(\widehat{\mathbf{x}}_i^+, \widehat{\mathbf{x}}_i^+)\quad (19)$$

The a posteriori uncertainty in the state estimate \mathbf{P}_i^+ is diagnosed from the analysis ensemble, which typically contracts during the assimilation.

6.1.4 Brightness Temperature Data Assimilation Illustrated with 3D Ensemble Kalman Filtering

Classical retrieval assimilation is appealing because of its relatively straightforward implementation, but there is a serious concern about observation biases. The inversion process from brightness temperatures to soil moisture retrievals relies on parameters and ancillary information that may be inconsistent with that used in the land surface model within the assimilation system. It is thus more natural to couple a radiative transfer or backscatter model to a land surface model, to forecast \widehat{Tb}_c or $\widehat{\sigma}_c^{0-}$ along with soil moisture, and then assimilate $Tb_{c,\text{obs}}$ or $\sigma_{c,\text{obs}}^0$ (rather than the soil moisture retrievals) as in Entekhabi et al. (1994), Reichle et al. (2001), and Balsamo et al. (2006), among others.

In the following example, the above EnKF equations are further illustrated for spatial (or “3D”) filtering (Reichle and Koster 2003) and using brightness temperature observations Tb_{obs} at a coarser resolution than the fine-scale model simulations. This will highlight that (i) brightness temperature information can be translated into soil moisture updates, (ii) soil moisture can be updated in unobserved areas, and (iii) fine-scale soil moisture estimates can be obtained, through dynamic disaggregation of coarse-scale observations. The observation

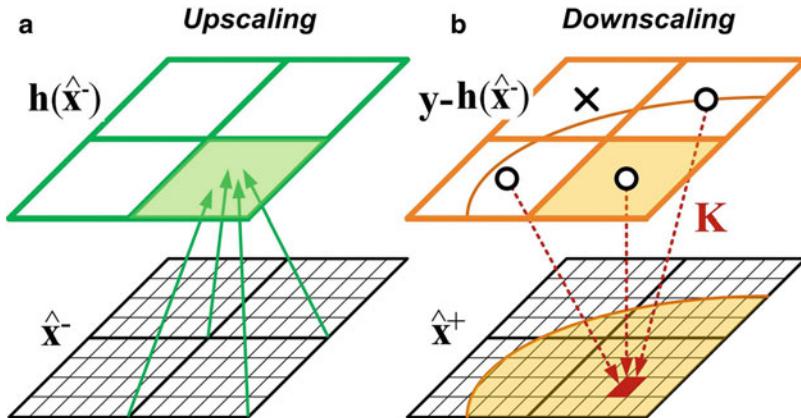


Fig. 2 Schematic of a multi-scale 3D filter. (a) State variables in fine-scale model grid cells are aggregated to coarse-scale observation predictions through the observation operator $\mathbf{h}(\cdot)$. (b) Coarse-scale innovations (observation-minus-forecasts, $\mathbf{y}_{\text{obs},i}\mathbf{h}(\hat{\mathbf{x}}_i^-)$) located within the influence radius (*curved shaded area*) around a fine-scale model grid cell are indicated by circles and will contribute to the fine-scale state update

predictions \widehat{Tb}_i^- require model information about soil moisture, temperature, and vegetation. The state vector therefore contains surface and soil temperature and possibly vegetation, especially if the latter is dynamically evolving in the model. Yet, here it is excluded for simplicity.

The state update is presented at a single fine-scale location k , and a single ensemble member j , using (possibly multiple) coarse-scale observations $Tb_{\text{obs},\kappa}$ ($\kappa = 1, \dots, m$) that are within a chosen influence area around the fine-scale location, as illustrated in Fig. 2:

$$\begin{bmatrix} \widehat{\text{sfmc}}_{i,j} \\ \widehat{\text{rzmc}}_{i,j} \\ \widehat{\text{tsurf}}_{i,j} \\ \widehat{\text{tsoil}}_{i,j} \end{bmatrix}_k^+ = \begin{bmatrix} \widehat{\text{sfmc}}_{i,j} \\ \widehat{\text{rzmc}}_{i,j} \\ \widehat{\text{tsurf}}_{i,j} \\ \widehat{\text{tsoil}}_{i,j} \end{bmatrix}_k^- + [\mathbf{K}_{i,j}]_k \left(\begin{bmatrix} Tb_1 \\ \dots \\ Tb_\kappa \\ \dots \\ Tb_m \end{bmatrix}_{\text{obs}} - \begin{bmatrix} \widehat{Tb}_1 \\ \dots \\ \widehat{Tb}_\kappa \\ \dots \\ \widehat{Tb}_m \end{bmatrix} \right)_{i,j} \quad (20)$$

The coarse-scale observation predictions are calculated by (i) transforming the fine-scale model state variables into fine-scale $\widehat{Tb}_{i,j,k}^- = \mathbf{h}_p\left(\left[\widehat{\mathbf{x}}_{i,j}\right]_k\right)$ using a radiative transfer model $\mathbf{h}_p(\cdot)$ and (ii) aggregating the fine-scale $\widehat{Tb}_{i,j,k}^-$ ($k = 1, \dots, N_\kappa$) to a coarse-scale $\widehat{Tb}_{i,j,\kappa}^-$ (Fig. 2a). The observation operator $\mathbf{h}(\cdot)$ combines these two operations, so that

$$\hat{\mathbf{y}}_{i,j}^- = \mathbf{h}(\hat{\mathbf{x}}_{i,j}^-) = \begin{bmatrix} \mathbf{Tb}_1 \\ \vdots \\ \mathbf{Tb}_k \\ \vdots \\ \mathbf{Tb}_m \end{bmatrix}_{i,j}^- = \begin{bmatrix} \frac{1}{N_1} \sum_{k=1}^{N_1} \mathbf{h}_p(\hat{\mathbf{x}}_{i,j}^-)_k \\ \vdots \\ \frac{1}{N_k} \sum_{k=1}^{N_k} \mathbf{h}_p(\hat{\mathbf{x}}_{i,j}^-)_k \\ \vdots \\ \frac{1}{N_m} \sum_{k=1}^{N_m} \mathbf{h}_p(\hat{\mathbf{x}}_{i,j}^-)_k \end{bmatrix} \quad \text{with}$$

$$\hat{\mathbf{x}}_{i,j}^- = \begin{bmatrix} \widehat{\text{sfmc}}_{i,j} \\ \widehat{\text{rzmc}}_{i,j} \\ \widehat{\text{tsurf}}_{i,j} \\ \widehat{\text{tsoil}}_{i,j} \end{bmatrix}_k \in \hat{\mathbf{x}}_{i,j}^-.$$
(21)

The Kalman gain to update each fine-scale $\hat{\mathbf{x}}_{i,j}^-_k$ is found as

$$[\mathbf{K}_i]_k = \text{Cov}([\hat{\mathbf{x}}_i^-]_k, \hat{\mathbf{y}}_i^-) [\text{Cov}(\hat{\mathbf{y}}_i^-, \hat{\mathbf{y}}_i^-) + \mathbf{R}_i]^{-1} \quad (22)$$

where $\text{Cov}([\hat{\mathbf{x}}_i^-]_k, \hat{\mathbf{y}}_i^-)$ is the error covariance between the fine-scale state variables and the coarse-scale observation predictions. The Kalman gain thus effectively partitions (Fig. 2b) the information in the coarse-scale \mathbf{Tb}_{obs} observations to fine-scale increments in soil moisture and temperature, using the error cross-correlations between fine-scale state variables, such as soil moisture, and coarse-scale \mathbf{Tb}_k observation predictions. The satellite-observed \mathbf{Tb}_{obs} only provides information about the top (~ 5 cm) layer soil moisture, but data assimilation propagates this surface information to deeper soil moisture layers through the model soil profile dynamics and the vertical error correlations between state variables.

The added advantage of 3D filtering is primarily in data-sparse regions, because information is horizontally propagated through the spatial forecast error structure. The spatial error correlations that are expressed in the off-diagonal elements of the cross-covariance matrix $\text{Cov}([\hat{\mathbf{x}}_i^-]_k, \hat{\mathbf{y}}_i^-)$ allow updating the state $\hat{\mathbf{x}}_{i,j}^-_k$ at the fine-scale location k with surrounding multiple coarse-scale innovations, as long as the latter are within the influence area around the fine-scale location (even if the state variable $\hat{\mathbf{x}}_{i,j}^-_k$ is “unobserved” and not part of any coarse-scale observation prediction). The influence area is typically obtained by localizing the spatial error correlations (discussed in Sect. 7.2.1) to limit the impact of distant observations.

Finally, it should be noted that the (spatial) observation vector can be further expanded by assimilating multiple types of observations simultaneously. For example, brightness temperatures are typically observed at two polarizations (H and V), and possibly at multiple incidence angles (e.g., SMOS). A 3D EnKF using both H- and V-polarized brightness temperatures is used for an operational SMAP data assimilation product as discussed in Sect. 9.2.

6.2 Smoothing

Smoothers update state vectors that are distributed in time. Smoothers have the potential to improve soil moisture reanalyses by assimilating multiple observations in time or time-integrated observations, such as, for example, TWS or river discharge. Here, a smoother is illustrated for the assimilation of monthly coarse-scale TWS as an extension of a spatially distributed (“3D”) ensemble Kalman filter in which the fine-scale state is distributed in time. The relevant model prognostic variables included in the state vector are the depth to the groundwater table (gwt) and root-zone soil moisture (rzmc). The concept is first introduced by updating each member j of the time-augmented ($i = 1, \dots, T$) state vector at a fine-scale location k using the traditional ensemble Kalman filter equations:

$$\begin{bmatrix} \widehat{\text{gwt}}_{1,j} \\ \text{rzmc}_{1,j} \\ \dots \\ \widehat{\text{gwt}}_{i,j} \\ \text{rzmc}_{i,j} \\ \dots \\ \widehat{\text{gwt}}_T \\ \text{rzmc}_T \end{bmatrix}_k^+ = \begin{bmatrix} \widehat{\text{gwt}}_{1,j} \\ \text{rzmc}_{1,j} \\ \dots \\ \widehat{\text{gwt}}_{i,j} \\ \text{rzmc}_{i,j} \\ \dots \\ \widehat{\text{gwt}}_T \\ \text{rzmc}_T \end{bmatrix}_k^- [\mathbf{K}]_k \left(\begin{bmatrix} \text{TWS}_1 \\ \dots \\ \text{TWS}_\kappa \\ \dots \\ \text{TWS}_m \end{bmatrix}_{\text{obs}} - \begin{bmatrix} \widehat{\text{TWS}}_1 \\ \dots \\ \text{TWS}_\kappa \\ \dots \\ \text{TWS}_m \end{bmatrix}_j \right) \quad (23)$$

The coarse-scale monthly (time index omitted) observation predictions ($\kappa = 1, \dots, m$) are obtained by first calculating fine-scale vertically integrated TWS using $\mathbf{h}_{\text{TWS}}(\cdot)$ and then aggregating the fine-scale state variables in space ($k = 1, \dots, N_\kappa$, for observation prediction κ) and time ($i = 1, \dots, T$):

$$\begin{aligned} \widehat{\mathbf{y}}_j^- &= \mathbf{h}\left(\widehat{\mathbf{x}}_j^-\right) = \begin{bmatrix} \text{TWS}_1 \\ \dots \\ \text{TWS}_\kappa \\ \dots \\ \text{TWS}_m \end{bmatrix}_j^- \\ &= \begin{bmatrix} \frac{1}{TN_1} \sum_{k=1}^{N_1} \sum_{i=1}^T \mathbf{h}_{\text{TWS}}\left(\left[\widehat{\mathbf{x}}_{i,j}^-\right]_k\right) \\ \dots \\ \frac{1}{TN_\kappa} \sum_{k=1}^{N_\kappa} \sum_{i=1}^T \mathbf{h}_{\text{TWS}}\left(\left[\widehat{\mathbf{x}}_{i,j}^-\right]_k\right) \\ \dots \\ \frac{1}{TN_m} \sum_{k=1}^{N_m} \sum_{i=1}^T \mathbf{h}_{\text{TWS}}\left(\left[\widehat{\mathbf{x}}_{i,j}^-\right]_k\right) \end{bmatrix} \quad \text{with } \left[\widehat{\mathbf{x}}_{i,j}^-\right]_k \\ &= \begin{bmatrix} \text{gwt}_{i,j} \\ \text{rzmc}_{i,j} \\ \dots \end{bmatrix}_k^- \end{aligned} \quad (24)$$

The above update equation (Eq. 23) could potentially involve a large-dimensional Kalman gain for which the error covariances are only valid if the number of included time steps (T) is small relative to the ensemble size, as in Dunne and Entekhabi (2006) who assimilated a batch of temporally distributed soil moisture observations and in Pauwels and De Lannoy (2009) who assimilated time-integrated discharge observations. Alternatively, and in analogy with strong constraint variational assimilation (Chap. X632X), one can limit the update to the initial conditions at the beginning (time step i_0) of the assimilation window ($\widehat{\mathbf{x}}_{i_0, k, j}^-$) so that the model trajectory over the entire smoothing window best fits the observations. The $[\mathbf{K}_i]_k$ matrix then becomes

$$[\mathbf{K}_{i_0}]_k = \text{Cov}\left(\left[\widehat{\mathbf{x}}_{i_0}^-\right]_k, \widehat{\mathbf{y}}^-\right) [\text{Cov}(\widehat{\mathbf{y}}^-, \widehat{\mathbf{y}}^-) + \mathbf{R}]^{-1} \quad (25)$$

Yet, adding a full increment to the initial soil moisture state $\widehat{\mathbf{x}}_{i_0, k, j}^-$ alone may not cause the desired persistent trajectory shift in hydrologic models. Instead, Zaitchik et al. 2008 computed an effective increment at each time step i by equally distributing the increment over the T time steps in the smoothing window, so that

$$\left[\widehat{\mathbf{x}}_{i,j}^+\right]_k = \left[\widehat{\mathbf{x}}_{i,j}^-\right]_k + \frac{1}{T} [\mathbf{K}_{i_0}]_k \begin{pmatrix} \begin{bmatrix} \text{TWS}_1 \\ \dots \\ \text{TWS}_k \\ \dots \\ \text{TWS}_m \end{bmatrix}_{\text{obs}} - \begin{bmatrix} \widehat{\text{TWS}}_1 \\ \dots \\ \widehat{\text{TWS}}_k \\ \dots \\ \widehat{\text{TWS}}_m \end{bmatrix} \end{pmatrix}, \text{ for each } i = 1, \dots, T \quad (26)$$

The TWS innovations are thus partitioned in time and space and into different water storage components. The methodological development for smoothing GRACE observations in the context of soil moisture assimilation is still in its infancy. Alternative formulations are under investigation.

6.3 Joint State and Parameter Updating

The above examples update the land surface state in response to observations related to soil moisture. The term “data assimilation” can also be used to estimate model parameters using observations related to soil moisture. Such parameter estimates could be static, as typically obtained after optimization of long-term statistics that involve differences between long time series of simulations and observations. Slowly varying parameters could be updated dynamically through recursive filtering, using similar techniques as described above, but after replacing (i) the state variables with parameters and (ii) the prognostic state propagation model with a persistent model. A combined state and parameter estimation has also been explored

(e.g., Montzka et al. 2013), but not for large-scale soil moisture modeling systems: the realism and observability of evolving parameters in interaction with dynamic state updates may pose difficulties.

7 Random Errors and Biases

The key to successful data assimilation is to qualify and quantify the errors in the model forecasts and observations. The errors include both random and systematic error, or bias. In the absence of knowledge about the true ensemble error at each time instant or location, land surface modelers often turn to the ergodicity principle and analyze errors in time series or spatial patterns. In that sense, bias can be defined as autocorrelated error, and the correlation length determines the temporal or spatial scale of bias. For large-scale (continental, global) hydrometeorological modeling, random errors have autocorrelation lengths of less than a few days in time (micro-scale, mesoscale), whereas bias or systematic error has time scales of several weeks or more. The following sections discuss the treatment of random errors and biases specifically for soil moisture data assimilation.

7.1 Bias, Autocorrelated Error

Statistically “optimal” data assimilation techniques, such as Kalman filtering, rely on observations and forecasts with zero-mean errors (first moment). A typical problem with assimilating satellite observations of soil moisture, however, is that their climatology differs from that of the land model integrations. If biases cannot be addressed through model calibration, then one can either treat the bias *a priori* and perform anomaly assimilation (i.e., after mapping the observations to the model climatology) or estimate the bias dynamically inside the assimilation scheme. Either approach to dealing with bias reduces the average magnitude of the innovations and avoids that the model is pushed away from its own climatology through data filtering. The climatological rescaling techniques are based on *a priori* knowledge and thus require historical data, whereas online bias estimation methods update the bias estimates dynamically at each assimilation event. Another difference is that rescaling techniques could address discrepancies between datasets in higher order moments under the assumption of stationary differences in the first moment, whereas the online bias estimation methods are focused on resolving nonstationary differences in the first moments.

7.1.1 A Priori Static Bias Treatment

A commonly used approach in soil moisture assimilation systems is to remove long-term differences between observations and forecasts by rescaling the observations to the model climatology. Rescaling does not per se assign the systematic errors to either the model or the observations, but rather removes the total bias from the innovations.

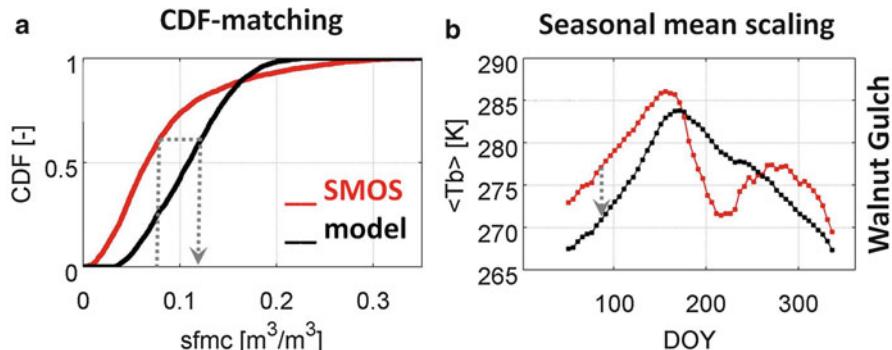


Fig. 3 Illustration of scaling techniques to remove bias from innovations in the assimilation soil moisture observations at a single 36-km grid cell in Walnut Gulch (WG3, see Table 3). **(a)** CDFs of soil moisture from SMOS retrievals and GEOS-5 simulations. **(b)** Seasonal climatology of SMOS-observed and simulated brightness temperatures (horizontal polarization, 40° incidence angle, ascending overpass, smoothed, and multiyear averaged). Both figures are based on 3 years (1 July 2010–1 July 2013) of data

A first approach to rescaling is to match the cumulative distribution function (CDF) of observed soil moisture values to the CDF of the soil moisture simulations (Reichle and Koster 2004; Drusch et al. 2005), thereby matching the long-term first, second, and higher order moments of the observations to the model. Figure 3a illustrates the CDF-matching approach. CDFs are calculated using a multiyear historical dataset of observations and land surface simulations for each location in space. This is illustrated by sampling SMOS retrievals and GEOS-5 simulations for ascending orbits (6:00 a.m. local time) at a 36-km grid cell inside the Walnut Gulch Experimental Watershed in Arizona (the USA). To complement the temporal sampling, additional sampling can be performed in a spatial window, which effectively smoothes the statistics. Here, all 36-km grid cells within a 0.5° radius around the central Walnut Gulch pixel are sampled. In this example, the SMOS retrievals are systematically drier than the simulations. To rescale an individual soil moisture retrieval (e.g., at $0.075 \text{ m}^3/\text{m}^3$), the cumulative probability density for this value is found (e.g., 0.62 [—]). Then, this probability is transferred to the model CDF, and the corresponding modeled soil moisture (e.g., $0.12 \text{ m}^3/\text{m}^3$) is found. In practice, this mapping happens by fitting a relationship between the observed and simulated soil moisture at identical cumulative probability densities. The observation rescaling possibly involves a change in the dynamical range of the observations, and therefore it is important to also rescale the original observation errors (\mathbf{R}). The error standard deviation $\sigma_{\text{sfmc,obs}}$ is rescaled to $\sigma'_{\text{sfmc,obs}}$, using the long-term (climatological) standard deviation of the observations $S[\widehat{\text{sfmc}}_{\text{obs}}]$ and simulations $S[\widehat{\text{sfmc}}]$, i.e.,

$$\sigma'_{\text{sfmc,obs}} = \sigma_{\text{sfmc,obs}} \cdot S[\widehat{\text{sfmc}}] / S[\widehat{\text{sfmc}}_{\text{obs}}] \quad (27)$$

The static nature of the CDF-matching approach is not always ideal, because it discards seasonal cycles in biases and may therefore even introduce biases into the system. In some applications, the CDF-matching approach has therefore been applied by season (de Rosnay et al. 2014).

The seasonality in biases is of particular concern when assimilating brightness temperatures, which are strongly impacted by seasonally varying surface temperature (Sect. 9.2). When it is important to resolve the seasonal and diurnal cycles of the climatology, rescaling is usually limited to the first moment and possibly the second moment, rather than the complete CDF, due to the limited availability of historical data. Figure 3b illustrates temporally variable biases between multiyear-averaged brightness temperatures from SMOS at 40° incidence angle and simulations with the GEOS-5 land system, both at 6:00 a.m. local time (ascending overpass). The temporally variable climatology of the mean brightness temperature is calculated by temporally smoothing the datasets (with the model cross masked for the availability of observations), and calculating a mean brightness temperature for each pentad (5-day period) averaged across 3 years. The climatological mean values of observed and modeled brightness temperature for each pentad p are given by $\langle \text{Tb}_{\text{obs}, p} \rangle$ and $\langle \widehat{\text{Tb}}_p^- \rangle$. Each individual $\text{Tb}_{\text{obs}, i}$ observation is then rescaled to $\text{Tb}'_{\text{obs}, i}$ using climatological information from the closest pentad p :

$$\text{Tb}'_{\text{obs}, i} = \text{Tb}_{\text{obs}, i} - \langle \text{Tb}_{\text{obs}, p} \rangle + \langle \widehat{\text{Tb}}_p^- \rangle = \text{Tb}_{\text{obs_anom}, i} + \langle \widehat{\text{Tb}}_p^- \rangle \quad (28)$$

where $\text{Tb}_{\text{obs_anom}, i}$ is the anomaly of the assimilated observations. The observation predictions can be similarly written as $\widehat{\text{Tb}}_i^- = \widehat{\text{Tb}}_{\text{anom}, i} + \langle \widehat{\text{Tb}}_p^- \rangle$, with $\widehat{\text{Tb}}_{\text{anom}, i}$ denoting the anomaly in the simulations. Consequently, the innovations can be written as

$$\text{Tb}'_{\text{obs}, i} - \widehat{\text{Tb}}_i^- = \text{Tb}_{\text{obs_anom}, i} - \widehat{\text{Tb}}_{\text{anom}, i} \quad (29)$$

This effectively means that anomaly information is assimilated when a rescaling approach is implemented.

7.1.2 Online Dynamic Bias Estimation

Bias can be estimated dynamically inside the data assimilation system, using information in the innovations. Without knowledge of the origin of persistent errors in the innovations, it is impossible to assign the bias to state forecast errors or observation errors.

Observation bias estimation removes bias from the innovations, and leaves the model and assimilation output in its own climatology. Similarly to observation rescaling techniques, dynamic observation bias estimation effectively aims at anomaly assimilation. The difference is that the climatological differences between the observation predictions and observations (e.g., $\langle \widehat{\text{Tb}}_p^- \rangle - \langle \text{Tb}_{\text{obs}, p} \rangle$ in Eq. 28)

are replaced by temporally variable bias estimates $\widehat{\mathbf{b}}_i^{\text{obs}}$ at each assimilation time step i . As an example, the state update equation (Eq. 5) would become

$$\widehat{\mathbf{x}}_i^+ = \widehat{\mathbf{x}}_i^- + \mathbf{K}_i \left[\mathbf{y}_{\text{obs}, i} - \mathbf{h}_i(\widehat{\mathbf{x}}_i^-) - \widehat{\mathbf{b}}_i^{\text{obs}} \right] \quad (30)$$

Forecast bias estimation techniques, on the other hand, remove state bias from the innovations (similar to rescaling techniques) and correct the model simulations for bias (i.e., they assign bias to the model, unlike rescaling techniques), either only in post-processing or with feedback into the model. Given a dynamic estimate of the forecast bias $\widehat{\mathbf{b}}_i^f$, the state update equation will be

$$\widehat{\mathbf{x}}_i^+ = \left(\widehat{\mathbf{x}}_i^- - \widehat{\mathbf{b}}_i^f \right) + \mathbf{K}_i \left[\mathbf{y}_{\text{obs}, i} - \mathbf{h}_i \left(\widehat{\mathbf{x}}_i^- - \widehat{\mathbf{b}}_i^f \right) \right] \quad (31)$$

For soil moisture data assimilation, the online forecast bias estimation technique has only been used in small-scale research studies (De Lannoy et al. 2007), but not yet in large-scale or operational applications. Forecast bias estimation and a correction of soil moisture simulations are particularly useful if knowledge of absolute levels of soil moisture is important in addition to knowledge of the temporal evolution of soil moisture. However, the main problem is that simple forecast bias models do not allow to propagate information from observed (e.g., surface soil moisture) to unobserved (e.g., root-zone soil moisture) variables.

7.2 Random Error

The observation and forecast error covariances (second moments) determine the relative weight of the observations in the assimilation scheme as well as the spatial (horizontal and vertical) distribution of the analysis increments. The optimality of data assimilation system depends on how these uncertainties are defined upon input.

7.2.1 Forecast Error Covariance

The various data assimilation techniques described above differ in how they handle the forecast error covariance matrix \mathbf{P}_i^- . For soil moisture assimilation, ensemble simulations have arguably become the most popular method to dynamically estimate the forecast error variance and the inter-variable error correlations (Reichle et al. 2002). This approach is also most directly relevant for ensemble hydrometeorological forecasting (chapters ▶ “Major Operational Ensemble Prediction Systems (EPS) and the Future of EPS” and ▶ “Hydrological Ensemble Prediction Systems around the Globe”). In ensemble soil moisture assimilation, the ensemble is usually generated by perturbing the forcings, model prognostic variables, and possibly land model parameters. In practice, the data assimilation system is calibrated by tuning of the magnitude of the perturbations to ensure that the innovations and increments

show the expected behavior (Sect. 7.2.3) and the assimilation results perform well (Sect. 8). Some a priori ensemble verification (De Lannoy et al. 2006, chapter ▶ “Verification Metrics for Hydrological Ensemble Forecasts”) could also be performed before activating the data assimilation. Good ensembles should envelop the observations with an ensemble standard deviation that is comparable to the root-mean-square error between the ensemble mean predictions and the observations.

Perturbations to model forcing and prognostic variables are usually applied at regular time intervals during the model integration, because the dispersion in soil moisture simulations is bounded and would collapse unless model forcing or prognostic variables are applied. Alternatively, perturbing land model parameters ensures a persistent ensemble spread, but this approach creates temporal error autocorrelation and may not reflect random errors properly.

As an example, Table 2 shows preliminary perturbations that are currently used for brightness temperature assimilation in the GEOS-5 land data assimilation system (Sect. 9.2). Select model prognostic variables and forcings are perturbed at each model time step, and land model parameters are not perturbed. Perturbations to radiation, temperature, and humidity are normally distributed and additive, whereas precipitation perturbations are lognormally distributed and multiplicative, because of the skewed nature of precipitation distributions. The forcing perturbations are applied with a temporal autocorrelation of 1 day and a spatial correlation scale of 50 km. Perturbations to the soil moisture (catdef, rzexc, srfexc) model prognostic variables (Sect. 4) are additive and applied with shorter correlation scales in space and time. Soil temperature is not explicitly perturbed to avoid excessive ensemble spreads in areas without vegetation, but it is indirectly perturbed through the perturbation of the radiation. Cross-correlations are also imposed to ensure some physical realism between the perturbed fields. The ensemble mean for all perturbations is constrained to zero for additive perturbations and to one for multiplicative perturbations.

When spatial filtering is applied, spatial error correlations are critical to ensure meaningful updates to the fine-scale state variables. The coarse-scale observation prediction error variance obtained by spatially aggregating N-independent fine-scale state variables equals the fine-scale error variance divided by N, i.e., aggregation causes a reduction in uncertainty. With inclusion of appropriate spatial correlations, the spatially aggregated uncertainty of the observation predictions is effectively increased. The same holds for temporal smoother applications, where temporal autocorrelation is important to ensure sufficient uncertainty in the observation predictions (relative to the observations).

As opposed to the useful spatial autocorrelations, the spurious long-range correlations are detrimental statistical artifacts in diagnosed ensemble \mathbf{P}_i^- matrices. To limit these spurious correlations and to reduce the undesirable impact of distant observations, ensemble filter techniques mostly include some covariance localization (Reichle and Koster 2003). Correlations beyond a certain separation distance are suppressed by using a Hadamard product with a local compactly supported correlation function that reaches zero beyond a given distance. This product is applied to the sample covariance terms $\text{Cov}(\hat{\mathbf{x}}_i^-, \hat{\mathbf{y}}_i^-)$ and $\text{Cov}(\hat{\mathbf{y}}_i^-, \hat{\mathbf{y}}_i^-)$ in the expression for the Kalman gain.

Table 2 Example of perturbations to forcing and model prognostic variable in the GEOS-5 land data assimilation system for brightness temperature assimilation. Values are from the preliminary system calibration used for the results discussed in Sect. 9.2

Perturbation	Additive (A) or multiplicative (M)	Standard deviation	AR(1) time series correlation scale	Spatial correlation scale	Cross-correlation with perturbations in P	Cross-correlation with perturbations in SW	Cross-correlation with perturbations in LW
Precipitation (P)	M	0.5	24 h	50 km	n/a	-0.8	0.5
Downward shortwave (SW)	M	0.3	24 h	50 km	-0.8	n/a	-0.5
Downward longwave (LW)	A	20 W/m ²	24 h	50 km	0.5	-0.5	n/a
Catchment deficit (catdef)	A	0.24 kg/m ² /h	3 h	50 km	n/a	0.0	
Surface excess (srfexc)	A	0.16 kg/m ² /h	3 h	50 km	0.0	n/a	

Finally, the number of ensemble members needs to allow a statistically valid diagnosis of the forecast error covariance matrix and should thus increase with the number of variables in the state. Most soil moisture assimilation applications use 10–100 ensemble members.

7.2.2 Observation Error Covariance

Observation uncertainty refers to all errors present in observation space, i.e., uncertainty due to sensor error, retrieval error, and representativeness error. Sensor error is given by instrument design specifications. For example, the Microwave Imaging Radiometer with Aperture Synthesis instrument on board SMOS measures multi-angular brightness temperatures with a radiometric error of 4 K (Kerr et al. 2010), and the radiometer on board SMAP is designed to measure brightness temperatures at a single 40° incidence angle with an error of 1.3 K (Enthekabi et al., 2014). Retrieval error depends on the radiative transfer or backscattering model, dynamic auxiliary information, and parameters. Representativeness error is often due to a horizontal or vertical mismatch between the observation and the observation predictions. For example, a collection of model grid cells may not accurately capture the actual sensor field of view area. Moreover, the penetration depth of microwaves decreases with increasing soil moisture content, whereas the simulated soil moisture is valid for a surface layer with a fixed thickness. Observation error also includes representativeness errors in the observation operator $\mathbf{h}_i(\cdot)$. This error specifically refers to uncertainty in the observation operator (e.g., due to erroneous parameters) and does not include errors in the state forecast $\widehat{\mathbf{x}}_i^-$. One promising method to obtain an estimate of soil moisture observation errors is the triple collocation procedure (Scipal et al. 2008). This method uses three independent estimates of soil moisture in order to estimate the uncertainty in one of them. However, the success of this method relies on conditions that are often difficult to meet in practice.

7.2.3 Error Optimization and Adaptive Filtering

The specification of error parameters as input to the data assimilation system is often user-defined and not necessarily optimal. The innovations provide a means to assess the filter behavior: for a linear system, the innovations should have a Gaussian distribution with zero mean and a covariance of $[\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]_i$. In practice, land surface models are nonlinear, and the Gaussian assumption is difficult to meet. However, the consistency of the innovations with the imposed a priori error covariances and observation error covariances (Reichle et al. 2002) can be (approximately) verified by normalizing the (ensemble mean) innovation $\langle \mathbf{y}_{\text{obs},i} - \mathbf{H}_i \widehat{\mathbf{x}}_i^- \rangle$ for each point in space and at each assimilation time step by the imposed $[\mathbf{H}\mathbf{P}^-\mathbf{H}^T + \mathbf{R}]_i$. The distribution of these normalized innovations (in time and/or space, ergodic sampling) should follow a normal distribution. If the standard deviation is less (larger) than one, then the values for \mathbf{R}_i and/or \mathbf{P}_i^- are too large (small). This can be used to manually optimize the filter. The alternative is to automatically tune the forecast error covariance \mathbf{P}_i^- , or more specifically the model error component \mathbf{Q}_i , during the online cycling of the data assimilation system through adaptive filtering

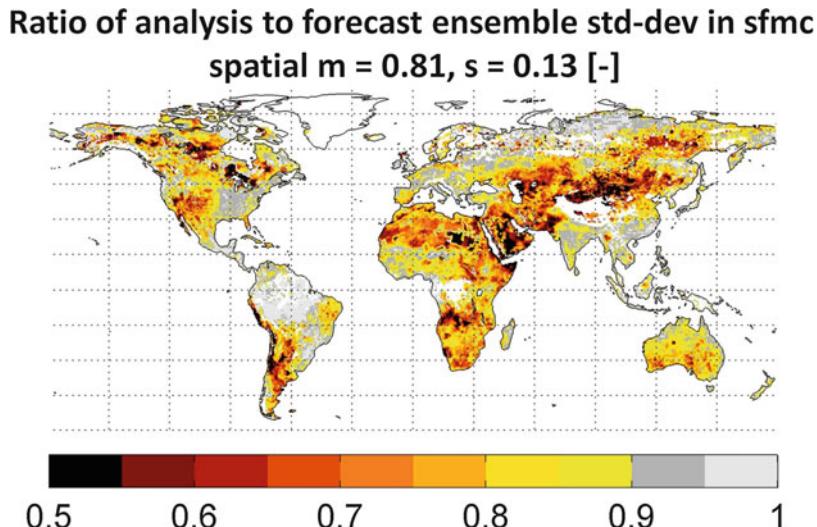


Fig. 4 Ratio of the analysis error standard deviation (std-dev) to the forecast ensemble standard deviation in surface soil moisture averaged over 3 years (July 2010–July 2013). The analysis is obtained by assimilating both ascending and descending SMOS soil moisture retrievals (v551) in the GEOS-5 land data assimilation system. Locations with less than 50 SMOS retrievals available for assimilation are shown in *white*. The spatial mean and standard deviation are denoted by *m* and *s* in the figure title

(Crow and Yilmaz 2014). However, the adaptive methods have not yet been thoroughly tested for large-scale data assimilation systems.

7.2.4 Analysis Error Covariance

A typical feature of data assimilation is that it reduces the analysis errors, both in terms of ensemble errors and in time series errors (see Sect. 8). Figure 4 illustrates how the ensemble analysis uncertainty in GEOS-5-simulated soil moisture is reduced compared to the forecast uncertainty, with a globally averaged fraction of 0.8, when assimilating SMOS retrievals during the period July 2010–July 2013.

8 Evaluation of Soil Moisture Estimates from Data Assimilation

8.1 In Situ Soil Moisture

Large-scale soil moisture data assimilation results are typically validated with in situ soil moisture observations, or some other measurements that depend on soil moisture, such as turbulent fluxes or river discharge. This section specifically

focuses on validation against independent in situ soil moisture measurements, i.e., observations that were not used in the data assimilation.

The most accurate measurements of soil moisture involve collecting soil samples that are weighed before and after drying to determine the amount of water present in the soil matrix. Yet, such destructive gravimetric measurements cannot be frequent in time or space, and therefore, they mainly serve to calibrate automated in situ soil moisture measuring devices, such as capacitance probes, time domain reflectometry probes, and neutron probes. Across the world, thousands of measurement stations are equipped with soil probes that record time series of soil moisture at various depths in the ground. Select examples of networks with more than 50 soil moisture measuring sites are the US Natural Resources Conservation Service Soil Climate Analysis Network (SCAN), the Snowpack Telemetry network (SNOWTELE), and the US Climate Reference Network (USCRN). Other in situ soil moisture networks include the Cosmic-Ray Soil Moisture Observing System (COSMOS) and Global Positioning System receivers (GPS), both of which use relationships between signals collected above the ground and vertically and horizontally integrated soil moisture, rather than using probes inserted in the ground. The limited spatial support of the point measurements complicates the comparison of spatial patterns or absolute values against gridded regional or global LSM simulations. However, the temporal variability in soil moisture at point locations provides useful information to validate the dynamics of model and assimilation results.

In addition, some watershed-averaged in situ measurements of surface soil moisture are available from core validation sites (Entekhabi et al. 2014, Chap. 7, pp.119–150) listed in Table 3. These USDA watersheds are equipped with locally dense sensor networks covering the area of a satellite footprint, which makes them directly relevant for the validation of remote-sensing retrievals, but are also very attractive to validate coarse-scale model or assimilation results. Table 3 provides details about the 36-km reference grid cells that are identified within each watershed for the evaluation of SMOS data assimilation in the example of Sect. 8.3 below. The soil moisture in each reference grid cell is computed as the average soil moisture across at least five profile sensors.

8.2 Validation Metrics

A number of metrics exist to validate simulation and assimilation results with in situ observations. It is often advised to focus on the temporal variability and use bias-free metrics for two reasons: (i) a comparison of gridded coarse-scale model output against point-scale in situ observations will suffer from representativeness biases and (ii) data assimilation for state updating alone is meant to correct for random errors and is not designed to fix any long-term biases between the model and observations. Examples of suitable metrics include the unbiased root-mean-square error (ubRMSE) (Entekhabi et al. 2010; Albergel et al. 2013) and the anomaly correlation coefficient.

Table 3 Details of core validation sites (Entekhabi et al. 2014) and 36-km reference grid cells within each site used for independent validation of the SMOS data assimilation experiment in Sect. 8.3. Each reference grid cell contains a minimum of 5 sensors and a maximum of N sensors. The latitude and longitude refer to the center of 36-km grid cells on the Equal-Area Scalable Earth Grid version 2

Core site	State (USA)	Reference grid cell	Latitude (°N)	Longitude (°W)	Maximum N sensors
Reynolds Creek	Idaho	RC1	43.33	116.70	6
		RC2	42.95	116.70	9
Walnut Gulch	Arizona	WG1	31.96	110.73	6
		WG2	31.62	110.35	14
		WG3	31.62	109.98	22
Little Washita	Oklahoma	LW	34.99	98.03	15
Fort Cobb	Oklahoma	FC	35.34	98.40	10
Little River	Georgia	LR1	31.62	83.84	15
		LR2	31.62	83.46	12
Saint Joseph	Indiana	SJ	41.43	84.96	15
South Fork	Iowa	SF	42.57	93.55	12

The RMSE between time series ($i = 1, \dots, T$) of simulated soil moisture content (mc_{est} , either forecasts or analyses) and in situ soil moisture measurements (mc_{insitu}) can be expanded as

$$RMSE = \sqrt{\frac{1}{T} \sum_{i=1}^T (mc_{est,i} - mc_{insitu,i})^2} \quad (32)$$

$$= \sqrt{S^2[mc_{est}] + S^2[mc_{insitu}] - 2 \cdot R \cdot S[mc_{est}] \cdot S[mc_{insitu}] + bias^2} \quad (33)$$

where $S[.]$ is the temporal standard deviation, R is the time series correlation coefficient, and “bias” is the difference between the long-term mean simulations and observations. This metric thus focuses on both similarities in temporal variability and on bias. The ubRMSE removes the bias term, or

$$ubRMSE^2 = RMSE^2 - bias^2 \quad (34)$$

and measures only the random error component of the RMSE. The anomaly time series correlation coefficient measures the linear correlation between simulations and observations after subtracting their respective seasonal climatologies, i.e., seasonally varying climatological mean values are subtracted from each data point. The temporal mean of the anomalies is zero by definition. The climatologies are obtained by smoothing the datasets (for a particular time of day, in case subdiurnal output is validated) and then calculating a multiyear average ($\bar{v}_{(i)}$) for each day, week, or month, depending on the temporal resolution. Note that the calculation of a climatology requires at least a few years of data. The anomaly correlation is thus given by

$$\text{anomR} = \frac{\sum_{i=1}^T (mc_{\text{est}, i} - \bar{mc}_{\text{est}(i)}) (mc_{\text{insitu}, i} - \bar{mc}_{\text{insitu}(i)})}{\sqrt{\sum_{i=1}^T (mc_{\text{est}, i} - \bar{mc}_{\text{est}(i)})^2} \sqrt{\sum_{i=1}^T (mc_{\text{insitu}, i} - \bar{mc}_{\text{insitu}(i)})^2}} \quad (35)$$

Examples of soil moisture data assimilation studies that used the anomaly correlation include Liu et al. (2011) and Draper et al. (2012).

8.3 Example

Figure 5 illustrates how SMOS retrieval assimilation, using a 1D EnKF with CDF matching, improves the ubRMSE of surface and root-zone soil moisture at select 36-km reference grid cells located in core validation watersheds across the USA (Table 3). The metrics are computed using 3-hourly model output and in situ data during the period July 2010–July 2013 at analysis time steps only (i.e., when SMOS observations are available) and excluding frozen conditions. The dark gray bars show that the current GEOS-5 system without assimilation (open loop) performs very well, with the ubRMSE at or below 0.04 m³/m³. When assimilating SMOS retrievals, the ubRMSE is further reduced in both the surface and

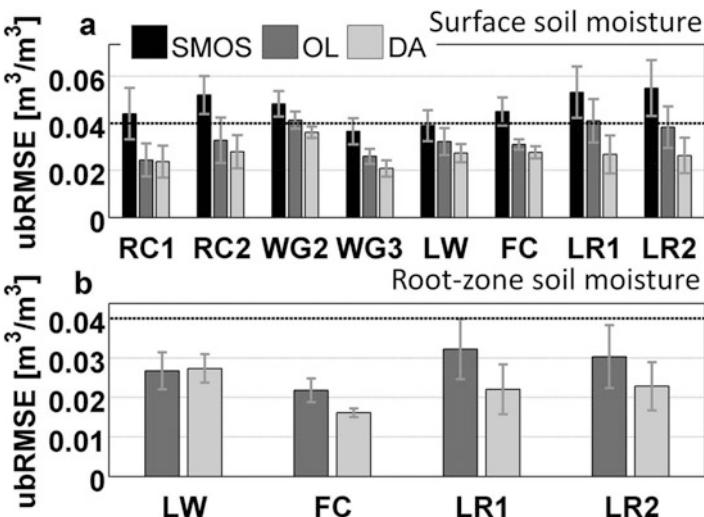


Fig. 5 Performance of (a) surface and (b) root-zone soil moisture in terms of ubRMSE for (black) both ascending and descending SMOS retrievals, (dark gray) open loop, and (light gray) SMOS retrieval data assimilation at various 36-km core validation sites across the USA (Table 3). The metrics are based on 3 years (1 July 2010–1 July 2013) of 3-hourly output at the analysis time steps only. The error bars indicate the 99% confidence intervals

root-zone, even though the SMOS retrievals have an uncertainty at or higher than $0.04 \text{ m}^3/\text{m}^3$. This highlights that data assimilation has the potential to improve model results, even if the assimilated observations are very uncertain. The confidence intervals are relatively large in this example, because only 3 years of data is used at the analysis time steps only. When including forecast time steps (not shown), the reductions in ubRMSE become statistically significant.

9 Toward Operational Soil Moisture Data Assimilation

A number of research and operational centers routinely generate data products that include a soil moisture analysis (in the broadest sense) at continental or global scales, either in reanalysis mode or to support operational prediction systems. Examples of data products generated with land-only systems include those from the North American Land Data Assimilation System (NLDAS, Xia et al. 2012) and the Global Land Data Assimilation System (GLDAS, Rodell et al. 2004) as well as the MERRA-Land and ERA-Land data products (Sect. 4). These data products primarily rely on the use of precipitation observations from gauges and satellites to improve the soil moisture simulations from atmospheric assimilation systems. The NLDAS and GLDAS data products use the NASA Land Information System (LIS, Kumar et al. 2008) software infrastructure, which is also integrated into the atmospheric assimilation systems at the US National Oceanic and Atmospheric Administration's (NOAA) National Centers for Environmental Prediction (NCEP) and the US Air Force Weather Agency. At NCEP, this combined land-atmosphere assimilation system is the basis for the Climate Forecast System Reanalysis (Saha et al. 2010).

The reanalysis data products benefit from the use of precipitation gauge information mainly because they are not subject to the strict latency constraints of atmospheric assimilation systems used for numerical weather prediction (NWP), which require observations to be available within hours. For NWP, the objective of soil moisture data assimilation is to initialize soil moisture conditions in order to provide the best possible accuracy and consistency of surface and near-surface weather forecasts in near-real time. Operational centers such as the UK Met Office (Dharssi et al. 2011), Deutscher Wetterdienst (Hess et al. 2008), Météo-France (Mahfouf et al. 2009), ECMWF (de Rosnay et al. 2014), and Environment Canada (Bélair et al. 2003) assimilate satellite soil moisture retrievals and/or screen-level observations to update the soil moisture state. This section describes two examples of soil moisture assimilation systems: (i) the soil moisture analysis in the ECMWF Integrated Forecasting System (IFS) and (ii) the NASA Goddard Earth Observing System Model, Version 5 (GEOS-5) land data assimilation system used to generate a soil moisture data assimilation product for the SMAP mission.

9.1 ECMWF Soil Moisture Data Assimilation for NWP

The ECMWF operational soil moisture analysis relies on a point-wise “simplified (E)KF” approach (de Rosnay et al. 2013, Sect. 6.1.1). For each grid point, the first three layers of soil moisture (of depth 0–7 cm, 7–28 cm, and 28–100 cm in the ECMWF IFS) are analyzed using observations of screen-level temperature and relative humidity, as well as ASCAT soil moisture retrievals. In this system, the observation operator is provided by the land surface model which gives the relation between screen-level temperature and humidity and soil moisture. Using screen-level observations as proxy information to analyze soil moisture has proved to be very relevant for NWP applications, because it consistently improves screen-level variables whose accurate forecast is a key objective of NWP. However, since screen-level observations are indirectly related to soil moisture, their assimilation is effective only in areas with strong coupling between soil moisture and screen-level variables. Recent developments at ECMWF focused on using satellite information from active and passive microwave sensors in addition to screen-level observations, to analyze soil moisture. In particular, ASCAT soil moisture data assimilation was recently implemented in operations in May 2015.

Figure 6 illustrates ECMWF soil moisture analysis components when screen-level observations are assimilated together with ASCAT surface soil moisture data, for a 6-day numerical experiment from 25 through 30 June 2013. The global NWP experiment was conducted at a resolution of 40 km as part of preoperational tests. The soil moisture data assimilation system accounts for soil moisture background errors fixed at $0.01 \text{ m}^3/\text{m}^3$ and observation errors of 1 K, 4%, and $0.04 \text{ m}^3/\text{m}^3$ for two-meter temperature, relative humidity, and ASCAT surface soil moisture, respectively. The ASCAT soil moisture index mean and range are rescaled to those of the ECMWF volumetric soil moisture using a seasonal (3-month moving window) CDF-matching approach. Innovations of two-meter temperature and humidity (Fig. 6c, e) indicate a good complementarity between two types of screen-level observations. They generally show warmer and/or drier conditions than the model over India and China, Eastern Siberia, and the northern part of South America, whereas colder and wetter conditions are observed in northern Canada, Eastern Europe, and Kazakhstan. In some areas two-meter temperature and humidity innovations indicate relatively patchy discrepancies between observations and the model (e.g., Australia). There is also good agreement between ASCAT and screen-level temperature and relative humidity innovations: in both cases, the observations indicate that the model is too wet over India and Eastern Siberia and too dry in the eastern part of South America. Over North America, ASCAT and relative humidity innovations are also consistent in the Great Lakes area, with wetter (drier) observed conditions north (south) of the lakes. However, in Russia, ASCAT and relative humidity innovations suggest soil moisture errors of opposite sign.

The right panel of Fig. 6 shows 6-day accumulated increments in the first soil layer (0–7 cm) due to ASCAT data assimilation (Fig. 6b), due to screen-level observations assimilation (Fig. 6d) and total surface soil moisture increments (Fig. 6f). It shows a good complementarity between ASCAT and screen-level data

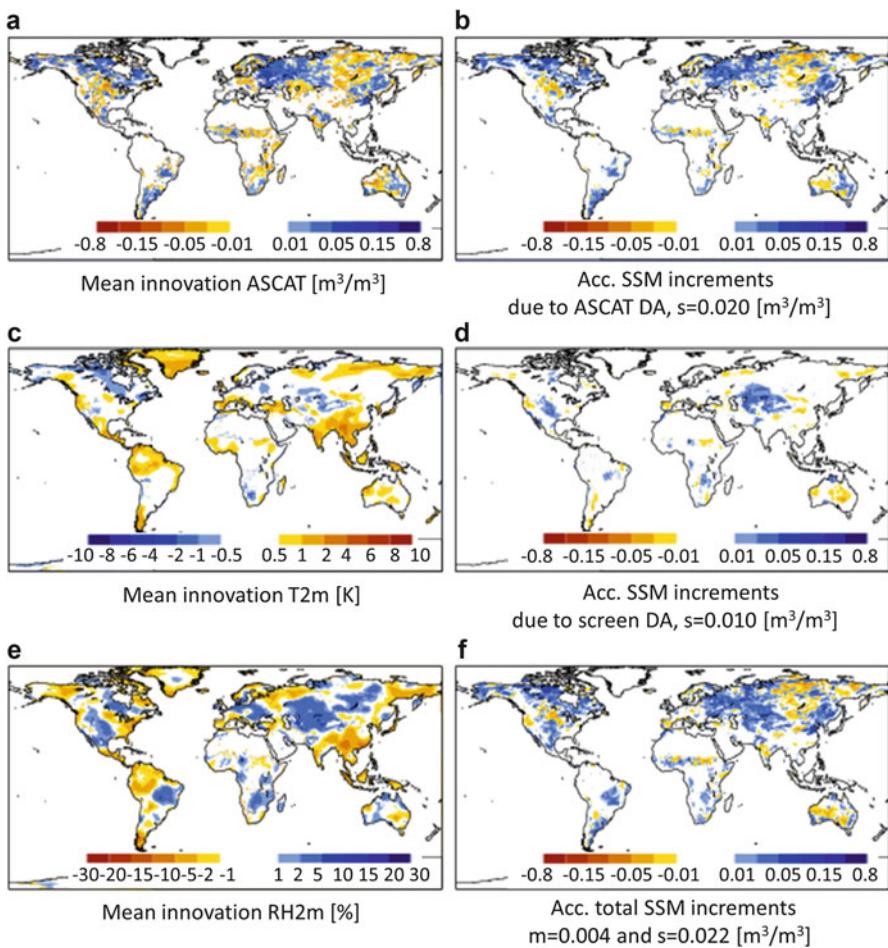


Fig. 6 Illustration of ECMWF soil moisture analysis (*left*) mean innovations and (*right*) accumulated soil moisture increments for a 6-day period from 25 through 30 June 2013. Innovations show the differences between observations and ECMWF first guess for (a) scaled ASCAT observations and ECMWF surface soil moisture (in m^3/m^3), (c) two-meter temperature (T2m), and (e) two-meter relative humidity (RH2m). Right panel shows accumulated surface soil moisture increments (m^3/m^3) due to (b) ASCAT data assimilation, (d) screen-level data assimilation, and (f) total accumulated increments. The spatial mean and standard deviation are denoted by m and s in the figure titles

assimilation in term of spatial distribution of the increments. ASCAT-induced increments are prominent at relatively high latitude, as well as in Argentina and Australia. Screen-level data assimilation mainly contributes to soil moisture increments in Kazakhstan and North America. It is interesting to notice that there is no contribution of screen-level observations to the soil moisture increments over India. This is because of rainy (monsoon) conditions that prevail at this time of the year in this region, leading to small values of the Jacobians (not shown) for screen-level

variables. The spatially averaged value of the combined accumulated increments is $0.004 \text{ m}^3/\text{m}^3$ for the first soil layer, with a standard deviation of $0.022 \text{ m}^3/\text{m}^3$ across the map. Spatially averaged (time series) standard deviation values of $0.02 \text{ m}^3/\text{m}^3$ and $0.01 \text{ m}^3/\text{m}^3$ are obtained for increments due ASCAT and screen-level assimilation, respectively. These values indicate the average magnitude of the increments and show that the contribution of ASCAT to the top layer soil moisture increments is larger than that of screen-level observations. This is consistent with the fact that satellite data provides direct information on surface soil moisture, whereas screen-level observations are only indirectly related to soil moisture (but including root-zone soil moisture).

For NWP applications the assimilation window length is generally relatively short (12 hours at ECMWF). Such a time window is shorter than most of the soil diffusion processes, and therefore the relation between observed surface soil moisture and root-zone soil moisture is weak. This results in low values of the Jacobian matrix elements that relate deep soil moisture to the surface soil moisture. Therefore ASCAT data assimilation mostly provides increments at surface. In the second soil layer (not shown), increments are smaller than in the first layer, with standard deviation of $0.0032 \text{ m}^3/\text{m}^3$ and $0.007 \text{ m}^3/\text{m}^3$ for ASCAT and screen-level contributions. In the third layer, most of the increments result from screen-level data assimilation with negligible increments due to ASCAT data assimilation. Although the screen-level observations provide only indirect information on soil moisture, their relation with root-zone soil moisture through soil-plant interaction processes makes them relevant to analyze root-zone soil moisture profile.

These results illustrate the complementarity between satellite information related to surface soil moisture and conventional observations related to root-zone soil moisture profiles, both in terms of spatial and vertical distributions of the increments. Future implementation plans for soil moisture data assimilation at ECMWF include the combined use of screen-level observations, ASCAT soil moisture, and SMOS and SMAP brightness temperature observations.

9.2 NASA SMAP Surface and Root-Zone Soil Moisture Product

Global estimates of surface soil moisture estimates are routinely obtained through satellite remote sensing, but many applications need an estimate of root-zone soil moisture. The NASA SMAP mission (Entekhabi et al. 2014), launched on 31 January 2015, provides a number of operational soil moisture data products, including Level 2 (half-orbit) and Level 3 (daily composite) soil moisture retrievals and a value-added Level 4 surface and root-zone soil moisture (L4_SM) data product (Entekhabi et al. 2014, Chap. 5, pp.89–100). This latter product is based on the assimilation of SMAP 36-km brightness temperatures into the NASA GEOS-5 Catchment land surface model, using a 3D ensemble Kalman filter.

Recall from Sect. 4 that the model prognostic variables related to soil moisture in the Catchment land surface model are the catchment deficit (*catdef*), root-zone excess (*rzexc*), and surface excess (*srfexc*). Other relevant model prognostic

variables are the land surface temperature (tsurf) and the near-surface ground heat content (ght) which determines the near-surface soil temperature (tsoil). These soil moisture variables as well as the forcings are suitably perturbed to generate an ensemble of forecasts (Table 2, Sect. 7.2.1). A radiative transfer model (Sect. 5.2) diagnoses brightness temperature based on the surface soil moisture and temperature in the assimilation system, i.e., $T_{b,i,j}^- = h(\hat{\mathbf{x}}_{i,j}^-)$ for each time step i and ensemble member j . This radiative transfer model is optimized to simulate a realistic long-term mean and variability in brightness temperature. The optimization is based on multi-angular SMOS observations and uses statistically optimized estimates of simulation and observation errors (De Lannoy et al. 2013). Consequently, long-term biases between observations and observation predictions [$T_{obs,i,j} - \widehat{T}_{b,i,j}^-$] are small by design, but seasonal biases are still present. To deal with the remaining biases, the instantaneous $T_{b,obs,i}$ are rescaled to the model climatology using seasonally varying means, as discussed in Sect. 7.1.1 (Fig. 3b). The data assimilation then maps the differences between the rescaled brightness temperature observations and simulations to increments in the prognostic variables of the Catchment land surface model.

Figure 7 illustrates the concept of assimilating coarse-scale (36 km) brightness temperatures into the GEOS-5 Catchment land surface model. In the absence of SMAP observations at the time of writing, SMOS (Kerr et al. 2010) brightness temperature observations are assimilated at 40° incidence angle, for both H- and V-polarization. For simplicity, it is assumed here that the state variables and observations are at the same coarse-scale 36-km resolution. Note however that the SMAP L4_SM product is produced at 9 km.

Figures 7a–b show the ensemble mean innovations $\langle T'_{b,obs,i} - \widehat{T}_{b,i}^- \rangle$ for (a) horizontal polarization and (b) vertical polarization and for both ascending and descending overpasses during 3 days. The swaths are relatively narrow, because of the strict quality control against aliased data and quality control on the angular fitting from multi-angular data to 40° incidence brightness temperature. Some swaths are also incomplete, because insufficient historical data are available (often due to radiofrequency interference, primarily over China, the Middle East, and Eastern Europe) to ensure a statistically reliable rescaling.

Figures 7c–f illustrate how the brightness temperature innovations are mapped to state increments. Negative brightness temperature innovations in Figs. 7a–b indicate that the model is too warm and/or too dry. Consequently, the negative innovations in the central western region of the USA and in the southwestern region of Australia result in an increase in soil moisture and a decrease in soil temperature. Likewise, positive innovations result in a decrease in soil moisture and an increase in soil temperature. Note that an increase in soil moisture corresponds to an increase in the water excess terms (srfe, rzex) and a decrease in the water deficit term (catdef). The magnitude of the increments to srfe, rzex, and catdef should be interpreted in relation to the storage capacity of each of these components (with equivalent soil layer thicknesses of 0.05, 1, and 1.3–10 m,

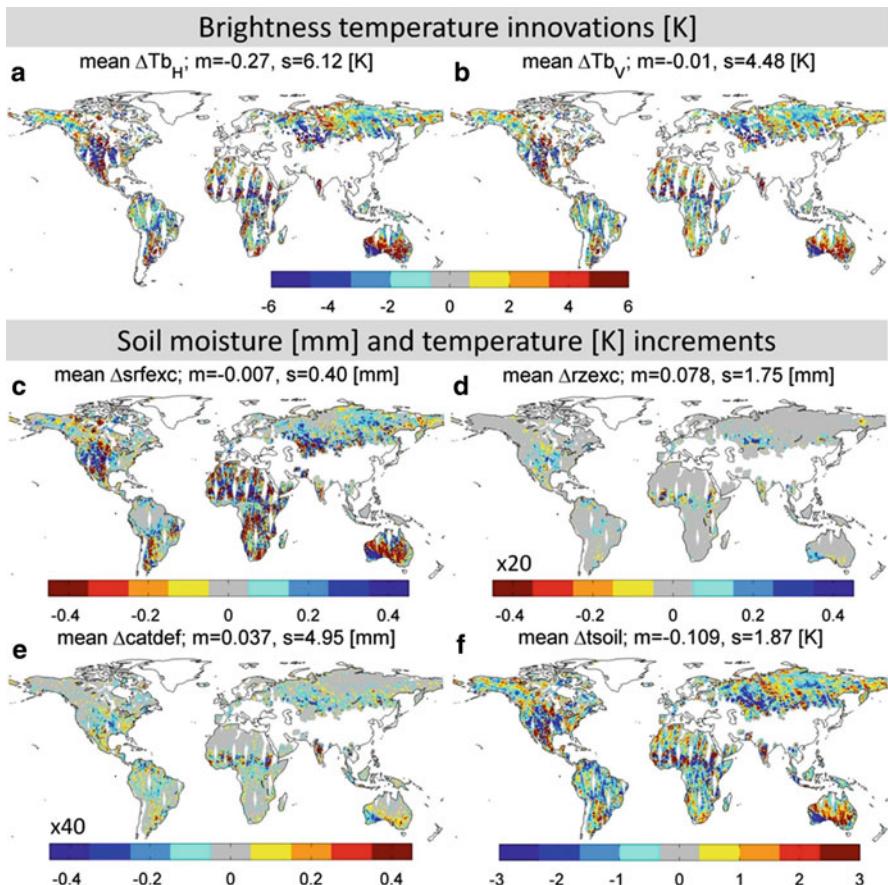


Fig. 7 (Top) Illustration of differences between scaled SMOS-observed and NASA GEOS-5 simulated brightness temperatures at 40° incidence angle for (a) H- and (b) V-polarization. (Bottom) Temporally averaged increments (Δ) to the Catchment land surface model (c) srfecc, (d) rzexc, (e) catdef, and (f) tsoil, after assimilation of SMOS brightness temperatures. The range of the color bars for rzexc is 20 times that of the srfecc, because rzexc applies to a 1-m root-zone, while srfecc applies to a 0.05-m surface layer. For catdef, the color-bar scaling factor of 40 is motivated by the average depth of bedrock (or profile soil moisture layer thickness) of 2 m. The time period covers ascending and descending orbits for 3 days from 27 through 29 July 2013. The spatial mean and standard deviation are denoted by m and s in the figure titles

respectively), and the color-bar range is scaled accordingly. While the absolute values of the increments are largest for catdef (global standard deviation of about 5 mm, pertaining to the entire profile), the increments relative to the layer depth are largest in the 5-cm surface layer (srfecc).

The spatial filtering introduces some additional important features: the increments are spatially smoothed, and the spatial coverage of the increments is wider than the coverage of brightness temperature innovations. The spatial error correlations also

introduce a horizontal propagation of information to unobserved areas. In this example, the support of the forecast error correlation function is limited to 1.25° , and increments thus taper off over a distance of 1.25° away from the observed swath. The L4_SM product is generated on a 9-km model grid, and the spatial filtering further improves soil moisture estimation through downscaling of 36-km brightness temperature innovations.

The L4_SM product is not limited to global surface and root-zone soil moisture. Research output includes other land surface state variables such as soil temperature and snow, as well as land surface fluxes and meteorological forcings. In addition, ensemble-derived error estimates are provided.

10 Conclusions

Soil moisture is a key variable in hydrological and Earth modeling and assimilation systems. Good estimates of soil moisture at regional to global scales are important for predictions of weather and climate, agricultural productivity, and natural hazards and for various other environmental and socioeconomic applications. Data assimilation provides a means to obtain enhanced soil moisture estimates by combining (often indirect) observations of soil moisture with land surface modeling. At large spatial scales, observations related to soil moisture are mainly provided by global surface observational networks or through remote sensing, either in the form of satellite retrievals, microwave radiances, or backscatter values. This chapter provides conceptual examples on how to assimilate these observations with widely accepted assimilation techniques, such as optimal interpolation and various types of Kalman filtering and smoothing. Special attention is paid to practical issues, such as dealing with multiple scales, the treatment of typical biases, and the characterization of random forecast and observation errors.

Operational centers have used screen-level observations of temperature and relative humidity to update soil moisture and temperature for numerical weather prediction. Satellite-based microwave observations provide more direct measurements of surface soil moisture and recent satellite missions such as ASCAT, SMOS, and SMAP are aiming at continued and improved surface soil moisture observations. The assimilation of satellite-based surface soil moisture retrievals has the capability to improve both surface and root-zone soil moisture, as illustrated in various research applications. However, the soil moisture retrieval process relies on parameters and a priori information that may be inconsistent with the land surface model used in the assimilation system. It is thus more natural to couple a radiative transfer or backscatter model to a land surface model, and then directly assimilate microwave observations such as brightness temperature or backscatter. Further improvements in continental-scale root-zone soil moisture can perhaps be obtained from the assimilation of integrated terrestrial water storage observations (e.g., from GRACE).

The growing experience with assimilation of soil moisture observations is reflected in the preparation of new cutting-edge data assimilation systems for operational applications. This chapter provides details on two of these systems. A first example shows how ASCAT surface soil moisture retrievals will be assimilated along with screen-level observations for numerical weather prediction at ECMWF. A second example discusses the assimilation of brightness temperature observations from SMOS to prepare for the operational global SMAP surface and root-zone product (L4_SM) at NASA. Both these systems benefit from increasingly available computational power as well as from recent and future satellite missions that are specifically designed to measure soil moisture. These operational systems are able to provide improved soil moisture estimates that have the potential to improve hydro-meteorological forecasting across a range of applications such as weather forecasting and the monitoring and prediction of droughts.

References

- C. Albergel, W. Dorigo, R. Reichle, G. Balsamo, P. de Rosnay, J. Muñoz-Sabater, L. Isaksen, R. de Jeu, W. Wagner, Skill and global trend analysis of soil moisture from reanalyses and microwave remote sensing. *J. Hydrometeorol.* **14**, 1259–1277 (2013). <https://doi.org/10.1175/JHM-D-12-0161.1>
- G. Balsamo, J.F. Mahfouf, S. Bélair, G. Deblonde, A global root-zone soil moisture analysis using simulated L-band brightness temperature in preparation for the Hydros satellite mission. *J. Hydrometeorol.* **7**, 1126–1146 (2006)
- G. Balsamo, P. Viterbo, A. Beljaars, B. van den Hurk, M. Hirschi, A.K. Betts, K. Scipal, A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system. *J. Hydrometeorol.* **10**, 623–643 (2009). <https://doi.org/10.1175/2008JHM1068.1>
- G. Balsamo, C. Albergel, A. Beljaars, S. Boussetta, H. Cloke, D. Dee, E. Dutra, J. Muñoz-Sabater, F. Pappenberger, P. de Rosnay, T. Stockdale, F. Vitart, ERA-Interim/Land: a global land water resources dataset. *Hydroclim. Earth Syst. Sci.* **10**, 14705–14745 (2013). <https://doi.org/10.5194/hessd-10-14705-2013>
- Z. Bartalis, W. Wagner, V. Naeimi, S. Hasenauer, K. Scipal, H. Bonekamp, J. Figa, C. Anderson, Initial soil moisture retrievals from the METOP-A advanced scatterometer (ASCAT). *Geophys. Res. Lett.* **34**, L20401 (2007). <https://doi.org/10.1029/2007GL031088>
- S. Bélair, L.P. Crevier, J. Mailhot, B. Bilodeau, Y. Delage, Operational implementation of the ISBA land surface scheme in the Canadian regional weather forecast model. Part I: warm season results. *J. Hydrometeorol.* **4**, 352–370 (2003)
- W.T. Crow, M.T. Yilmaz, The auto-tuned land data assimilation system (ATLAS). *Water Resour. Res.* **50**, 371–385 (2014). <https://doi.org/10.1002/2013WR014550>
- G.J.M. De Lannoy, P.R. Houser, V.R.N. Pauwels, N.E. Verhoest, Assessment of model uncertainty for soil moisture through ensemble verification. *J. Geophys. Res.* **111**, D10101 (2009). <https://doi.org/10.1029/2005JD006367>
- G.J.M. De Lannoy, R.H. Reichle, P.R. Houser, V.R.N. Pauwels, N.E.C. Verhoest, Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter. *Water Resour. Res.* **43**, W09410 (2007). <https://doi.org/10.1029/2006WR00544>
- G.J.M. De Lannoy, R.H. Reichle, V.N.R. Pauwels, Global calibration of the GEOS-5 L-band microwave radiative transfer model over non-frozen land using SMOS observations. *J. Hydrometeorol.* **14**, 765–785 (2013). <https://doi.org/10.1175/JHM-D-12-092.1>

- P. de Rosnay, M. Drusch, D. Vasiljevic, G. Balsamo, C. Albergel, L. Isaksen, A simplified extended Kalman filter for the global operational soil moisture analysis at ECMWF. *Q. J. Roy. Meteorol. Soc.* **139**(674), 1199–1213 (2013). <https://doi.org/10.1002/qj.2023>
- P. de Rosnay, G. Balsamo, C. Albergel, J. Muñoz-Sabater, L. Isaksen, Initialisation of land surface variables for numerical weather prediction. *Surv. Geophys.* **35**(3), 607–621 (2014). <https://doi.org/10.1007/s10712-012-9207-x>
- I. Dharssi, K.J. Bovis, B. Macpherson, C.P. Jones, Operational assimilation of ASCAT surface soil wetness at the Met Office. *Hydrol. Earth Syst. Sci.* **15**, 2729–2746 (2011). <https://doi.org/10.5194/hess-15-2729-2011>
- P. Dirmeyer, Using a global soil wetness dataset to improve seasonal climate simulation. *J. Climate* **13**, 2900–2921 (2000)
- W.A. Dorigo, A. Gruber, R.A.M. de Jeu, W. Wagner, T. Stacke, A. Loew, C. Albergel, L. Brocca, D. Chung, R. Parinussa, R. Kidd, Evaluation of the ESA CCI soil moisture product using ground-based observations. *Remote Sens. Environ.* **162**, 380–395 (2015). <https://doi.org/10.1016/j.rse.2014.07.023>
- C.S. Draper, R.H. Reichle, G.J.M. De Lannoy, Q. Liu, Assimilation of passive and active microwave soil moisture retrievals. *Geophys. Res. Lett.* **39**, L04401 (2012). <https://doi.org/10.1029/2011GL050655>
- M. Drusch, E.F. Wood, H. Gao, Observation operators for the direct assimilation of TRMM microwave imager retrieved soil moisture. *Geophys. Res. Lett.* **32**, L15403 (2005). <https://doi.org/10.1029/2005GL023623>
- M. Drusch, K. Scipal, P. de Rosnay, G. Balsamo, E. Andersson, P. Bougeault, P. Viterbo, Towards a Kalman filter-based soil moisture analysis system for the operational ECMWF Integrated Forecast System. *Geophys. Res. Lett.* **36**, L10401 (2009). <https://doi.org/10.1029/2009GL037716>
- S. Dunne, D. Entekhabi, Land surface state and flux estimation using the ensemble Kalman smoother during the Southern Great Plains 1997 field experiment. *Water Resour. Res.* **42**, W01407 (2006)
- D. Entekhabi, H. Nakamura, E.G. Njoku, Solving the inverse problems for soil moisture and temperature profiles by sequential assimilation of multifrequency remotely-sensed observations. *IEEE Trans. Geosci. Remote Sens.* **32**, 438–448 (1994)
- D. Entekhabi, R.H. Reichle, R.D. Koster, W.T. Crow, Performance metrics for soil moisture retrievals and application requirements. *J. Hydrometeorol.* **11**, 832–840 (2010). <https://doi.org/10.1175/2010JHM1223.1>
- D. Entekhabi, S. Yueh, P. O'Neill, K. Kellogg, SMAP Handbook, NASA/JPL Publication JPL 400-1567, Pasadena, CA, USA, p. 182 (2014).
- A.K. Fung, Z. Li, K.S. Chen, Backscattering from a randomly rough dielectric surface. *IEEE Trans. Geosci. Remote Sens.* **30**, 356–369 (1992)
- D. Giard, E. Bazile, Implementation of a new assimilation scheme for soil and surface variables in a global NWP model. *Mon. Weather Rev.* **128**, 997–1015 (2000)
- P.H. Gleick, Water resources, in *Encyclopedia of climate and weather*, ed. by S.H. Schneider, vol. 2 (Oxford University Press, New York, 1996), pp. 817–823
- R. Hess, M. Lange, W. Werner, Evaluation of the variational soil moisture assimilation scheme at Deutscher Wetterdienst. *Hydrol. Earth Syst. Sci.* **134**(635), 1499–1512 (2008)
- Y. Kerr et al., The SMOS mission: new tool for monitoring key elements of the global water cycle. *Proc. IEEE* **98**, 666–687 (2010)
- R.D. Koster, M.J. Suarez, A. Ducharme, M. Stieglitz, P. Kumar, A catchment-based approach to modeling land surface processes in a general circulation model 1. Model structure. *J. Geophys. Res.* **105**(D20), 24,809–24,822 (2000)
- R.D. Koster, P.A. Dirmeyer, Z. Guo, G. Bonan, P. Cox, C. Gordon, S. Kanae, E. Kowalczyk, D. Lawrence, P. Liu, C. Lu, S. Malyshev, B. McAvaney, K. Mitchell, D. Mocko, T. Oki, K. Oleson, A. Pitman, Y. Sud, C. Taylor, D. Verseghy, R. Vasic, Y. Xue, T. Yamada, Regions of strong coupling between soil moisture and precipitation. *Science* **305**, 1138–1140 (2004)
- S. Kumar, C. Peters-Lidard, Y. Tian, R. Reichle, J. Geiger, C. Alonge, J. Eylander, P. Houser, An integrated hydrologic modeling and data assimilation framework. *IEEE Comput.* **41**, 52–59 (2008). <https://doi.org/10.1109/MC.2008.511>

- Q. Liu, R.H. Reichle, R. Bindlish, M.H. Cosh, W.T. Crow, R. de Jeu, G.J.M. De Lannoy, G.J. Huffman, T.J. Jackson, The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in a land data assimilation system. *J. Hydrometeorol.* **12**, 750–765 (2011). <https://doi.org/10.1175/JHM-D-10-05000>
- J.F. Mahfouf, K. Bergaoui, C. Draper, F. Bouyssel, F. Taillefer, L. Taseva, A comparison of two off-line soil analysis schemes for assimilation of screen level observations. *J. Geophys. Res.* **114**, D08105 (2009). <https://doi.org/10.1029/2008JD011077>
- T. Mo, B.J. Choudhury, T.J. Schmugge, J.R. Wang, T.J. Jackson, A model for microwave emission from vegetation-covered fields. *J. Geophys. Res. Oceans Atmos.* **87**(C13), 1229–1237 (1982)
- C. Montzka, J.P. Grant, J. Moradkhani, H.J. Hendricks-Franssen, L. Weihermüller, M. Drusch, H. Vereecken, Estimation of radiative transfer parameters from L-band passive microwave brightness temperatures using advanced data assimilation. *Vadose Zone J.* **12**(3), 1–17 (2013). <https://dl.sciencesocieties.org/publications/vzj/pdfs/12/3/vzj2012.0040>
- M. Pan, E.F. Wood, R. Wojcik, M.F. McCabe, Estimation of regional terrestrial water cycle using multi-sensor remote sensing observations and data assimilation. *Remote Sens. Environ.* **112**, 1282–1294 (2008)
- V.R.N. Pauwels, G.J.M. De Lannoy, Ensemble-based assimilation of discharge into rainfall-runoff models: a comparison of approaches to mapping observational information to state space. *Water Resour. Res.* **45**(8), W08428 (2009). <https://doi.org/10.1029/2008WR007590>
- R.H. Reichle, R.D. Koster, Assessing the impact of horizontal error correlations in background fields on soil moisture estimation. *J. Hydrometeorol.* **4**(6), 1229–1242 (2003)
- R.H. Reichle, R.D. Koster, Bias reduction in short records of satellite soil moisture. *Geophys. Res. Lett.* **31**, L19501 (2004). <https://doi.org/10.1029/2004GL020938>
- R.H. Reichle, D. Entekhabi, D. McLaughlin, Downscaling of radio brightness measurements for soil moisture estimation: a four dimensional variational data assimilation approach. *Water Resour. Res.* **37**, 2353–2364 (2001)
- R.H. Reichle, D.B. McLaughlin, D. Entekhabi, Hydrologic data assimilation with the ensemble Kalman filter. *Mon. Weather Rev.* **120**, 103–114 (2002)
- R.H. Reichle, R.D. Koster, G.J.M. De Lannoy, B.A. Forman, Q. Liu, S.P.P. Mahanama, A. Toure, Assessment and enhancement of MERRA land surface hydrology estimates. *J. Climate* **24**, 6322–6338 (2011)
- R.H. Reichle, G.J.M. De Lannoy, B.A. Forman, C.S. Draper, Q. Liu, Connecting satellite observations with water cycle variables through land data assimilation: examples using the NASA GEOS-5 LDAS. *Surv. Geophys.* **35**, 577–606 (2014). <https://doi.org/10.1007/s10712-013-9220-8>
- M. Rodell, P.R. Houser, U. Jambor, J. Gottschalck, K. Mitchell, C.J. Meng, K. Arsenault, B. Cosgrove, J. Radakovich, M. Bosilovich, J.K. Entin, J.P. Walker, D. Lohmann, D. Toll, The global land data assimilation system. *Bull. Am. Meteorol. Soc.* **85**(3), 381–394 (2004)
- M. Rodell, J. Chen, H. Kato, J.S. Famiglietti, J. Nigro, C.R. Wilson, Estimating groundwater storage changes in the Mississippi River basin (USA) using GRACE. *Hydrogeol. J.* **15**, 159–166 (2007)
- J. Sabater, L. Jarlan, J. Calvet, F. Bouyssel, P. de Rosnay, From near-surface to root-zone soil moisture using different assimilation techniques. *J. Hydrometeorol.* **8**(2), 194–206 (2007)
- S. Saha et al., The NCEP climate forecast system reanalysis. *Bull. Am. Meteorol. Soc.* ES9–ES24 (2010). <https://doi.org/10.1175/2010Bams3001.1>
- K. Scipal, T. Holmes, R. de Jeu, V. Naeimi, W. Wagner, A possible solution for the problem of estimating the error structure of global soil moisture data sets. *Geophys. Res. Lett.* **35**, L24403.1–L24403.4 (2008)
- J.P. Wigneron et al., L-band microwave emission of the biosphere (L-MEB) model: description and calibration against experimental data sets over crop fields. *Remote Sens. Environ.* **107**, 639–655 (2007)

- Y. Xia, K. Mitchell, M. Ek, J. Sheffield, B. Cosgrove, E. Wood, L. Luo, C. Alonge, H. Wei, J. Meng, B. Livneh, D. Lettenmaier, V. Koren, Q. Duan, K. Mo, Y. Fan, D. Mocko, Continental scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.* **117**, D03109 (2012). <https://doi.org/10.1029/2011JD016048>
- B.F. Zaitchik, M. Rodell, R.H. Reichle, Assimilation of GRACE terrestrial water storage data into a land surface model: results for the Mississippi river basin. *J. Hydrometeorol.* **9**, 535–548 (2008)



Assimilation of Streamflow Observations

Seong Jin Noh, Albrecht H. Weerts, Oldrich Rakovec, Haksu Lee,
and Dong-Jun Seo

Contents

1	Introduction	746
2	Streamflow Measurement and Associated Uncertainties	748
2.1	In Situ Stage-Discharge Measurement	748
2.2	Remote Sensing Measurement	749
2.3	Direct and Crowdsourced Measurement	750
2.4	Observational Uncertainty in Streamflow DA	750
3	Applications of Streamflow Assimilation	751
3.1	Input Uncertainty Modeling	751
3.2	Asynchronous EnKF (AEnKF) with DHM	754
3.3	VAR with Hydrologic Routing Model	756
3.4	Weak- Versus Strong-Constrained 4DVAR with DHM	757

S. J. Noh (✉)

Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX, USA
e-mail: seongjin.noh@gmail.com

A. H. Weerts

Operational Water Management, Inland Water Systems, Deltares, Delft, The Netherlands

Hydrology and Quantitative Water Management Group, Wageningen University and Research,
Wageningen, The Netherlands

e-mail: albrecht.weerts@deltares.nl

O. Rakovec

Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

Faculty of Environmental Sciences, Czech University of Life Sciences, Prague, Czech Republic
e-mail: oldrich.rakovec@ufz.de

H. Lee

National Oceanic and Atmospheric Administration, Silver Spring, MD, USA
e-mail: haksu.lee@noaa.gov

D.-J. Seo

Department of Civil Engineering, The University of Texas at Arlington, Arlington, TX, USA
e-mail: djseo@uta.edu

3.5	Lagged PF with DHM	759
3.6	Multiscale Bias Correction	763
3.7	Comparison of DA Methods	766
4	Benefits and Challenges	774
4.1	Large-Scale Streamflow DA	774
4.2	Multi-Data Assimilation	775
4.3	Timing Errors	776
	References	777

Abstract

Streamflow is arguably the most important predictor in operational hydrologic forecasting and water resources management. Assimilation of streamflow observations into hydrologic models has received growing attention in recent decades as a cost-effective means to improve prediction accuracy. Whereas the methods used for streamflow data assimilation (DA) originated and were popularized in atmospheric and ocean sciences, the nature of streamflow DA is significantly different from that of atmospheric or oceanic DA. Compared to the atmospheric processes modeled in weather forecasting, the hydrologic processes for surface and groundwater flow operate over a much wider range of time scales. Also, most hydrologic systems are severely under-observed. The purpose of this chapter is to provide a review on streamflow measurements and associated uncertainty and to share the latest advances, experiences gained, and science issues and challenges in streamflow DA. Toward this end, we discuss the following aspects of streamflow observations and assimilation methods: (1) measurement methods and uncertainty of streamflow observations, (2) streamflow assimilation applications, and (3) benefits and challenges streamflow DA with regard to large-scale DA, multi-data assimilation, and dealing with timing errors.

Keywords

Streamflow · Observation · Data assimilation · Hydrologic modeling · Ensemble Kalman filtering · Particle filtering · Variational assimilation · Multiscale bias correction · Maximum likelihood ensemble filtering

1 Introduction

Streamflow, defined as the flow rate or discharge of water along a river channel (Maidment 1993), is arguably the most important predictor in operational hydrologic forecasting. Streamflow is also the most commonly used, and often the only routinely observed, prognostic variable in hydrologic modeling (Sun et al. 2016). A component of hydrologic cycle, streamflow is generated by a combination of rainfall-runoff and subsurface processes that produce baseflow, interflow, and overland flow and cascades through networks of channels and rivers. Streamflow observations hence reflect various states, elements, and processes that exist and occur in a catchment such as precipitation, temperature, soil, vegetation, land use, surface runoff, groundwater, river morphology, and man-made structures and operations.

Streamflow may be observed using in situ, remote sensing, and alternative methods. To date, the most common method which remains is the use of rating curve with which one may estimate discharge from in situ measurement of stage (McMillan et al. 2012). For instance, in the continental United States (CONUS), there are approximately 7000 in situ stream gauges that report stage measurements in real time which are then converted to streamflow via rating curves. Remote sensing has been expanding the spatial coverage of streamflow observation over much of the globe to include rivers at multiple scales in remote areas (Biancamaria et al. 2017; Bjerkli et al. 2003, 2005; Lettenmaier et al. 2015; Pan et al. 2016; Tourian et al. 2017). Uncertainty characteristics in streamflow observations are dependent entirely on the technique used to measure or estimate discharge (McMillan et al. 2012).

Streamflow data assimilation (DA) uses streamflow observations to update model states and/or parameters to improve model prediction and to quantify predictive uncertainty. Since the early applications of modern DA techniques such as Kalman filter (KF, Kalman 1960) and linear or linearized hydrolB34ogic models (Kitanidis and Bras 1980; Shiiba and Takasao 1980; Wood and Szöllösi-Nagy 1978), the past 40 years have seen a rapid increase in streamflow DA applications. The DA methods used include KF (Wang and Bai 2008), ensemble Kalman filter (EnKF; particle filter (PF; Moradkhani et al. 2005; Noh et al. 2014; Weerts and El Serafy 2006; Yan and Moradkhani 2016)), variational assimilation (VAR; Ercolani and Castelli 2017; Lee and Seo 2014; Seo et al. 2009), maximum likelihood ensemble filter (MLEF; Rafieeinab et al. 2014; Zupanski 2005), and variants of the aforementioned filters and smoothers (McMillan et al. 2013; Noh et al. 2011). The models used include lumped and semi-lumped and distributed hydrologic and hydraulic models. The applications include real-time forecasting (Habets et al. 2008; Tao et al. 2016) and parameter and model structural uncertainty identification (Pathiraja et al. 2016; Vrugt et al. 2013).

Whereas the methods used for streamflow data assimilation originated and were popularized in atmospheric and ocean sciences, the nature of streamflow DA is significantly different from that of atmospheric or oceanic DA (Liu et al. 2012; Seo et al. 2014). Compared to the atmospheric processes modeled in weather forecasting, the hydrologic processes for surface and groundwater flow operate over a much wider range of time scales. In addition to differences among multiple hydrologic components, the memory issue also occurs within the same state variable due to spatial transfer and aggregation from the headwater to downstream locations. The effect of different hydrologic memories or timing errors could be more significant especially concerning large catchments and distributed hydrologic models. Not surprisingly, most hydrologic systems are also severely under-observed. The common strategy in streamflow DA is to update the catchment state variables or parameters, which are often not observed, using streamflow observations at specific locations along the channels.

A number of approaches have been suggested to take into account explicitly or implicitly the memory of the hydrologic system in streamflow DA (McMillan et al. 2013; Noh et al. 2014; Rafieeinab et al. 2014; Rakovec et al. 2015). The key idea for the above accounting is to relate the past observations with the current state variables in filtering or to smooth both past and current state variables in smoothing and variational approaches using observations within an assimilation window. In streamflow DA with distributed hydrologic model (DHM), high dimensionality

is also a challenge. Updating model states at all grids in a catchment in multiple layers using a very small number of streamflow observations may pose a severely underdetermined inverse problem. There have been a number of attempts to specify uncertainty of a high-dimensional hydrologic model in a scrutinized way or to decrease the number of model states while maintaining the forecast quality (McMillan et al. 2013; Rakovec et al. 2015; Xie and Zhang 2013).

The purpose of this chapter is to provide a comprehensive review on streamflow measurements and associated uncertainty and to share the latest advances, experiences gained, and science issues and challenges in streamflow DA. Toward this end, the chapter is organized as follows: Sect. 2 provides a brief review of streamflow measurement including in situ stage-discharge methods based on rating curve, remote sensing methods, and direct and crowdsourced methods. Then, Sect. 3 presents illustrative real-world applications to show capabilities and various aspects of streamflow DA. Lastly, Sect. 4 discusses benefits and challenges of streamflow DA especially focused on large-scale, multi-data, and timing errors.

2 Streamflow Measurement and Associated Uncertainties

Streamflow DA uses real-time streamflow measurements to produce the best possible estimates of initial hydrological conditions or parameters at the start of the model prediction explicitly taking into account input uncertainty, model structural uncertainty, and observation uncertainty. Therefore, uncertainty in streamflow observations significantly affects the distribution of states updated via DA. In this section, we review the commonly used discharge measurement methods and typical sources of uncertainties in streamflow observation.

2.1 In Situ Stage-Discharge Measurement

The most commonly used method for continuous observation of streamflow at a fixed location is via stage measurement whose observations are converted to discharge using a rating curve. The rating curve is a hydraulic model of the stage-discharge relation at the gauging location and is estimated by fitting a curve through a set of gauging points (McMillan and Westerberg 2015). A widely used rating curve model is the power law function:

$$q = a(h + c)^b \quad (1)$$

where h and q are the gauged stage and the estimated discharge, respectively. The coefficients a , b , and c are the model parameters. Equation (1) is similar to the weir equations where the parameters a and b are related to the shape of the weir and c is related to the stage of zero flow (Coxon et al. 2015). If a single rating curve cannot represent the observations, multiple sets of parameters may be assigned to different flow regimes.

Uncertainties in the rating curve arise from both aleatory (random) and epistemic (of unknown character, nonrandom) sources (Beven 2016; Beven et al. 2011; Coxon et al. 2015). Aleatory uncertainties include point measurement errors in both stage and discharge (McMillan and Westerberg 2015). Uncertainty in the stage measurement is related to instrument precision and is generally considered small, rarely exceeding 10 mm. The discharge measurement is prone to errors associated with both instrumentation and quality control (McMillan et al. 2012) and hence is more uncertain. Epistemic uncertainties include changes in morphology and channel cross section, seasonal variations of the state of vegetation and ice, backwater and hysteresis effects, and human intervention and regulation. According to a benchmarking study by McMillan et al. (2012), the total streamflow observation errors associated with the in situ stage and discharge measurement method is 50–100% for low flows, 10–20% for medium or high (in-bank) flows, and 40% for out-of-bank flows. Coxon et al. (2015) also reported that the discharge errors range from 25% for low flows to 13% for the highest normalized flows in the analysis on UK gauging stations.

Uncertainties in stage and discharge measurements transfer to parametric and structural uncertainties in the rating curve. For example, if multiple segments are specified in a rating curve, the number of segments and the power law parameters, a , b , and c in Eq. (1) at each segment and a range of each segment are parameters for estimation. The fitting procedure of the compound rating curve is a piecewise regression problem where the number of segments and the associated change points are assumed unknown. A number of different methods may be used to estimate the rating curve parameters and the uncertainty associated with the resulting streamflow estimates (Juston et al. 2014; Kuczera 1996; Le Coz et al. 2014; Petersen-Øverleir and Reitan 2005; Westerberg et al. 2011). To date, discharge estimates based on the rating curve have been the most widely used observations in streamflow DA.

2.2 Remote Sensing Measurement

Remote sensing techniques such as satellite and radar altimetry have the ability to monitor variations in surface water stage for large wetlands, rivers, and associated floodplains where direct gauging is not available (Birkett et al. 2002). The fundamental challenges in using air-/spaceborne observations for discharge estimation are (a) obtaining river stage measurements with sufficient accuracy and (b) obtaining the cross-sectional velocity measurements needed to establish rating curves (Lettenmaier et al. 2015). Despite great potential for global applications, remote sensing-based streamflow observation is considered highly uncertain and is not widely used in DA. A number of studies have attempted to assimilate directly stage measurements from spaceborne synthetic aperture radar (SAR) to update hydrodynamic or hydraulic model states (García-Pintado et al. 2013; Giustarini et al. 2011; Matgen et al. 2010). Significant advances in streamflow DA may occur in the near future owing to the increasing accuracy and frequency of remotely sensed observations on a global scale. For instance, the joint US-French Surface

Water and Ocean Topography (SWOT) mission scheduled for launch in 2020 aims to produce images of surface water elevation rather than one-dimensional track returns. With a vertical accuracy of about 1 cm, SWOT is intended to produce estimates of discharge for rivers with width of at least 100 m. Multiple synthetic experiments have been conducted to assess potential gains from assimilating streamflow estimates from SWOT (Andreadis et al. 2007; Andreadis and Schumann 2014). In addition to observations of stage and discharge in the river channels, low-cost spaceborne observations of flood extent is expected to be assimilated into hydrologic and hydrodynamic models (Yan et al. 2015).

2.3 Direct and Crowdsourced Measurement

New instruments and techniques continue to be developed to measure streamflow more directly using riverbed-mounted sensors such as acoustic Doppler velocimetry (ADV) and acoustic Doppler current profiling (ADCP) (McMillan et al. 2012). These techniques benefit from direct streamflow measurement without using rating curve, especially at locations where stage-discharge relationship may not hold. However, the streamflow observations obtained from these techniques have their own individual uncertainty characteristics which should be properly taken into account for use in DA. McMillan et al. (2012) reported that the errors in streamflow data ranged 20% for the ADV technique. These techniques for direct streamflow measurement are also utilized to complement the rating curve method.

In recent years, crowdsourcing of flood observation based on cellular communication and digital imaging technologies has emerged with a main focus on rapid near real-time mapping of the reports of flood damages and emergencies (Fohringer et al. 2015; Koswatte et al. 2015; Le Coz et al. 2016; Lowry and Fienen 2013). Hydrologic and hydraulic information such as the extent and depths of inundated areas and flow rate can be estimated from messages, images, and videos produced and shared by citizens. Le Coz et al. (2016) showed that flood videos recorded by citizen could be processed to estimate river flow velocity and discharge using image velocimetry techniques such as large-scale particle image velocimetry (LSPIV). Mazzoleni et al. (2017) present a framework for assimilating crowdsourced data with irregular availability and variable accuracy into hydrologic models. Despite quality and credibility issues, crowdsourced data are expected to extend spatiotemporal coverage of flood observations and may impact the DA paradigm used in hydrologic forecasting in the near future.

2.4 Observational Uncertainty in Streamflow DA

Observational uncertainty in streamflow data is usually modeled as having a Gaussian distribution $N(0, \sigma_{\text{obs}_k}^2)$ (e.g., Georgakakos 1986; Salamon and Feyen 2010). The Gaussian observation error may be modeled as heteroscedastic or

constant. If heteroscedastic, the standard deviation σ_{obs_k} is assumed to increase in proportion to the magnitude of the observed streamflow:

$$\sigma_{\text{obs}_k} = \alpha_{\text{obs}} y_k + \beta_{\text{obs}} \quad (2)$$

where α_{obs_k} is the coefficient which relates the heteroscedastic error to observation at the current time k , y_k , and β_{obs} is the coefficient which represents constant uncertainty especially in periods of low flow. These coefficients are to be determined according to the statistical characteristics of observed streamflow.

Although characteristics of observational errors is closely related with measurement methods and rating curve uncertainty, observation uncertainty is usually expressed in a simplified way in streamflow DA shown above. However, recent advances in the definition of rating curve uncertainty (e.g., aleatory and epistemic uncertainties) allow incorporating a flow measurement error explicitly in DA. In one example, Ocio et al. (2017) presented a probabilistic framework to consider the flow measurement error originated from the rating curve as well as input forcing and model structural errors using multiple DA approaches.

3 Applications of Streamflow Assimilation

In this section, we describe streamflow DA applications involving both lumped and distributed hydrologic models. The latter cases also illustrate how hydrologic memory and high dimensionality may be handled in different DA approaches. Then, we explore multiscale bias correction for streamflow DA in order to relax assumptions behind the existing DA methods. Finally, two comparative evaluation studies are then reviewed to discuss strength and weakness of different DA methodologies. Among numerous applications, we selected some example cases to glimpse various aspects of streamflow DA especially with regard to updating of model states. It is worthy to note that parameter or dual parameter-state updating approaches have been drawing attention in research community. However, online parameter estimation in theory is based upon a premise that parameters are unknown but static (Kantas et al. 2015). Mathematical theory and derivations are presented minimally in this section to focus on application and contextual understanding of different streamflow DA methods. Fundamental theory and mathematical derivations of basic DA methods may be found in other chapters in this volume.

3.1 Input Uncertainty Modeling

DA combines information from a model and observations in an objective way by taking into account associated uncertainty. Primarily due to the large degrees of freedom associated with the DA setup (given limited observations and computing resources), it is impossible to consider all different uncertainties from input, parameters, model structures, and observations. Effective representation of uncertainty is,

therefore, a crucial issue in DA. In particular, rainfall (including temperature in snow-dominated basins) is the dominant input term determining the hydrological response. For instance, the study of Arnaud et al. (2011) on 500 basins in the southeast of France found that small catchments were more sensitive to the catchment rainfall input uncertainties with regard to sampling of rainfall data in a gauge network, while the largest catchments were more sensitive to uncertainties generated when the spatial variability was not taken into account. As such, sound spatial and temporal representation of uncertainty in rainfall is also crucial in streamflow DA. A common practice to consider uncertainty of input forcings is to perturb the point or spatially distributed estimates using Gaussian white noise. From a hydrometeorological perspective, however, the realizations based on the random perturbation are not very realistic because of a lack in coherent temporal evolution of each individual rainfall field. A method to obtain sound spatially distributed rainfall fields including a proper error structure is conditional simulation (e.g., Goovaerts 1997). The rainfall ensemble generator presented here is based on multivariate conditional simulation (Rakovec et al. 2012), which maintains spatial covariance structure given the rain gauge observations and also ensures that the temporal correlation structure is maintained for each ensemble member. Multivariate conditional simulation consists of the following steps: (1) draw time-independent univariate ensemble realization i at time step k , (2) transform rainfall observations into normal space, (3) define a random path throughout the simulation domain such that each grid node is visited once, (4) generate a random number from a Gaussian distribution at each cell given the kriging mean value and corresponding variance of the rain gauge observations, and (5) back-transform the normally distributed values to the original rainfall distributions after all grid cells are simulated. Figure 1 shows two ensemble realizations in comparison with the Thiessen polygon. The magnitudes of the simulated and interpolated rainfall fields are approximately the same, although the spatial variation is much greater in the simulated fields. Considering uncertainty in deriving spatially distributed forcings this way prevents averaging out of extreme values that are of particular interest for estimating hydrologic extreme values.

In addition to one specific example shown above, a number of studies have presented methods to produce a gridded ensemble of precipitation which allows for the estimation of input forcing uncertainty with regard to streamflow DA. For instance, Clark and Slater (2006) presented a method to produce probabilistic estimates of precipitation based on locally weighted regression, where topographic attributes are used as explanatory variables to estimate spatial variations in occurrence and amounts of precipitation. Clark et al. (2008) used the matrix decomposition method outlined by Clark and Slater (2006) to generate the spatially correlated normally distributed random noises for applying EnKF to update states in a DHM. Recently, Newman et al. (2015) modified and extended the work of Clark and Slater (2006) to develop a daily, station-based, ensemble dataset of precipitation and temperature at one-eighth degree resolution for CONUS, Northern Mexico, and southern Canada. They concluded that the ensemble product produced a more realistic occurrence of precipitation statistics especially in wet day fraction, but future work is required to address temporal correlation of precipitation anomalies.

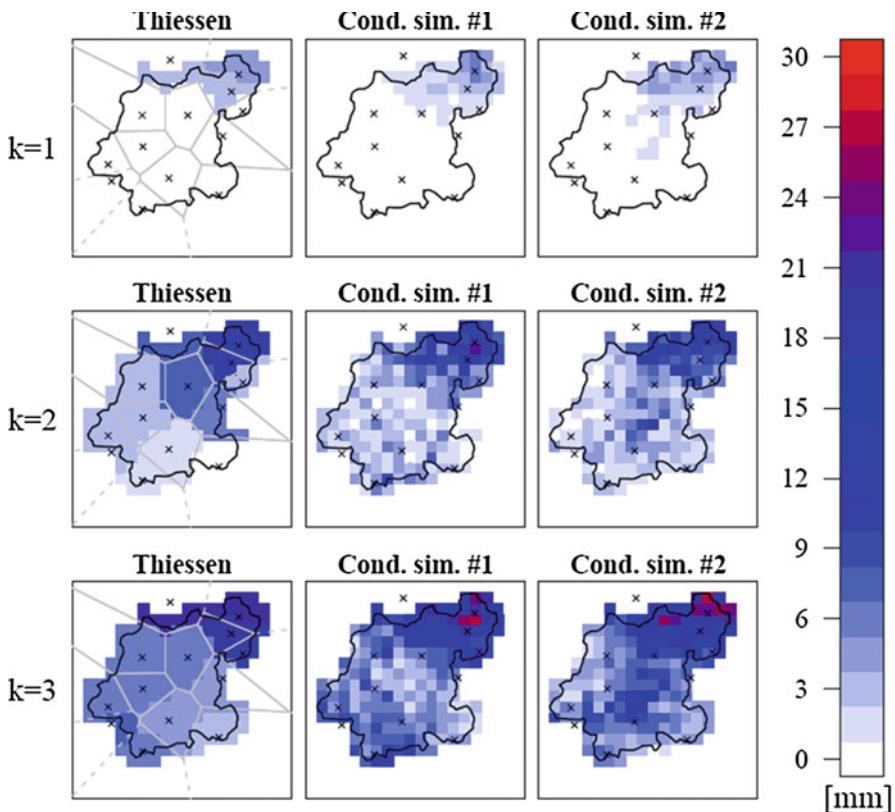


Fig. 1 Example of spatial rainfall fields for three consecutive time steps obtained by the Thiessen interpolation (left) and two realizations of conditional simulation (center and right) for the Katsura catchment on July 17, 2006. (From Noh et al. 2014)

In addition to gauge-based ensemble methods, there is another alternative to use the uncertainty estimation methods developed for remote sensing-based data. In one specific example, Kirstetter et al. (2015) developed a probabilistic approach for estimating precipitation from a radar measurement which explicitly accounts for precipitation types and uses a probabilistic model quantifying relation between radar reflectivity and the corresponding reference precipitation.

Despite a lack of spatial coherence, the global multiplier based on normal or log-normal random noises is one of the widely used methods to perturb input forcings in streamflow DA due to the relative simplicity and ease of implementation (Chen et al. 2013; DeChant and Moradkhani 2012). In addition to ensemble methods based on spatiotemporal correlation structure at large scale, simplified approaches such as the global multiplier may be effective at small scale. However, simplified input uncertainty approaches need to be constrained in a more objective way, which suggests further development and validation of error models on precipitation measured by heterogeneous methods. In one specific example, the

work of McMillan et al. (2011) in the Mahurangi catchment (New Zealand) found that the log-normal distribution provided a relatively close approximation to the true error characteristics of observed rainfall but did not capture the distribution tails, especially during heavy rainfall. In addition, they concluded that a multiplicative error formulation was appropriate for the two measurement types: a rain gauge network and a high-resolution radar rainfall.

3.2 Asynchronous EnKF (AEnKF) with DHM

Besides the traditional classification of DA methods into sequential and variational, they may also be grouped into synchronous and asynchronous. The synchronous methods include three-dimensional (3D) VAR, EnKF, and PF which assimilate observations valid at the prediction time only. The asynchronous approaches include four-dimensional (4D) VAR, ensemble Kalman smoother (EnKS), and asynchronous EnKF (AEnKF) in which observations are assimilated into the model valid at times other than the prediction time as well.

The EnKF estimates the true pdf of model states based on available observations in probabilistic manner. At a given time step t , the i th ensemble member is approximated as follows:

$$\mathbf{x}_t^{+,i} = \mathbf{x}_t^{-,i} + \mathbf{K}_k \left(\mathbf{y}_k - \mathbf{h}(\mathbf{x}_t^{-,i}) + \boldsymbol{\nu}_k \right)$$

where $\mathbf{x}_t^{+,i}$ is the analysis (update) model state vector and $\mathbf{x}_t^{-,i}$ is the corresponding forecast (prior) model state vector. \mathbf{K}_k is the Kalman gain, which weights the errors in model and observations, and \mathbf{y}_k is the vector of model observations. $\mathbf{h}(\cdot)$ is the model operator, which translates model states into the model output, and $\boldsymbol{\nu}_k$ represents the observation uncertainty through the Gaussian white noise. While the synchronous EnKF equation assimilates observations only at the current time step t , the AEnKF employs past observations in order to update the model states valid at the current time step (Sakov et al. 2010). The AEnKF model state vector $\tilde{\mathbf{x}}_t^{-,i}$ is augmented with the past simulations from the W previous time steps:

$$\tilde{\mathbf{x}}_t^{-,i} = \begin{bmatrix} \mathbf{x}_t^{-,i} \\ h(x_{t-1}^{-,i}) \\ h(x_{t-2}^{-,i}) \\ \vdots \\ h(x_{t-W}^{-,i}) \end{bmatrix} \quad (3)$$

Similarly, the other terms get augmented accordingly; see Rakovec et al. (2015) for detailed definitions.

The characteristic of the AEnKF of adding past observations to improve the DA procedure at the current time step is attractive for operational use given its relatively

small additional computational costs over EnKF. Rakovec et al. (2015) investigated performance of AEnKF for assimilating streamflow observations into a gridded hydrologic model for the Upper Ourthe catchment in the Belgian Ardennes. Figure 2 shows the mean difference between the forecast and updated model states for four scenarios with different assimilation windows and number of observations. These examples scrutinize the model behavior of the updated internal states from two perspectives: (1) the effect of assimilating past asynchronous observations in addition to those valid at the current time and (2) the impact of spatially distributed

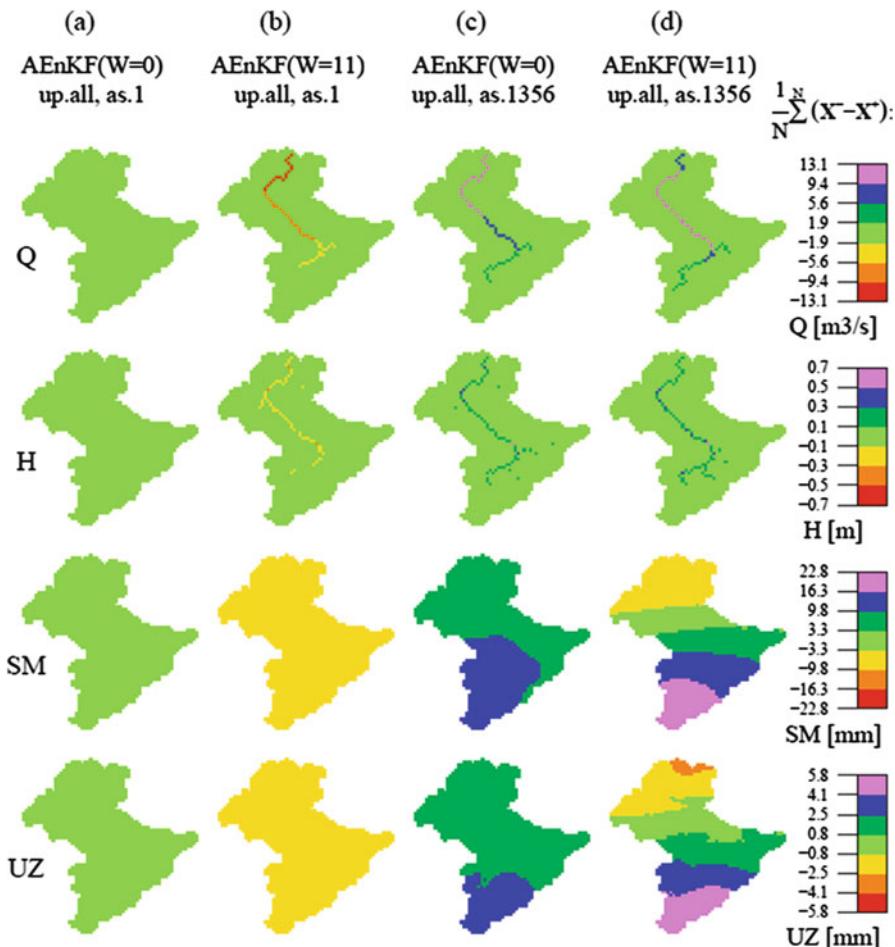


Fig. 2 Mean difference between the forecast (X^-) and updated (X^+) model states (shown in horizontal panels) on December 31, 2002, for four scenarios (shown in vertical panels). Four most sensitive model states are shown: discharge (Q), water level (H), soil moisture (SM), and upper zone (UZ). Notations $W = 0$ and $W = 11$ represent the size of the augmented state variable. Notation up all indicates that all of the model states are updated. Notation as “xx” indicates the gauge numbers which are assimilated. (From Rakovec et al. 2015)

observations on the updated states. The innovation of the model states is spatially differentiated most widely when asynchronous observations from multiple locations are considered (Fig. 2d) compared to the EnKF with the synchronous observations from multiple locations (see Fig. 2c) or compared to the AEnKF which assimilates a single discharge observation (see Fig. 2b). Thus, the modeler can identify exact locations within the model domain, where water is removed from the catchment during the update. Additionally, abrupt changes between the forecast and updated states can possibly point out to streamflow observation biases.

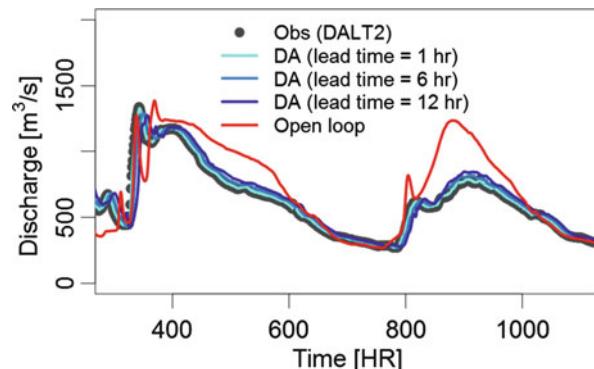
3.3 VAR with Hydrologic Routing Model

For streamflow DA, VAR solves a smoothing, rather than filtering, problem. That is, VAR solves for the states valid within the entire assimilation window rather than those valid at the prediction time only when minimizing the objective function. Accordingly, VAR naturally accounts for the process memory within the scale of the assimilation window. Methods to get the solution of the objective function include the gradient method and the adjoint method. The gradient method is based on the fact that the gradient is equal to 0 at the minimum of the objective function (Blayo et al. 2014). Since an analytical expression for the gradient cannot be obtained for most of nonlinear hydrologic models, the gradient must be determined numerically using the finite difference approximation. However, the gradient method is not practical for high-dimensional systems due to expensive computation. On the other hand, the adjoint method provides a mathematically efficient algorithmic framework computing the gradient of a differential scalar function (i.e., objective function) with respect to its arguments (i.e., initial conditions and/or parameters) (Blayo et al. 2014). Adjoint code can be generated using automatic differentiation engines such as TAPENADE (<http://tapenade.inria.fr:8080/tapenade/index.jsp>).

The following provides an example of assimilating streamflow observations with a distributed hydrologic routing model using VAR. The three-parameter Muskingum model (O'Donnell 1985) was used for the Upper Trinity River in Texas, USA (Noh et al. 2016). In this example, streamflow observations at the two upstream locations (FWOT2, CART2) were used as input, while observations at the two downstream locations (GPRT2, DALT2) were assimilated into the routing model (Noh et al. 2016). The objective function included uncertainties in the upstream BCs, routing model parameters, lateral inflow, and the ICs. The resulting constrained minimization problem was solved numerically using the Fletcher-Reeves-Polak-Ribiere minimization (FRPRMN) algorithm (Press et al. 1992), a conjugate gradient method. Figure 3 shows the observed and simulated streamflow with and without VAR at downtown Dallas (DALT2), a highly important National Weather Service (NWS) forecast point in a large urban center. Compared to the open-loop (i.e., DA-less) solution, the DA-aided simulations show greatly improved accuracy (Noh et al. 2016).

Due to potentially significant errors from model structure and parameters, predictions by hydrologic routing models can benefit from DA if streamflow

Fig. 3 Comparison of observed and simulated streamflow with (in blue) and without (in red) DA versus observations (black solid circles)



observations are assimilated in real time. Note that VAR has benefits over filtering methods in routing DA problems from its smoothing nature within the assimilation window and the efficient estimation of the objective function using the adjoint method.

3.4 Weak- Versus Strong-Constrained 4DVAR with DHM

From the perspective of model uncertainty, VAR may be classified into weak- and strong-constrained approaches. The weakly constrained VAR explicitly introduces model error terms in model equations and/or an objective function, while the strong-constrained VAR assumes that the prediction model is exactly satisfied by the sequence of estimated state variables without considering random errors (Blayo et al. 2014). Lee et al. (2016) examined the capability of weakly constrained (WC) DA approaches for the distributed SAC-SMA model. In the WC VAR approach, an error (\mathbf{X}_W) was added to run off components prior to the routing to model the inadequacy of a rainfall-runoff model and penalize in the objective function. WC DA produces state variables that are dynamically more consistent with a priori states than the strongly constrained (SC) DA by adjusting control variables to a smaller degree while retaining the quality of analyses. Here, we describe error modeling used in WC 4DVAR for the DHM which is comparatively evaluated against SC 4DVAR.

The study area used is a headwater basin with an area of 2258 km² that drains into the Elk River near Tiff City, MO (TIFM7). Two types of spatial structure of \mathbf{X}_W were examined, spatially heterogeneous or homogeneous. While heterogeneous \mathbf{X}_W provides a flexibility to model errors in runoff simulation at individual pixels, it increases dimensionality of the control variables very significantly and hence increases the under-determinedness of the DA problem. Figure 4 shows the mean absolute difference (MAD) of base and updated soil moisture at a pixel scale. Spatially homogeneous \mathbf{X}_W generally adjusts soil moisture less than heterogeneous \mathbf{X}_W in terms of MAD, indicating potentially spatially correlated model errors and/or benefits of solving less under-determined DA problem. It should be noted that,

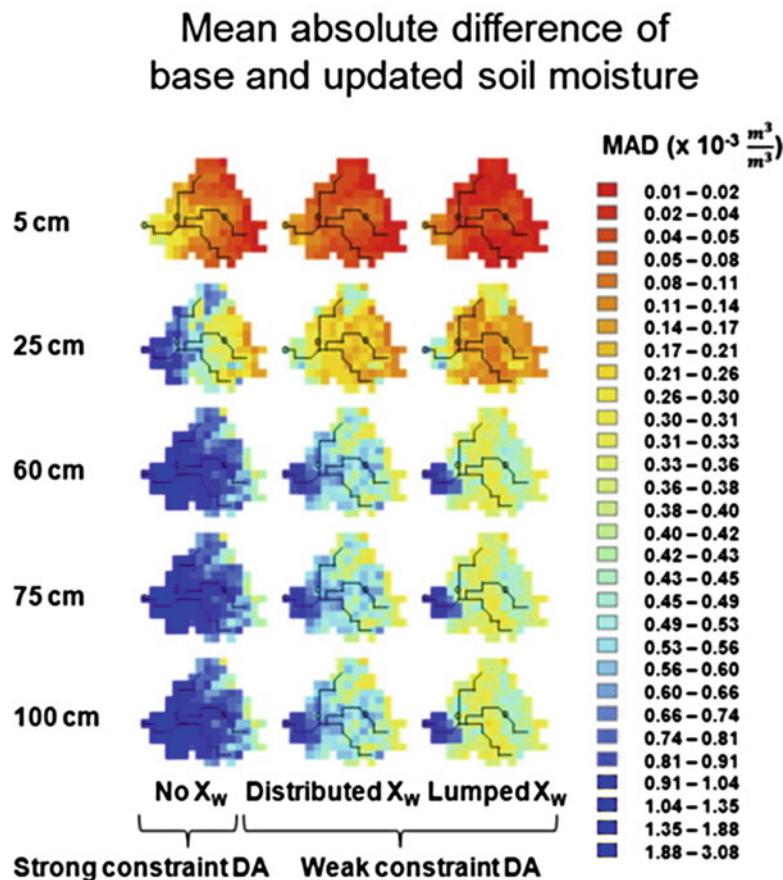


Fig. 4 Mean absolute difference (MAD) of a priori and updated soil moisture translated from the SAC-SMA model states. (From Lee et al. 2016)

while small in magnitude, homogeneous adjustment has a large impact on catchment-wide water balance. The effects of heterogeneous adjustment, on the other hand, tend to cancel out due to spatial averaging of random fluctuations over the catchment area.

Figure 5 shows the maps of temporal mean of two runoff components \mathbf{X}_W^{SURF} (surface runoff error) and \mathbf{X}_W^{GRND} (groundwater runoff error) in two \mathbf{X}_W modeling cases. Spatially homogeneous \mathbf{X}_W produced larger mean \mathbf{X}_W^{SURF} and \mathbf{X}_W^{GRND} than heterogeneous \mathbf{X}_W by approximately 2 and 19 times, respectively. In heterogeneous \mathbf{X}_W , both \mathbf{X}_W^{SURF} and \mathbf{X}_W^{GRND} for upstream pixels may be smaller because the information contents in outlet flow observations are too diluted through the routing process to inform the runoff processes at upstream locations. Also, the increased dimensionality of the state space in heterogeneous \mathbf{X}_W makes the DA problem ill-posed compared to homogeneous \mathbf{X}_W which may render the assimilation

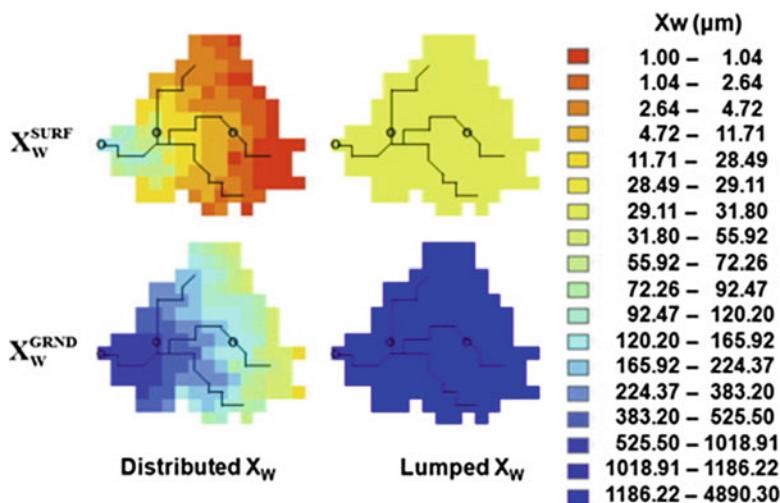


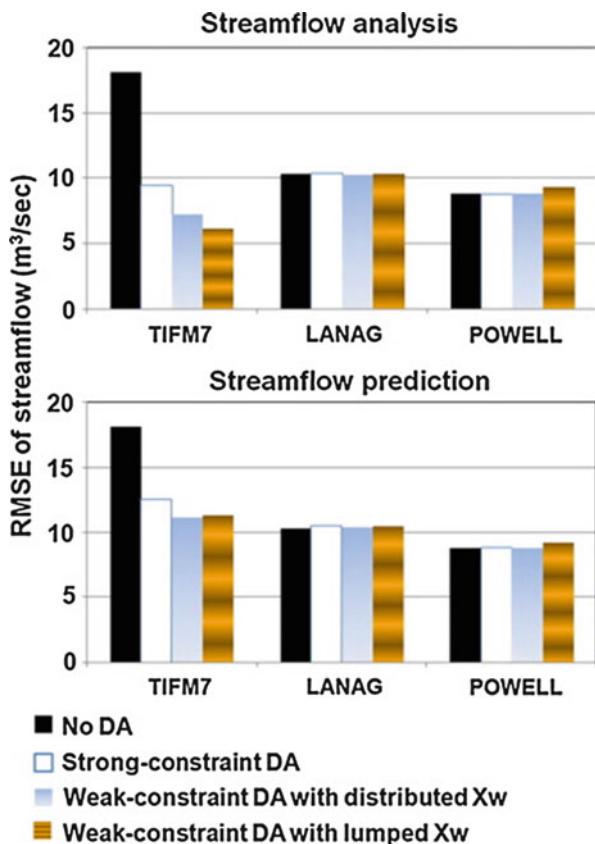
Fig. 5 Map of temporal mean $\mathbf{X}_W^{\text{SURF}}$ and $\mathbf{X}_W^{\text{GRND}}$ generated via WC DA applied to the distributed SAC-SMA. (From Lee et al. 2016)

procedure passive in adjusting $\mathbf{X}_W^{\text{SURF}}$ and $\mathbf{X}_W^{\text{GRND}}$. Figure 6 shows that the WC DA approach reduces RMSE of outlet flow analysis and prediction with smaller adjustment to soil moisture (Fig. 5) than in the SC DA approach.

3.5 Lagged PF with DHM

The major drawback of the DA approaches discussed in the previous sections (KF-based and VAR methods) is the underlying assumption that the model states and the observations have a Gaussian distribution (Lahoz and Schneider 2014). PF is a Bayesian learning process in which the propagation of all uncertainties is carried out by a suitable selection of randomly generated particles without any assumptions about the nature of the distributions (Gordon et al. 1993; Ristic et al. 2004). Due to this flexibility, PF has drawn an attention in hydrologic applications where the assumption on a Gaussian distribution is not always valid and models are highly nonlinear (Noh et al. 2014; Vrugt et al. 2013; Weerts and El Serafy 2006; Yan and Moradkhani 2016). Unlike KF-based methods, PF stochastically selects particles with respect to the associated weights instead of directly correcting state variables, which may alleviate numerical instability or a loss of mass balance especially in process-based models. On the other hand, PF is computationally expensive to represent non-Gaussian (e.g., multimodal, skewed, or fat-tailed) distributions using a large number of particles. In addition, especially in high-dimensional systems, PF may suffer from filter degeneracy which means a loss of sample diversity in particles. However, with rapid development of computing technology, ensemble

Fig. 6 RMSE of flow analysis (top) and prediction averaged over 1–6 h lead time (bottom) at the outlet and interior locations. (From Lee et al. 2016)



simulation can be implemented in parallel not to increase run-time linearly as the number of ensemble increases if high-performance computing (HPC) systems such as supercomputers are available. Additionally, if major uncertainty sources are represented by properly structured error models such as the multivariate rainfall generator, PF may be applied in a wide range of hydrologic applications without suffering from the curse of dimensionality.

In the meanwhile, whatever distribution is assumed, sequential methods have a similarity from the fact that model states are innovated using the latest measurement information. Therefore, in case that the observations and model states at the same time domain are not closely associated with each other due to long system memory (e.g., DHM), a frequent update in accordance with observation frequency may lead to incorrect adjustment of state variables (EnKF) or selection of incorrect particles (PF). It is also worthy to note that the significance of system memory due to residence time may differ in DA according to the structure of the hydrologic models. If the current state depends only on the previous state (i.e., the modeled process is Markovian), system memory is not an issue. Lagged

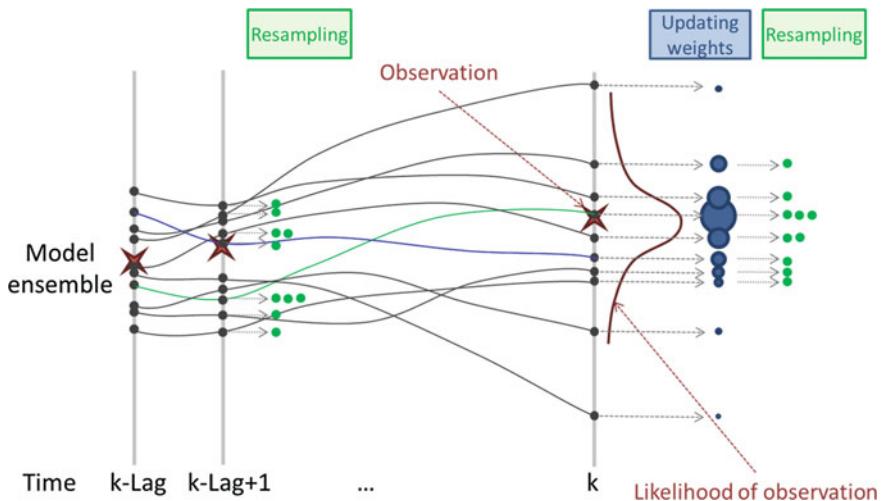


Fig. 7 A cycle of lagged PF in an assimilation window. (From Noh et al. 2014)

filtering (Noh et al. 2011) is an approach to consider different time scales of the hydrologic processes in DA and to use of all available observations within the assimilation window. The key idea of lagged filtering is to update the past state variables (lagged EnKF) or to select the past particles (lagged PF) using the measurement at the current time step.

Figure 7 illustrates a single cycle of lagged PF with ten ensemble members. The blue and green lines show trajectories of two ensemble members. In the resampling step, the trajectory associated with the green line is duplicated into three equally likely members according to the highest likelihood attained at the current time step, while the ensemble member in the blue line remains the same. Since the state variables in the past time (i.e., k -lag+1) may benefit from the latest observation, prediction in the next assimilation cycle could resume from improved initial conditions (IC). Note that as mentioned previously, the concept of updating the past state variables could be applied for any sequential DA methods such as EnKF (McMillan et al. 2013; Noh et al. 2014).

Figure 8 shows sensitivity of lagged PF to the size of the assimilation window which was assessed using the Nash-Sutcliffe efficiency (NSE) in the Katsura catchment (887 km^2), Japan (Noh et al. 2013). In this study, distributed soil moisture variables were perturbed by a global multiplier and updated using streamflow observed at the outlet. When the lead time was less than 2 h, the performance was similar regardless of the size of the assimilation window. With small assimilation windows ($< 4 \text{ h}$), however, the performance sharply deteriorated as the lead time increased. With large assimilation windows ($> 6 \text{ h}$), on the other hand, gains from updating last up to 20 h demonstrated that lagged filtering improves prediction. It is also observed that there is a threshold size for the assimilation window beyond which prediction does not improve. The optimal size

for the assimilation window appears to be associated with travel time of the flood wave in the river at catchment scales, but more rigorous verification remains as a future research endeavor.

A typical DA procedure for streamflow forecasting is illustrated in Fig. 9 (Noh et al. 2014). In real-time streamflow DA, the observed weather variables force hydrologic models in the analysis step, while the forecasted weather variables are applied with the updated state variable through DA in the forecast step. In the work of Noh et al. (2014), input and model uncertainties were considered by the multivariate rainfall generator described in the previous section and perturbing the model state variables (i.e., the soil moisture content) using multiplicative normal noises,

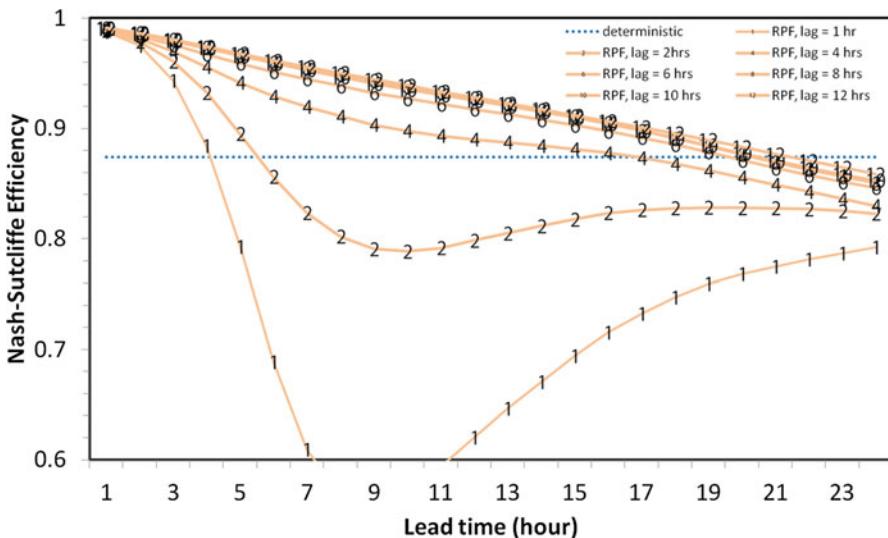


Fig. 8 Nash-Sutcliffe efficiency for varying lag-time windows with lagged PF. (From Noh et al. 2013)

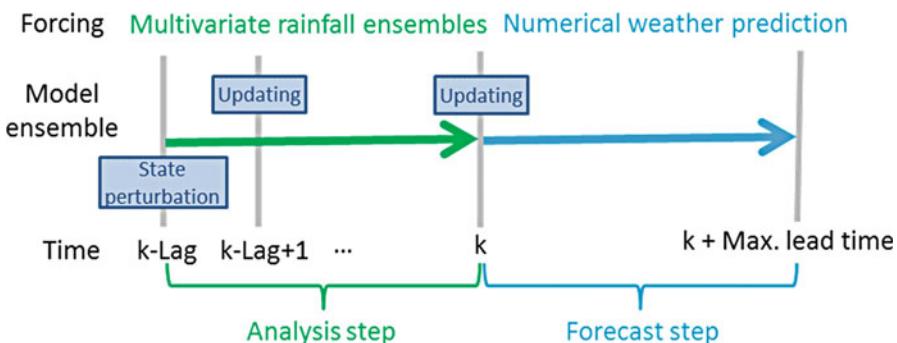


Fig. 9 Schematic diagram of DA procedure using rainfall ensemble and lagged particle filtering. (From Noh et al. 2014)

respectively. Then, the model response (i.e., streamflow at measurement locations) was analyzed within a lag-time window and updated based on comparisons with corresponding streamflow measurements. The model ensemble members updated in the analysis step are then forced with numerical weather prediction (NWP) for the given lead times. As shown in Fig. 10, streamflow forecasts based on this DA approach produced improved performance over the open loop simulation. The results indicate that observational uncertainty in streamflow DA may be specified via structured and objective approaches such as the rainfall ensemble generator with less subjective assumptions on prediction errors. This example demonstrates that, though often considered computationally too expensive, PF is capable of providing a stable solution even for high-dimensional DHMs with objectively specified input uncertainty and lagged filtering.

3.6 Multiscale Bias Correction

All modern estimation theory-based DA techniques assume that the statistical properties of the errors in the model and in the data are perfectly known. In hydrologic reality, however, the above assumption is rarely met. The errors associated with rainfall-runoff processes are nonlinear and flow- and scale-dependent and hence are very difficult to model accurately, particularly given the general paucity of hydrologic observations. It is well known that inaccurate modeling of heteroscedastic errors has large impact on the performance of DA (see, e.g., Seo et al. 2009; Rafieeinab et al. 2014). With less than accurate error modeling, the DA results would be at best suboptimal and may not even be representative of the model dynamics. Compared to the atmospheric processes modeled in weather forecasting, the hydrologic processes for surface and groundwater flow operate over a much wider range of time scales. If the statistical properties of the errors involved are perfectly known, and if the inverse problem is well-posed, one may expect the existing sequential DA techniques, or batch DA techniques with forward

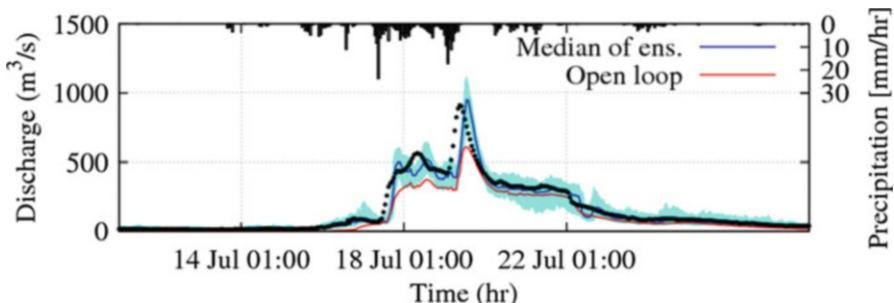


Fig. 10 Observation and 3-h-lead forecasts of streamflow. Blue lines and area represent the median and the 95% confidence interval of the streamflow ensemble by PF, respectively. Red lines represent the open loop simulation. (From Noh et al. 2014)

propagation of the updated initial conditions (IC), to account for time scale-dependent dynamics well. In reality, however, the above conditions are rarely met, and the DA solutions based on inaccurate modeling of scale-dependent error statistics may not be representative of the model dynamics even after a long warm-up period if they are associated with longer-memory processes.

Given the above, one may consider decomposing the inverse (i.e., the DA) problem into two parts of disparate time scales: a first-order bias estimation problem and a higher-order error modeling problem. The former operates at time scales where the random errors may average out, whereas the latter operates at a much smaller time scale. By scale decomposing the estimation problem, it is likely that the resulting inverse problems are of lower dimensionality and hence less likely to be underdetermined. Such an approach has been used successfully in parameter estimation and calibration of lumped hydrologic models (Seo et al. 2009). For assimilating streamflow observations to update the model soil moisture, the above approach would decompose the inverse problem into highly nonlinear soil moisture accounting and mildly nonlinear routing parts for which different DA approaches may be better suited.

The above suggests the following multiscale bias correction (MSBC) as an alternative approach for streamflow DA: (1) identify the time scales at which single-uniform multiplicative biases in observed precipitation, observed potential evaporation (PE), and model-simulated runoff are to be estimated, (2) solve for the biases at the largest temporal scale of aggregation such that the time-integrated simulated streamflow matches the time-integrated observed streamflow, (3) forward-integrate the rainfall-runoff and routing models from the beginning of the largest assimilation window to the midpoint of the window using the bias estimates from Step 2, and (4) repeat Steps 2 and 3 until the biases for the smallest time scale are estimated, and forward-integrate the rainfall-runoff model to the prediction time.

The above MSBC process yields simulated soil moisture states that reflect DA-adjusted biases in the forcing data and, if included, the model-simulated runoff at progressively smaller time scales as they approach the prediction time (i.e., the current time). This strong-constraint formulation does not require the knowledge of the statistical properties of the model errors but only those of the multiplicative biases in observed precipitation and PE. Such an approach may be justified for large assimilation windows because time-varying errors in soil moisture simulations from a well-calibrated rainfall-runoff model should average out over an assimilation window of significant length. The strong-constraint approach has been used successfully in estimation of long-term biases in precipitation and PE (Seo et al. 2009). If the assimilation window is large, one may expect that the impact of adjusting the soil moisture states at the beginning of the window is small compared to that of adjusting biases over the entire window and that the larger-scale bias is a good first guess for the smaller-scale bias. By prescribing the soil moisture states at the beginning of the window by running the model forward using the updated biases over the larger window, the model soil moisture states at the beginning of the smaller window may be expected to be of generally high quality, thereby reducing the need for adjusting, or updating, the initial soil moisture states. Note that, if the biases are

the same at all scales, the smallest-scale solution is the same as the largest-scale solution and that, if the biases are scale-dependent at all scales, each scale-dependent bias contributes to the updated soil moisture conditions over the smallest window ending at the prediction time. The MSBC operation results in soil moisture state that approximately volume-matches the simulated flows with the observed over the range of temporal scales of aggregation. As such, one may consider MSBC as a multiscale double mass analysis in which the analysis period is progressively reduced following bisection.

As described above, MSBC requires the second-order error statistics for precipitation and PE and, very likely, model-simulated runoff (e.g., TCI in SAC). Those for precipitation may be estimated from a combination of the literature (e.g., Smith and Krajewski 1991) and analysis of the NWS-produced radar-only, gauge-only, and radar-gauge mean areal precipitation (MAP) estimates in gauge-dense areas of the study areas (see Seo and Breidenbach 2002). Those for PE may be estimated from a combination of the literature (e.g., McNider et al. 2011) and sensitivity analysis. Those for the multiplicative biases in model runoff may be estimated from mass balance analysis using model-simulated and observed streamflow for the study basins using the highest-quality forcing data and routing models. One may assume very simple temporal correlation structure (e.g., autoregressive-1) and heteroscedasticity model (e.g., $\sigma_i^2 \propto z_i$) and estimate the necessary parameters via sensitivity analysis. Because the error terms in the above formulation impact the objective function only as aggregates over the entire assimilation window, the level of accuracy or sophistication in error modeling does not exert significant influences on the DA solution (Lee et al. 2011).

To glimpse into what may be expected from MSBC, we show below example results for which streamflow, precipitation, and PE data were assimilated into SAC and UH models via VAR (Seo et al. 2009) for the 212 km² catchment, GNVT2, which drains to the Cowleech Fork of the Sabine River at Greenville in North Texas. In the above, the soil moisture states updated at the previous time step were propagated forward to provide the updated IC for the current time step. This effectively renders VAR a sequential DA, which behaves similarly to Kalman smoother but with the added capacity to handle nonlinear observation equations. Figure 11 shows the RMSE versus lead time of streamflow prediction without DA (in black), with fixed-lag smoothing using VAR (in red), and with MSBC (in blue) under the assumption of clairvoyant precipitation and PE. For reference, prediction based on persistence (in green) is also shown. Note that MSBC reduces RMSE noticeably, an impressive result given the optimal nature of the VAR solution. The left panel in Fig. 12 shows the time series of simulated lower zone tension water content (LZTWC), one of Fig. 11 SAC state variables, without DA (in black), with regular VAR (in red) and with MSBC. The right panel in Fig. 12 shows the scatter plots of the two DA-aided simulation of LZTWC versus the DA-less. Note that the MSBC solution tracks closely the DA-less state, whereas the regular VAR solution departs very widely from the DA-less, a strong sign of an under-determinedness. Even though the streamflow results from regular VAR are reasonable for short lead times (see Fig. 11), it is very likely that the regular VAR solution is very poor for

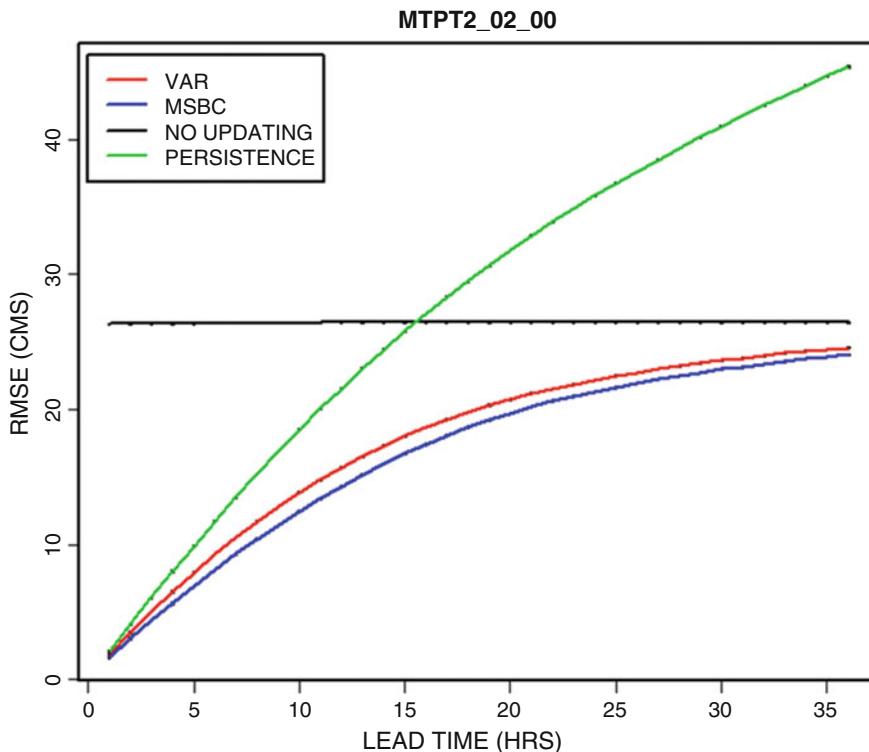


Fig. 11 Effect on MSBC on DA-aided streamflow prediction for MTPT2

longer lead times, which may render the solution unacceptable for, e.g., water supply forecasting. Such multiscale performance is an important criterion for DA to support hydrologic prediction from short to long time scales.

3.7 Comparison of DA Methods

Selection of different DA methods may lead to different performance and uncertainty quantification. Comparative studies of streamflow DA methodology are, therefore, of importance for development of robust algorithms as well as improved performance in operations. In this section, we review two comparative evaluations of DA methods such as EnKF, AEnKF, PF, and MLEF. In the former study, ensemble-based methods, EnKF, AEnKF, and PF, are compared for streamflow DA using the lumped hydrologic models. As discussed in the previous section, the main difference between EnKF and AEnKF is whether or not to include past observations in the updating. Both EnKF and AEnKF are optimal only in the second-order moment sense, while PF provides in theory an optimal solution regardless of the type of the

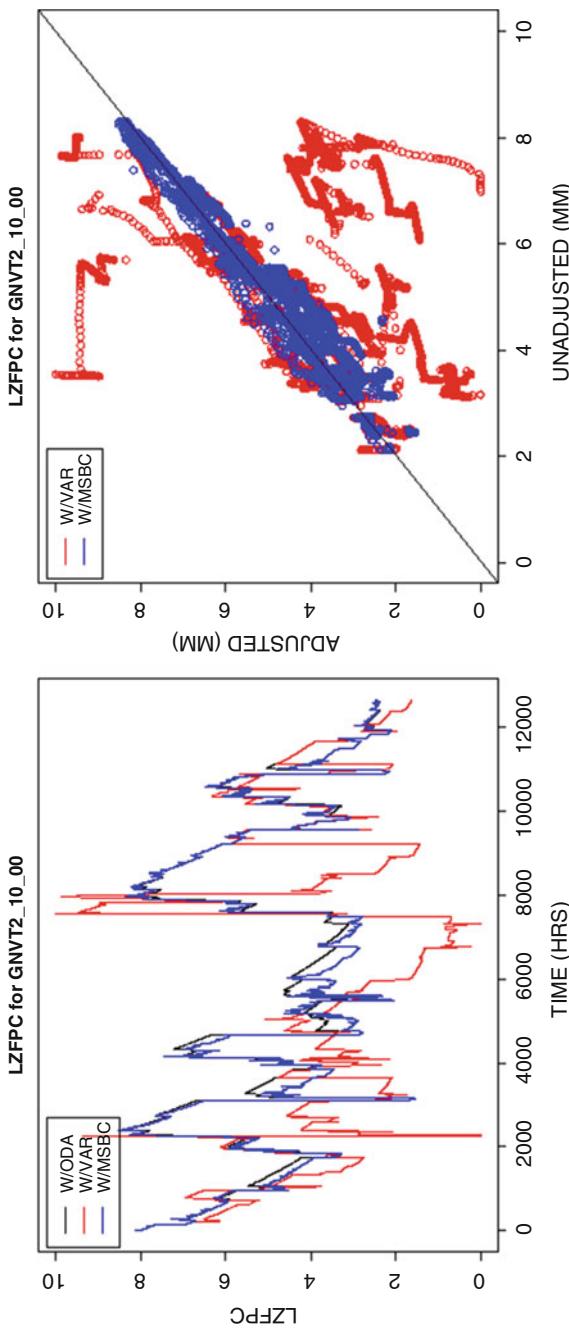


Fig. 12 Time series of DA-aided and DA-less lower-zone primary free water content for GNVT2 (left) and scatter plots of DA-aided state versus DA-less (right)

distribution. In the latter study, the ensemble members in MLEF are generated around the control solution obtained via nonlinear constrained minimization similarly to VAR. As such, for MLEF, the sensitivity to ensemble size is largely a reflection of the sampling uncertainty associated with limited ensemble size. The EnKF results, on the other hand, reflect not only the sampling uncertainty but also the diminished quality of the solution when the distributions involved are not symmetric, as well as the suboptimal nature of the solution when the observations are nonlinearly related to the model states.

3.7.1 Comparison of EnKF, AEnKF, and PF

The experiment described in this section illustrates different behaviors and performances of EnKF, AEnKF, and PF for streamflow DA using the lumped hydrologic models, the Sacramento soil moisture account model (SAC), and unit hydrograph (UH). The study domain includes multiple catchments in the Tennessee Valley, USA. The DA experiments were implemented in the FEWS-OpenDA framework. OpenDA is an open interface standard for (and free implementation of) a set of tools to quickly implement DA and automated calibration for arbitrary numerical models. As shown in Fig. 13, the FEWS operator client (OC) running the forecasting models calls the OpenDA software, which in turn executes DA through (1) running copies of the forecasting system models (termed “fewsasmodel”) in an ensemble (e.g., multiple, parallel simulation) mode and (2) analyzing and returning the results to the OC system.

Comparison of DA methods was done through synthetic experiments, where the true states are considered known. One of advantages of synthetic experiment is that it is possible to check the properties of a DA algorithm for the model being studied. For generating the synthetic data, a “true” flow was run as the model perturbed by one specific realization of the error processes mentioned above. The “true” model is run over the period of October 2008–October 2009, starting from a cold state. The output

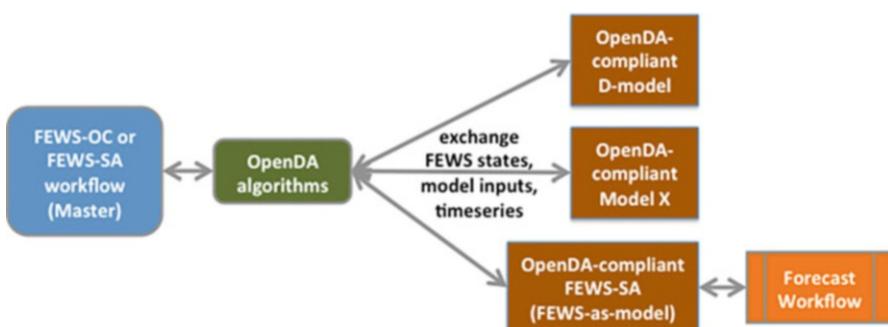


Fig. 13 Schematic design OpenDA- FEWS and other possible OpenDA compliant applications/ models. D-model stands for the Deltares software line, such as D-SOBEK, etc. Model X stands for any OpenDA compliant model, which includes the FEWS system of models

of this “true” model is a six hourly discharge time series. The synthetic data is generated by adding random numbers to this discharge time series. The random numbers are used to represent the observational error, which are assumed to be independent in time and normally distributed. It is further assumed that the standard deviation of the observational error varies with the actual discharge (5% of the actual discharge).

The DA evaluation experiments were performed for the period of Sep 15, 2009–Oct 06, 2009, starting from a cold state. A deterministic model simulation was run over this period as a reference for checking the impact of DA. Time series of the discharge and state variable were plotted to illustrate the results of the experiments. A DA technique will work best when the noise statistics are known. To test the techniques in this idealized situation, the same noise specifications were used in the DA experiments.

The results of the experiment corresponding to the high rainfall event around Sep 26, 2009, are presented in Fig. 14. The following can be observed from these results. Both EnKF and AEnKF successfully bring the model closer to the truth. The impact of DA on the accuracy improvement is significant as compared to the accuracy of the deterministic model. The PF results are however only slightly better than the deterministic run. The PF does not modify the model state based on the discharge residuals but selects/resamples model realizations (particles). Particles that are closer to the truth are more likely to get resampled. However, since the underlying deterministic model is far off from the truth in this case, the estimate is also far from the truth. As noted earlier, the PF approach was found to perform better with relatively larger noise settings than the EnKF approaches, and the same noise settings for all were used here. Overall, the PF does not drift far from the deterministic run. For events in which the deterministic run already performs well, the PF performs better than EnKF and AEnKF. The EnKF and AEnKF actually degrade the accuracy of certain components in this period. The longer-term impact of the accuracy degradation on these components should be tested further.

AEnKF performs similarly with EnKF. Here the analysis with AEnKF is performed with a time interval of 24 h. This means that at each analysis step, all observed discharge data from the last 24 h are assimilated at once. The results of the AEnKF are therefore rather stepwise or less smooth than EnKF.

Figure 15 shows the Talagrand plots: histograms showing the frequencies of observations falling within bins defined by the ensemble traces. For ensembles with equiprobable members, each bin should receive approximately the same number of observations. For all locations a disproportional number of observations fall in the first bin for all the DA-based methods (i.e., below the spread of the ensemble) but with EnKF doing better than AEnKF and AEnKF doing better than PF. Early lead forecasts (i.e., the first two columns) show a modest hump in the middle of the distribution, meaning that the forecasts may also be overdispersive, while longer lead forecasts appear more reliable but biased low. The reference forecasts are overdispersive and biased high (observations tended to verify in the lower ranges of the forecast).

3.7.2 Comparison of EnKF and MLEF

As an ensemble extension of variational assimilation providing the maximum likelihood solution, maximum likelihood ensemble filter (MLEF) accounts for nonlinear relationships between soil moisture and streamflow via a nonlinear observation equation, whereas EnKF solution is optimal in the second-order sense only in the case of a linear observation equation. To assess relative performance in the case of a highly nonlinear observation equation, MLEF and EnKF are comparatively evaluated for the MTPT2 basin which drains into

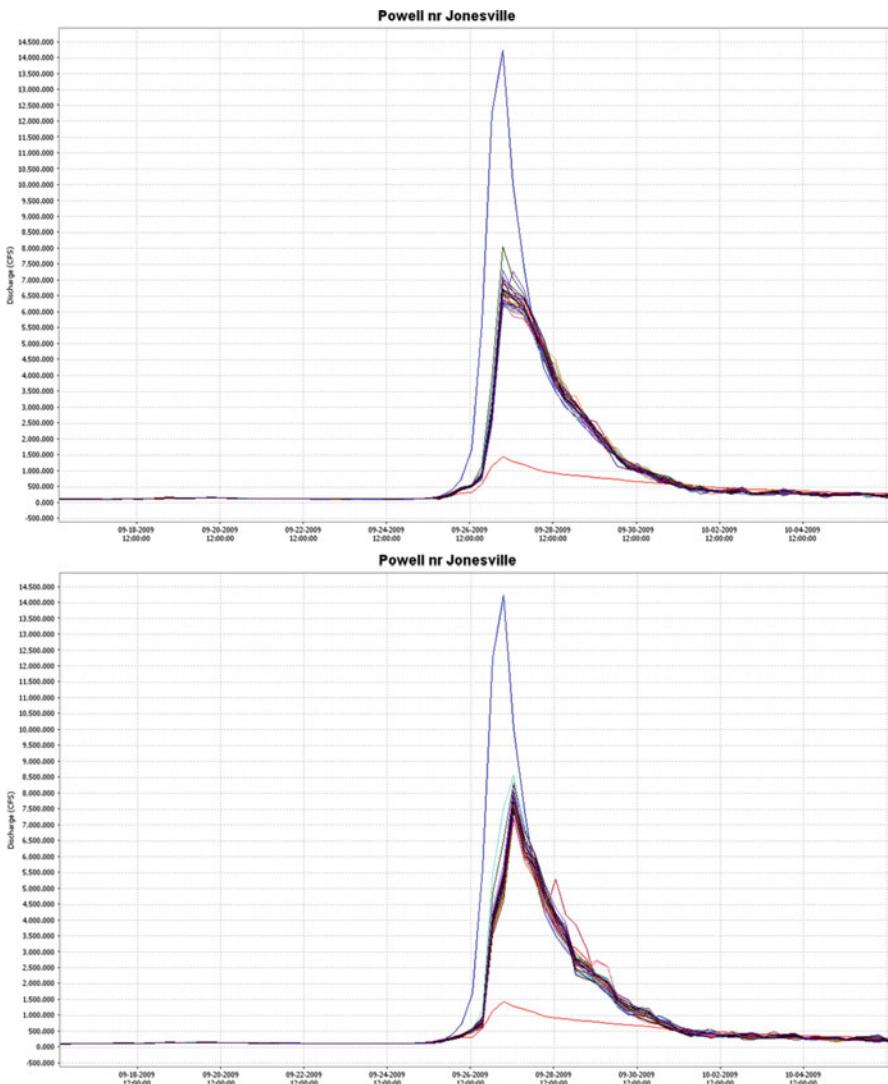


Fig. 14 (continued)

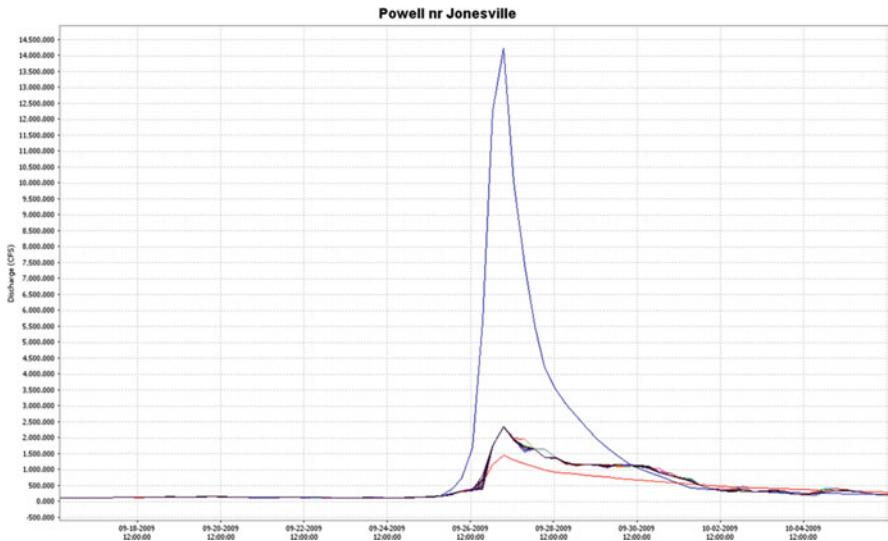


Fig. 14 Validation experiments: discharges by EnKF (top), AEnKF (middle), and PF (bottom). At each panel, blue line is the truth, red line is the deterministic run without data assimilation, and the rest are ensemble run with data assimilation

the Tres Palacios River near Midfield in southern Texas (Rafieeinasa et al. 2014). Streamflow, mean areal precipitation (MAP), and mean areal potential evaporation (MAPE) data were assimilated into SAC and UH in order to update soil moisture states for improved prediction of streamflow at the basin outlet.

Performance of MLEF and EnKF was compared based on the RMSE of streamflow predication as a function of lead time under varying observation errors in streamflow and MAP (Fig. 16a), model errors (Fig. 16b), ensemble size (Fig. 16c), and the number of flow observations assimilated (Fig. 16d). At Fig. 16a, heteroscedastic modeling of observation errors does not improve DA performance over homoscedastic modeling, indicating difficulties of the heteroscedastic modeling in practice. Also, Fig. 18a shows reasonably good performance of MLEF without very accurate modeling of observational error variances. At Fig. 16b, accounting for model errors in soil moisture dynamics noticeably reduces the RMSE of streamflow at short lead hours, and both MLEF and EnKF achieve their respective best with a fraction of 0.025.

At Fig. 16c, EnKF is more sensitive to ensemble size than MLEF and underperforms in comparison to MLEF. Since ensemble members in MLEF are generated around the control solution obtained via nonlinear constrained minimization, the sensitivity of MLEF solutions to ensemble size is largely a reflection of the sampling uncertainty associated with limited ensemble size. On the other hand, EnKF results reflect not only the sampling uncertainty but also the diminished

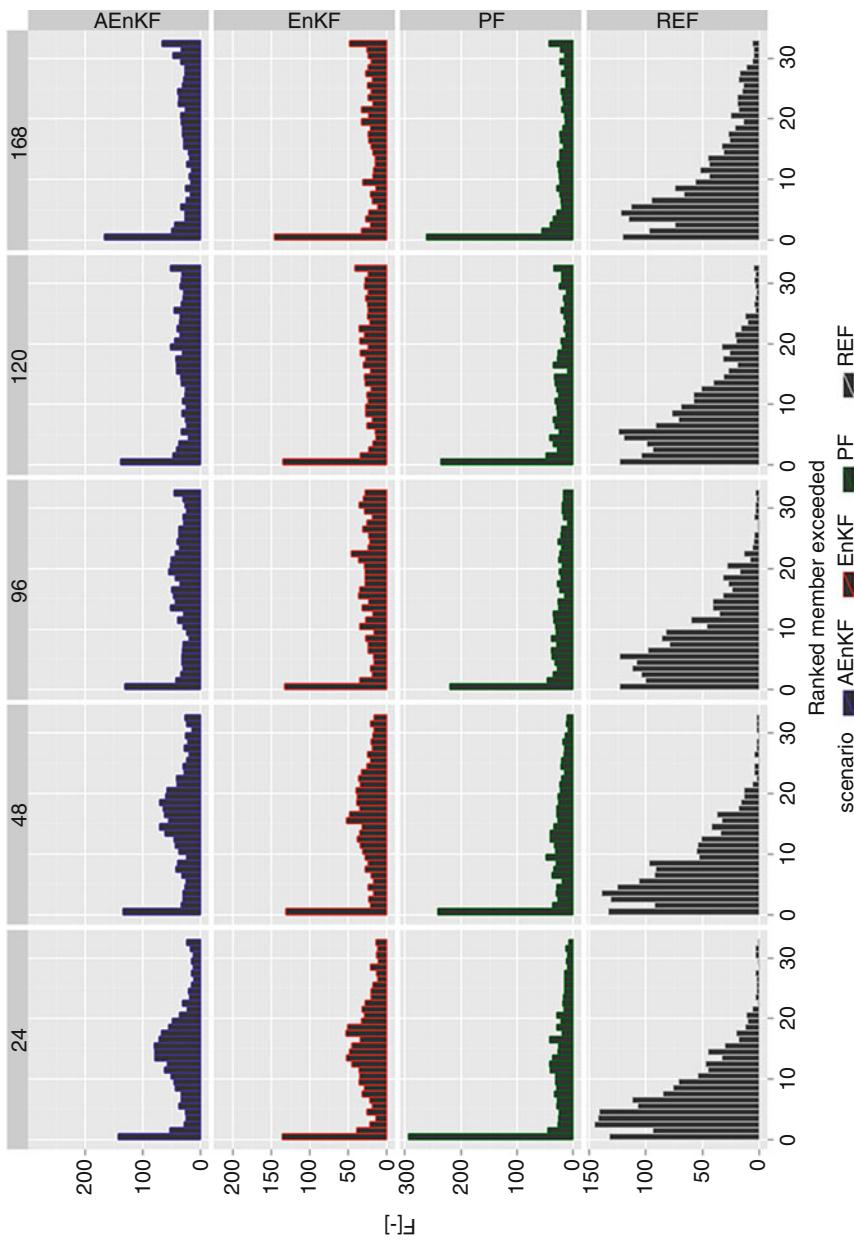


Fig. 15 Talagrand diagrams with AEnKF, EnKF, PF, and reference forecasts

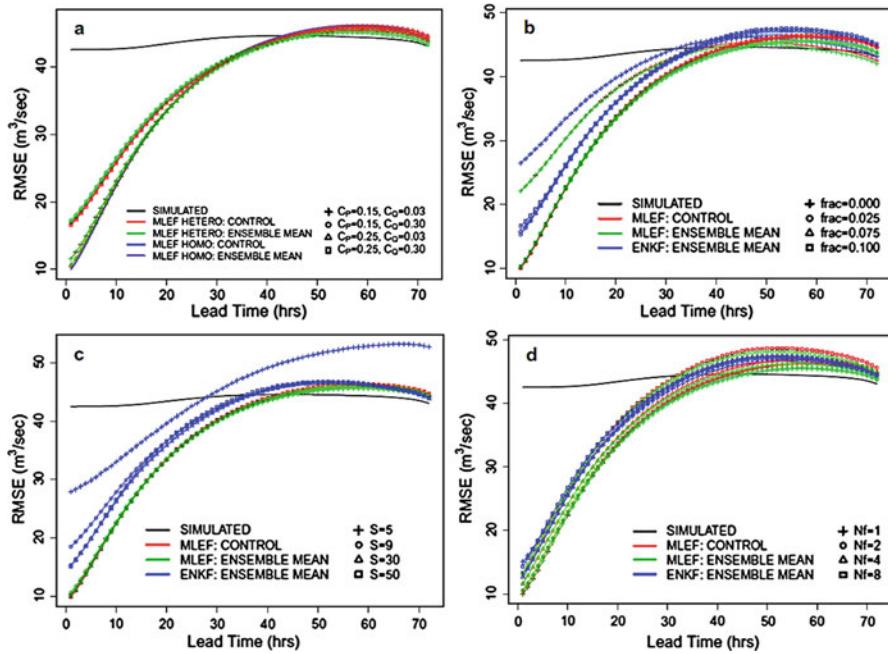


Fig. 16 (a) Effect of heteroscedastic modeling of observation errors in streamflow, MAP, and runoff on RMSE of streamflow prediction with MLEF. The variables CP and CQ denote the multiplicative coefficients for MAP and streamflow observations error standard deviation. The abbreviations HETERO and HOMO denote heteroscedastic and homoscedastic modeling of observation errors, respectively. (b) Performance of streamflow prediction with MLEF and EnKF at varying levels of the error assumed for the model dynamics. In EnKF, a fraction (frac) of soil water content was used to model the standard deviation of the model error for each state variable. In MLEF, frac of the soil water bucket size was used to model the standard deviation of the residual model error. (c) Sensitivity of DA performance on the ensemble size (S). (d) Sensitivity of DA performance on the number of streamflow observations (Nf) assimilated at a given assimilation cycle. (From Rafieeinab et al. 2014)

quality of the solution in the case of asymmetric distributions as well as the suboptimal nature of the solution due to the nonlinear relationship between flow observations and model states. At Fig. 16d, MLEF results deteriorate with assimilating a larger number of streamflow observations. On the other hand, EnKF results improve up to four hourly streamflow observations assimilated per cycle and decrease when the number is increased further. Since the convolution operation of UH renders assimilating a single observed flow already amount to adjusting runoff simulations at multiple time steps, multiple streamflow observations may not be readily translated via UH into a dynamically consistent runoff time series – the reason for the performance degradation in MLEF. The suboptimal nature of EnKF solution largely reduces its sensitivity to the number of streamflow observations assimilated.

4 Benefits and Challenges

Extensive research, including examples presented in the previous section, has illustrated capabilities of streamflow DA to improve model predictions and reduce associated uncertainty by combining information from observations and models. Many streamflow DA problems are, however, very nonlinear not only in the physical processes but also in the observational processes (Weerts et al. 2013). Despite real-time streamflow measurements at multiple locations, most of state variables in hydrologic systems remain unobserved and hence are subject to large degrees of freedom. In this section, we discuss benefits and challenges to further improve streamflow DA in terms of large-scale modeling, multi-data assimilation, and timing errors. One of the key efforts in hydrologic science is to build continental- and global-domain hydrologic models with fine spatiotemporal resolutions for water security assessments (Clark et al. 2016). Unlike the conventional tactics at catchment scale, high-resolution streamflow DA at large scale poses very underdetermined inverse problems. While DA may capitalize on increasing computational power and advances in observational techniques to resolve challenges, development of improved DA techniques will require a rigorous evaluation as well as major international interdisciplinary collaborations. Timing error estimation is also one of the key topics which have received scant attention despite its importance in streamflow DA.

4.1 Large-Scale Streamflow DA

There have been increasing advances in large-scale hydrologic modeling. Especially, high-resolution global hydrology and land surface modeling have received critical attention due to the availability of high-resolution terrain and input forcing data, rapidly increasing computational capabilities and growing demands by community and stake holders. For instance, the pan-European Flood Awareness System (EFAS) runs the hydrologic model, LISFLOOD, on a 5-km grid for the entire European domain forced by different ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (ECMWF; Knijff et al. 2010). In the United States, the National Water Model (NWM) generates streamflow analysis and forecast for 2.7 million river reaches and other hydrologic information on 1-km and 250-m grids over the CONUS (Gochis et al. 2014). Currently, the NWM employs a nudging scheme to assimilate more than 6000 USGS streamflow observations and provides initial conditions for its forecasts. More sophisticated streamflow assimilation strategies will be introduced in the next operational versions of the NWM. A summary of the various large-scale models can be also found in Bierkens et al. (2015) and references therein. The success of high-resolution modeling depends not only on the capacity for hyper-resolution (0.1–1 km) modeling but also on the ability to provide the model with the forcing inputs, parameters, initial conditions (IC), and boundary conditions (BC) that are skillful at the commensurate resolution. Effective translation of these data from coarse grids and points to a fine grid is the

key to making “locally relevant hydrologic models everywhere” (Bierkens et al. 2015) a reality.

Meanwhile, capability of streamflow DA has been commonly verified at catchment scales, while large-scale demonstration is still sparse. In recent years, a number of studies have shown regional-scale operational applications of streamflow DA using KF-based methods (Bourgin et al. 2014; Randrianasolo et al. 2014; Thirel et al. 2010a, b). For instance, Météo France (Thirel et al. 2010a, b) has demonstrated large improvement in streamflow prediction by assimilating streamflow observations at a regional scale; they assimilated streamflow data from 186 stream gauge stations in France into the distributed hydrometeorological model, SIM (Habets et al. 2008), via a form of extended Kalman filter (EKF). Two main challenges in streamflow DA at large scale are (1) expensive computation required for ensemble or iterative simulation and (2) suitability of streamflow DA methodology in high-dimensional domains. DA methods applied for DHMs at catchment scale are expected to be implemented for larger scale with coarse grids but efficient parameterization (Rakovec et al. 2016). In addition to exploiting advances in parallel computation, we can also make more effective use of the available computing through more efficient/agile models (e.g., use of hydrologic similarity concepts) within the framework of DA (Clark et al. 2017).

4.2 Multi-Data Assimilation

Hydrologic models are not limited to simulating streamflow. They can be also used to predict other hydrologic fluxes and states, such as evapotranspiration, groundwater storage, surface storage, and soil moisture. Reliable quantitative estimates of these variables may help decision-makers to cope with water, food, energy, and economic issues particularly in times of droughts and floods. Traditionally, the hydrologic models are constrained against river discharge only; however, if other hydrologic variables are of interest and their measurements are available, there remains enough room for multi-data assimilation to reduce uncertainties of hydrologic variables different from river discharge, such as microwave remote sensing products of soil moisture and snow (e.g., Moradkhani 2008). For example, it was shown that the short-term predictive skill of a rainfall-runoff model could be improved when remotely sensed estimates of soil moisture (e.g., Crow and Ryu 2009; Parajka et al. 2006; Wanders et al. 2014; Yan and Moradkhani 2016) or in situ soil moisture observations (e.g. Lee et al. 2011) were employed.

Additionally, the snow cover fraction (SCF) data from the moderate-resolution imaging spectroradiometer resulted in consistent improvements on snow and streamflow predictions in snow-dominated region (Liu et al. 2013). The long-term predictive skills of discharge can be enhanced by remotely sensed monthly anomalies of the Earth’s gravity field retrieved by the Gravity Recovery and Climate Experiment (GRACE). In addition, assimilation of long-memory GRACE data using a simple auto-regressive model could reveal predisposition of a river basin to flooding as much as 5–11 months in advance (Reager et al. 2014). In general,

GRACE data are beneficial in particular in data-sparse regions (e.g., Tangdamrongsub et al. 2015).

The assimilation of multiple types of observations is common in weather, ocean, and land surface modeling. For instance, observation types assimilated in the study of Aberson et al. (2014) included wind and temperature from different measurement devices such as airborne Doppler radar and flight-level dropwindsonde; the ocean assimilation system such as NEMOVAR (Mogensen and Balmaseda 2012) assimilated temperature and salinity profiles as well as sea level anomalies. In spite of potential gains and a couple of demonstration in synthetic experiments, multi-data assimilation still remains challenging in real-world streamflow modeling. Lee et al. (2011) discussed that a combination of structural and parametric errors in the hydrologic models, less than accurate modeling of scale-dependent and heteroscedastic uncertainties, and large observational uncertainty and microscale variability in in situ soil moisture data was responsible for the lack of additional improvement by multi-data assimilation in the real-world experiment. They also suggested that aggregation of state variables at different temporal scales might yield a more effective strategy for assimilating multiple data with different temporal memory. In the discussion of the synthetic experiment on combined assimilation of streamflow and snow water equivalent, Bergeron et al. (2016) indicated that biases and unknown errors in models and observations are the major challenge in real-world multi-data assimilation.

Last but not least, multiple spatial and temporal resolutions of data and model setups also make it difficult to adequately exploit the information content of the DA system.

4.3 Timing Errors

Streamflow DA techniques developed so far update states based on the difference in the magnitude of observed and simulated flow at the concurrent time step which cannot explicitly account for timing errors in simulated hydrographs. To account for flow timing errors in streamflow DA, timing error estimation should precede the DA procedure utilizing timing error information. Flow timing errors in a modeled hydrograph can be estimated by visual or wavelet-based approaches. The visual approach mimics the process of comparing two time series with human eye and brain in manual analyses. The wavelet-based approach calculates timing errors based on the phase difference between two time series computed from the cross wavelet transform (XWT). As examples of the former, the series distance (SD; Ehret and Zehe 2011) and the hydrograph matching algorithm (HMA; Ewen 2011) have shown usefulness in quantifying separately timing and magnitude errors in hydrographs and improving model calibration results. Using the latter approach, Liu et al. (2011) improved peak flow timing as well as the overall shape of hydrographs by simply shifting simulated hydrographs based on XWT-based timing error estimates.

Utilizing timing error estimates in streamflow DA may be achieved by penalizing timing errors directly in DA formulations or correcting timing errors prior to the

assimilation. The former entails modifying the objective function in the case of VAR. The latter consists of two steps that shift the hydrograph as much as the timing error estimate in the first step and then assimilate observations for updating states in the second step. This is similar to the DA by field alignment (Ravela et al. 2007) or phase-correcting DA (Brewster 2003) that corrects position and amplitude errors sequentially.

References

- S.D. Aberson, A. Aksoy, K.J. Sellwood, T. Vukicevic, X. Zhang, Mon. Weather Rev. **143**, 511 (2014)
- K.M. Andreadis, G.J.-P. Schumann, Adv. Water Resour. **73**, 44 (2014)
- K.M. Andreadis, E.A. Clark, D.P. Lettenmaier, D.E. Alsdorf, Geophys. Res. Lett. **34**, L10403 (2007)
- P. Arnaud, J. Lavabre, C. Fouchier, S. Diss, P. Javelle, Hydrol. Sci. J. **56**, 397 (2011)
- J.M. Bergeron, M. Trudel, R. Leconte, Hydrol. Earth Syst. Sci. **20**, 4375 (2016)
- K. Beven, Hydrol. Sci. J. **61**, 1652 (2016)
- K. Beven, P.J. Smith, A. Wood, Hydrol. Earth Syst. Sci. **15**, 3123 (2011)
- S. Biancamaria, F. Frappart, A.-S. Leleu, V. Marieu, D. Blumstein, J.-D. Desjonquères, F. Boy, A. Sottolichio, A. Valle-Levinson, Adv. Space Res. **59**, 128 (2017)
- M.F.P. Bierkens, V.A. Bell, P. Burek, N. Chaney, L.E. Condon, C.H. David, A. de Roo, P. Döll, N. Drost, J.S. Famiglietti, M. Flörke, D.J. Gochis, P. Houser, R. Hut, J. Keune, S. Kollet, R.M. Maxwell, J.T. Reager, L. Samaniego, E. Sudicky, E.H. Sutanudjaja, N. van de Giesen, H. Winsemius, E.F. Wood, Hydrol. Process. **29**, 310 (2015)
- C.M. Birkett, L.A.K. Mertes, T. Dunne, M.H. Costa, M.J. Jasinski, J. Geophys. Res. Atmos. **107**, 8059 (2002)
- D.M. Bjerklie, S. Lawrence Dingman, C.J. Vorosmarty, C.H. Bolster, R.G. Congalton, J. Hydrol. **278**, 17 (2003)
- D.M. Bjerklie, D. Moller, L.C. Smith, S.L. Dingman, J. Hydrol. **309**, 191 (2005)
- E. Blayo, M. Bocquet, E. Cosme, L. F. Cugliandolo (eds.), *Advanced Data Assimilation for Geosciences: Lecture Notes of the Les Houches School of Physics: Special Issue, June 2012*, 1st edn. (Oxford University Press, Oxford, 2014)
- F. Bourgin, M.H. Ramos, G. Thirel, V. Andréassian, J. Hydrol. **519**(Part D), 2775 (2014)
- K.A. Brewster, Mon. Weather Rev. **131**, 480 (2003)
- H. Chen, D. Yang, Y. Hong, J.J. Gourley, Y. Zhang, Adv. Water Resour. **59**, 209 (2013)
- M.P. Clark, A.G. Slater, J. Hydrometeorol. **7**, 3 (2006)
- M.P. Clark, D.E. Rupp, R.A. Woods, X. Zheng, R.P. Ibbitt, A.G. Slater, J. Schmidt, M.J. Uddstrom, Adv. Water Resour. **31**, 1309 (2008)
- M.P. Clark, B. Schaeffer, S.J. Schymanski, L. Samaniego, C.H. Luce, B.M. Jackson, J.E. Freer, J.R. Arnold, R.D. Moore, E. Istanbulluoglu, S. Ceola, Water Resour. Res. **52**, 2350 (2016)
- M.P. Clark, M.F.P. Bierkens, L. Samaniego, R.A. Woods, R. Uijenhoet, K.E. Bennet, V.R.N. Pauwels, X. Cai, A.W. Wood, C.D. Peters-Lidard, Hydrol. Earth Syst. Sci. Discuss. **2017**, 1 (2017)
- G. Coxon, J. Freer, I.K. Westerberg, T. Wagener, R. Woods, P.J. Smith, Water Resour. Res. **51**, 5531 (2015)
- W.T. Crow, D. Ryu, Hydrol. Earth Syst. Sci. **13**, 1 (2009)
- C.M. DeChant, H. Moradkhani, Water Resour. Res. **48**, W04518 (2012)
- U. Ehret, E. Zehe, Hydrol. Earth Syst. Sci. **15**, 877 (2011)
- G. Ercolani and F. Castelli, Water Resources Research **53**, 158 (2017)
- J. Ewen, J. Hydrol. **408**, 178 (2011)
- J. Fohringer, D. Dransch, H. Kreibich, K. Schröter, Nat. Hazards Earth Syst. Sci. **15**, 2725 (2015)

- J. García-Pintado, J.C. Neal, D.C. Mason, S.L. Dance, P.D. Bates, *J. Hydrol.* **495**, 252 (2013)
- K.P. Georgakatos, *Water Resour. Res.* **22**, 2083 (1986)
- L. Giustarini, P. Matgen, R. Hostache, M. Montanari, D. Plaza, V.R.N. Pauwels, G.J.M. De Lannoy, R. De Keyser, L. Pfister, L. Hoffmann, H.H.G. Savenije, *Hydrol. Earth Syst. Sci.* **15**, 2349 (2011)
- D.J. Gochis, W. Yu, D.N. Yates., Available at https://www.ral.ucar.edu/projects/wrf_hydro. (2014)
- P. Goovaerts, *Geostatistics for Natural Resources Evaluation* (Oxford University Press, New York, 1997)
- N.J. Gordon, D.J. Salmond, A.F. Smith, in *Radar Signal Process. IEE Proc. F* (1993), pp. 107–113
- F. Habets, A. Boone, J.L. Champeaux, P. Etchevers, L. Franchistéguy, E. Leblois, E. Ledoux, P. Le Moigne, E. Martin, S. Morel, J. Noilhan, P. Quintana Seguí, F. Rousset-Regimbeau, P. Viennot, *J. Geophys. Res. Atmos.* **113**, D06113 (2008)
- J. Juston, P.-E. Jansson, D. Gustafsson, *Hydrol. Process.* **28**, 2509 (2014)
- R.E. Kalman, *J. Basic Eng.* **82**, 35 (1960)
- N. Kantas, A. Doucet, S.S. Singh, J. Maciejowski, N. Chopin, *Stat. Sci.* **30**, 328 (2015)
- P.-E. Kirttetter, J.J. Gourley, Y. Hong, J. Zhang, S. Moazamigoodarzi, C. Langston, A. Arthur, *Water Resour. Res.* **51**, 1422 (2015)
- P.K. Kitanidis, R.L. Bras, *Water Resour. Res.* **16**, 1034 (1980)
- J.M.V.D. Knijff, J. Younis, A.P.J.D. Roo, *Int. J. Geogr. Inf. Sci.* **24**, 189 (2010)
- S. Koswatte, K. McDougall, X. Liu, *Surv. Rev.* **47**, 307 (2015)
- G. Kuczera, *Water Resour. Res.* **32**, 2119 (1996)
- W.A. Lahoz, P. Schneider, *Front. Environ. Sci.* **2**, 1 (2014)
- J. Le Coz, B. Renard, L. Bonnifait, F. Branger, R. Le Boursicaud, *J. Hydrol.* **509**, 573 (2014)
- J. Le Coz, A. Patalano, D. Collins, N.F. Guillén, C.M. Garcia, G.M. Smart, J. Bind, A. Chiaverini, R. Le Boursicaud, G. Dramais, I. Braud, *J. Hydrol.* **541**(Part B), 766 (2016)
- H. Lee, D.-J. Seo, *Adv. Water Resour.* **74**, 196 (2014)
- H. Lee, D.-J. Seo, V. Koren, *Adv. Water Resour.* **34**, 1597 (2011)
- H. Lee, D.-J. Seo, S.J. Noh, *J. Hydrol.* **542**, 373 (2016)
- D.P. Lettenmaier, D. Alsdorf, J. Dozier, G.J. Huffman, M. Pan, E.F. Wood, *Water Resour. Res.* **51**, 7309 (2015)
- Y. Liu, J. Brown, J. Demargne, D.-J. Seo, *J. Hydrol.* **397**, 210 (2011)
- Y. Liu, A.H. Weerts, M. Clark, H.-J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.-J. Seo, D. Schwanenberg, P. Smith, A.I.J.M. van Dijk, N. van Velzen, M. He, H. Lee, S.J. Noh, O. Rakovec, P. Restrepo, *Hydrol. Earth Syst. Sci.* **16**, 3863 (2012)
- Y. Liu, C.D. Peters-Lidard, S. Kumar, J.L. Foster, M. Shaw, Y. Tian, G.M. Fall, *Adv. Water Resour.* **54**, 208 (2013)
- C.S. Lowry, M.N. Fienen, *Ground Water* **51**, 151 (2013)
- D. Maidment, *Handbook of Hydrology*, 1st edn. (McGraw-Hill Education, New York, 1993)
- P. Matgen, M. Montanari, R. Hostache, L. Pfister, L. Hoffmann, D. Plaza, V.R.N. Pauwels, G.J.M. De Lannoy, R. De Keyser, H.H.G. Savenije, *Hydrol. Earth Syst. Sci.* **14**, 1773 (2010)
- M. Mazzoleni, M. Verlaan, L. Alfonso, M. Monego, D. Norbiato, M. Ferri, D.P. Solomatine, *Hydrol. Earth Syst. Sci.* **21**, 839 (2017)
- R. T. McNider, J. R. Christy, D. Moss, K. Doty, C. Handyside, A. Limaye, A. Garcia y Garcia, and G. Hoogenboom, *J. Appl. Meteor. Climatol.* **50**, 1459 (2011)
- H.K. McMillan, I.K. Westerberg, *Hydrol. Process.* **29**, 1873 (2015)
- H. McMillan, B. Jackson, M. Clark, D. Kavetski, R. Woods, *J. Hydrol.* **400**, 83 (2011)
- H. McMillan, T. Krueger, J. Freer, *Hydrol. Process.* **26**, 4078 (2012)
- H.K. McMillan, E.Ö. Hreinsson, M.P. Clark, S.K. Singh, C. Zammit, M.J. Uddstrom, *Hydrol. Earth Syst. Sci.* **17**, 21 (2013)
- K. Mogensen, W. Balmaseda, *The NEMOVAR Ocean Data Assimilation System as Implemented in the ECMWF Ocean Analysis for System 4* (ECMWF, Reading, 2012)
- H. Moradkhani, *Sensors* **8**, 2986 (2008)

- H. Moradkhani, K.-L. Hsu, H. Gupta, S. Sorooshian, Water Resour. Res. **41**, W05012 (2005)
- A.J. Newman, M.P. Clark, J. Craig, B. Nijssen, A. Wood, E. Gutmann, N. Mizukami, L. Brekke, J.R. Arnold, J. Hydrometeorol. **16**, 2481 (2015)
- S.J. Noh, Y. Tachikawa, M. Shiiba, S. Kim, Hydrol. Earth Syst. Sci. **15**, 3237 (2011)
- S.J. Noh, Y. Tachikawa, M. Shiiba, S. Kim, J. Hydrol. Eng. **18**, 1684 (2013)
- S.J. Noh, O. Rakovec, A.H. Weerts, Y. Tachikawa, J. Hydrol. **519**(Part D), 2707 (2014)
- S.J. Noh, M. Mazzoleni, H. Lee, Y. Liu, D.-J. Seo, D.P. Solomatine, in *Proceedings of Hydroinformatics 2016*, Seoul, 2016
- T. O'Donnell, Hydrol. Sci. J. **30**, 479 (1985)
- D. Ocio, N. Le Vine, I. Westerberg, F. Pappenberger, W. Buytaert, Water Resour. Res. **53**, 4197 (2017)
- F. Pan, C. Wang, X. Xi, J. Hydrol. **540**, 670 (2016)
- J. Parajka, V. Naeimi, G. Blöschl, W. Wagner, R. Merz, K. Scipal, Hydrol. Earth Syst. Sci. **10**, 353 (2006)
- S. Pathiraja, L. Marshall, A. Sharma, H. Moradkhani, Water Resour. Res. **52**, 3350 (2016)
- A. Petersen-Øverleir, T. Reitan, J. Hydrol. **311**, 188 (2005)
- W.H. Press, B.P. Flannery, S.A. Teukolsky, W.T. Vetterling, *Numerical Recipes in Fortran 77: The Art of Scientific Computing*, 2nd edn. (Cambridge University Press, Cambridge, UK, 1992)
- A. Rafieeinabab, D.-J. Seo, H. Lee, S. Kim, J. Hydrol. **519**(Part D), 2663 (2014)
- O. Rakovec, P. Hazenberg, P.J.J.F. Torfs, A.H. Weerts, R. Uijlenhoet, Hydrol. Earth Syst. Sci. **16**, 3419 (2012)
- O. Rakovec, A.H. Weerts, J. Sumihar, R. Uijlenhoet, Hydrol. Earth Syst. Sci. **19**, 2911 (2015)
- O. Rakovec, R. Kumar, J. Mai, M. Cuntz, S. Thober, M. Zink, S. Attinger, D. Schäfer, M. Schrön, L. Samaniego, J. Hydrometeorol. **17**, 287 (2016)
- A. Randrianasolo, G. Thirel, M.H. Ramos, E. Martin, J. Hydrol. **519**(Part D), 2676 (2014)
- S. Ravela, K. Emanuel, D. McLaughlin, Phys. Nonlinear Phenom. **230**, 127 (2007)
- J.T. Reager, B.F. Thomas, J.S. Famiglietti, Nat. Geosci. **7**, 588 (2014)
- B. Ristic, S. Arulampalam, N. Gordon, *Beyond the Kalman Filter: Particle Filters for Tracking Applications* (Artech House, Boston, 2004)
- P. Sakov, G. Evensen, and L. Bertino, Tellus A, **62** (2010)
- P. Salamon, L. Feyen, Water Resour. Res. **46**, W12501 (2010)
- D.-J. Seo, L. Cajina, R. Corby, T. Howieson, J. Hydrol. **367**, 255 (2009)
- D.-J. Seo and J. P. Breidenbach, J. Hydrometeor **3**, 93 (2002)
- D.-J. Seo, Y. Liu, H. Moradkhani, A. Weerts, J. Hydrol. **519**(Part D), 2661 (2014)
- M. Shiiba, T. Takasao, in *Proceedings of Third International Symposium on Stochastic Hydraulics*, Tokyo, 1980
- J. A. Smith and W. F. Krajewski, J. Appl. Meteor. **30**, 397 (1991)
- L. Sun, O. Seidou, I. Nistor, K. Liu, Hydrol. Sci. J. **61**, 2348 (2016)
- N. Tangdamrongsub, S.C. Steele-Dunne, B.C. Gunter, P.G. Ditmar, A.H. Weerts, Hydrol. Earth Syst. Sci. **19**, 2079 (2015)
- J. Tao, D. Wu, J. Gourley, S.Q. Zhang, W. Crow, C. Peters-Lidard, A.P. Barros, J. Hydrol. **541**(Part A), 434 (2016)
- G. Thirel, E. Martin, J.-F. Mahfouf, S. Massart, S. Ricci, F. Habets, Hydrol. Earth Syst. Sci. **14**, 1623 (2010a)
- G. Thirel, E. Martin, J.-F. Mahfouf, S. Massart, S. Ricci, F. Regimebeau, F. Habets, Hydrol. Earth Syst. Sci. **14**, 1639 (2010b)
- M.J. Tourian, C. Schwatke, N. Sneeuw, J. Hydrol. **546**, 230 (2017)
- J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, G. Schoups, Adv. Water Resour. **51**, 457 (2013)
- N. Wanders, D. Karssenberg, A. de Roo, S.M. de Jong, M.F.P. Bierkens, Hydrol. Earth Syst. Sci. **18**, 2343 (2014)
- C.-H. Wang, Y.-L. Bai, J. Hydrol. Eng. **13**, 290 (2008)
- A.H. Weerts, G.Y.H. El Serafy, Water Resour. Res. **42**, 1 (2006)

- A.H. Weerts, D.-J. Seo, M. Werner, J. Schaake, *Applied Uncertainty Analysis for Flood Risk Management* (Imperial College Press, London, 2013)
- I. Westerberg, J.-L. Guerrero, J. Seibert, K.J. Beven, S. Halldin, *Hydrol. Process.* **25**, 603 (2011)
- E.F. Wood, A. Szöllösi-Nagy, *Water Resour. Res.* **14**, 577 (1978)
- X. Xie, D. Zhang, *Water Resour. Res.* **49**, 7350 (2013)
- H. Yan, H. Moradkhani, *Adv. Water Resour.* **94**, 364 (2016)
- K. Yan, G. Di Baldassarre, D.P. Solomatine, G.J.-P. Schumann, *Hydrol. Process.* **29**, 3368 (2015)
- M. Zupanski, *Mon. Weather Rev.* **133**, 1710 (2005)

Part VII

Post-processing of Hydrological Ensemble Forecasts



Motivation and Overview of Hydrological Ensemble Post-processing

Thomas M. Hopson, Andy Wood, and Albrecht H. Weerts

Contents

1	Introduction and Motivation	784
2	Hydrological Versus Meteorological Post-processing	785
3	General Approaches to Hydrologic Post-processing	787
3.1	Total Hydrologic Uncertainty Versus Explicit Uncertainty Decomposition	787
3.2	Error Distributions Versus Ensemble Time Series	788
3.3	Overview of Approaches	789
4	Post-processing Requirements and Challenges	790
5	Overview of Section Contents	791
	References	791

Abstract

In this introduction to this chapter on hydrologic post-processing, we discuss the different but complementary directives that the “art” of post-processing must satisfy: the particular directive defined by specific applications and user needs; versus the general directive of making any ensemble member indistinguishable from the observations. Also discussed are the features of hydrologic post-processing that are similar and separate from meteorological post-processing,

T. M. Hopson

Research Applications Laboratory, National Center for Atmospheric Research, Boulder, CO, USA
e-mail: hopson@ucar.edu

A. Wood (✉)

National Center for Atmospheric Research, Boulder, CO, USA
e-mail: andywood@ucar.edu

A. H. Weerts

Operational Water Management, Inland Water Systems, Deltares, Delft, The Netherlands
Hydrology and Quantitative Water Management Group, Wageningen University and Research,
Wageningen, The Netherlands
e-mail: albrecht.weerts@deltares.nl

providing a tie-in to early chapters in this handbook. We also provide an overview of the different aspects the practitioner should keep in mind when developing and implementing algorithms to adequately “correct and calibrate” ensemble forecasts: when forecast uncertainties should be characterized separately versus maintaining a “lumped” approach; additional aspects of hydrological ensembles that need to be maintained to satisfy additional user requirements, such as temporal covariability in the ensemble time series, an overview of the different post-processing approaches being used in practice and in the literature, and concluding with a brief overview of more specific requirements and challenges implicit in the “art” of post-processing.

Keywords

Post-processing · Conditional forecast · Regression · Estimation · Uncertainty · Systematic error · Bias correction · Streamflow forecast · Hydrologic modeling · Covariance

1 Introduction and Motivation

The “art and practice” of post-processing hydrologic ensemble forecasts (primarily referring to streamflow forecasting in this chapter.) can be viewed as trying to satisfy two masters. On the one hand is the particular directive: post-processing needs to account for the user of the forecasts themselves and how to best provide the forecasts in a form that is most conducive and beneficial to the needs of the particular application. Will the forecasts be used, say, to provide the probability of exceeding a specific river stage height at a specific location at a specific lead time or be used as further input into a hydropower generation optimization program over a given season and a particular reservoir? The process of tailoring to the “particular” may require disaggregating or integrating the original hydrologic ensemble output over space and/or time scales; or it may involve thresholding the output to particular binary states that are of societal importance, such as bank-full exceedance levels.

On the other hand is the general directive: where the hydrologic modeling process inevitably falls short in its ability to simulate the full physics, post-processing is required to characterize the residual stochastic variability, as well as to step in to remove systemic deficiencies in the forecasting process. Post-processing, then, strives to satisfy the heuristic of making any randomly chosen ensemble member of the forecast indistinguishable from the future observation itself while retaining (and enhancing if possible) the skill in the dynamic, physically based model output and reproducing the spatial and temporal co-variability expected of the (natural) system being modeled.

It is the demands of the latter objective that distinguish the post-processing of hydrologic ensemble prediction from classical uncertainty estimation in hydrologic time-series analysis, which models the central tendency of a predictive distribution and parameterizes its associated estimation error. This idea of the ensemble member mimicking the behavior of the actual system itself is implied in ensemble prediction.

In both the context of atmospheric and climate science, as well as in land surface hydrology, physically and numerically based dynamical models are being developed and implemented to mimic and predict both the physical processes and uncertainties of each particular science. The models strive to generate outputs that mimic the “real world” in both space and time, with its commensurate observations and uncertainties – in essence, providing numerical approximations to the Fokker-Planck equation (Palmer 2000), which describes the time evolution of the probability density function (PDF) describing the future state of a system under the influence of both inertial forces and random forces. Inertial forces provide forecast signal and skill, while random forces diffuse predictability and increase uncertainty, widening the PDF.

But this is a high bar to achieve, and it is the job of post-processing to correct systematic biases in model-predicted future mean states and their uncertainty and to incorporate additional sources of uncertainty that are underrepresented in the dynamic, physically based model output. In some senses, this process can be viewed as trying to make the forecasts as generalizable as possible, such that the final ensemble outputs, as close to an accurate representation of the hydrological system being modeled as possible, can then be “sliced and diced” further to meet the needs of specific applications.

Hydrologic ensemble prediction also implies that ensembles themselves would provide a uniform and ideally a “complete” sampling of the uncertainty PDF of a hydrologic variable, which would then allow the user to map the ensembles to reliable probabilities that can be used for societal benefit, such as of the likelihood of damaging extremes (e.g., probability of the river cresting over a road vital for evacuation) and associated societal risks. However, given that practical and cost considerations limit the scope and fidelity of dynamical physically based model ensemble generation, and of observations to constrain them, a key task of post-processing is to translate limited ensemble model output into an informative and reliable conditional PDF while at the same time observing the post-processing equivalent to the Hippocratic oath of “doing no harm” to the inherent skill of the raw model outputs.

It follows, in some sense, from the dual requirements (with their various aspects and interpretations) imposed by these two “masters” on hydrologic ensemble forecast post-processing, that a great variety of algorithms have been developed over the years to satisfy their ends. The purpose of this chapter is to provide an introduction to this challenging “multi-objective” topic.

2 Hydrological Versus Meteorological Post-processing

Many aspects of the above discussion apply to both hydrologic and meteorological ensemble forecast post-processing. Both seek, from the most general standpoint, to provide a conditional predictive distribution $F(\cdot)$ that represents a reliable and unbiased expectation for the observed variable of interest y_i given a model forecast of that variable $X_i = x$ and given any other conditioning factors θ that are relevant to the model forecast errors.

$$F(y|x) = P(y_i \leq y | X_i = x, \theta) \quad (1)$$

Post-processing in both contexts strives to generate forecast ensembles that are indistinguishable, statistically, from the realizable observation. Yet greater complexity arises in the need to replicate the physical system's spatial, temporal, and multivariate covariance structure, which is important for event timing estimation and for forcing secondary applications that rely on the accurate representation of these complex relationships.

However, from a more particular standpoint, meteorological post-processing clearly differs from the hydrological in that the former is being used to drive the models of the latter (in the context of this handbook), with hydrologic models often requiring realistic spatially and temporally distributed input fields (in the case of distributed and semi-distributed modeling approaches). Hydrology is sensitive to meteorological variability across space and time; thus an acute meteorological post-processing need is to represent and preserve the space-time covariance of the meteorological fields of a given drainage area. To see the importance of this function, consider that many of the pathways of the movement of water through the land surface are fundamentally nonlinear; and as such, for a given basin-average rainfall amount, the spatial distribution of this rainfall can have a significant impact on channelized outflow and should be accurately represented in the forcing fields, either explicitly (through the physical modeling processes) or parameterized. Because hydrology provides space and time aggregation of inputs, such attention to covariance is often of lesser importance, allowing hydrologic post-processing to focus on errors at point locations for many use cases. There are, however, use cases in which space/time covariance structure must be accounted for in the generation of hydrologic ensembles. Consider the situation where the modeling process of higher-order stream channels (i.e., basin outlets) depends on separate modeling of upstream (lower-order) stream channel flows (Regonda et al. 2006) and a synchronization of their runoff generation. This issue of preserving temporal and spatial correlation structures is addressed throughout this chapter.

Another fundamental difference between hydrologic and meteorological post-processing deals with the end goal of the hydrologic ensemble forecasting chain. Hydrologic post-processing is often the last “gatekeeper” on generating the final outputs that are appropriate for the end user, implying the output needs to be tailored for specific user needs (examples of which are particular temporal aggregation lengths, probability thresholds, etc.). The output also must provide a complete and reliable (and ideally, skillful) uncertainty quantification in the hydrologic forecasts of underrepresented residual sources of uncertainty in the ensemble forecasting chain. In contrast, the meteorological ensembles are primarily focused on meeting the needs of the “next step” in the uncertainty forecasting chain and, as such, are more likely geared toward generating ensembles that are physically realizable from the vantage point of the hydrologic model.

A final, notable difference between meteorological and hydrologic forecast post-processing is that hydrological systems, to a much greater extent, balance predictability related to the inertia of a watershed, which are often detectable in the

initial hydrologic conditions of the forecast, with predictability or uncertainty in the boundary forcing of the forecast – i.e., the future meteorological inputs to the system (see, e.g., Wood and Lettenmaier 2008). Identical boundary forcing during a forecast period may lead to different systematic biases because the initial conditions of the forecast are quite different. For example, a forecast may have a systematic low bias in predicting the watershed response to the first storm of the season (because the watershed model has a systematically low soil moisture); but the same forecast may show a systematic high bias for the same type of storm hitting the basin later in the wet season, (if the watershed model has saturations that are systematically too high). The atmosphere, by contrast, has less memory and inertia. (With the caveat that the atmospheric *boundary conditions* may have long memory and inertia (in contrast to initial conditions), aka ocean states, and for that matter, the land surface soil moisture states as well (as treated as a boundary condition for the atmosphere in this context).) This dual dependence on initial conditions and boundary forcing means that hydrologic post-processing often needs to represent more complex variations in regimes and their associated systematic uncertainties. This increased dimensionality to the error structure also means that the available sample sizes for characterizing error (across regimes) can become a limiting factor in training robust post-processing techniques.

3 General Approaches to Hydrologic Post-processing

3.1 Total Hydrologic Uncertainty Versus Explicit Uncertainty Decomposition

Error and uncertainty are introduced at essentially every step in the hydrologic ensemble forecasting chain (e.g., errors in the observed and forecasted meteorological forcing fields, hydrologic model initial state uncertainty, inaccuracies in the parameterization and modeling of dynamical processes). To account for the errors and uncertainties in the final forecast outcome, estimated error/uncertainty can be addressed at each step of a deterministic forecast generation chain through the introduction of parameterized error corrections and uncertainty distributions. However, the numerical accounting of such parameterized error distributions is often far more easily accounted for through discrete sampling of the error distributions in the form of ensembles. In many operational hydrologic forecasting chains, the treatment of such uncertainties is readily manifest in the form of ensembles (an example of which would be ensemble rainfall forecast fields derived from numerical weather prediction models). However, often in operational systems, only certain steps in the hydrologic forecasting chain have their uncertainties accounted for up to the point of post-processing. Typically, for instance, hydrologic ensemble forecast systems represent weather and climate uncertainties with meteorological ensemble forecast inputs but lack ensemble treatments of hydrologic model structure and parameter uncertainty and also typically lack ensemble estimates of initial hydrologic states (which themselves are driven by meteorological forcing ensembles).

Whether all or none of the “links” in the hydrologic forecasting chain have tried to account for their introduced uncertainty (up to the point of post-processing), it is typical that significant systematic errors and unaccounted uncertainty will still reside in the ensemble outputs. This is true whether or not these hydrologic outputs derive from a single, multi-model, deterministic, or ensemble modeling process. This residual, uncorrected error and uncertainty become the central challenge for hydrologic post-processing. A reality of all current operational short-to-medium range systems is that irrespective of whether explicit uncertainty and error corrections have been applied at any (or all) of the steps in the forecasting chain, hydrologic forecast post-processing is still one of the most essential system components.

Given the necessity of hydrologic post-processing, a reasonable alternative to explicit uncertainty decomposition and correction throughout the forecast workflow is simply to utilize the final post-processing step to account for the total hydrologic forecasting uncertainty. A special case is single-value hydrologic model post-processing, in which case the final post-processing step accounts for all uncertainty estimation and may take the form of an uncertainty “dressing” that estimates predictive uncertainty from observed past single-value errors (e.g., Hoss et al. 2015; Verkade et al. 2017). It is important to note that the post-processing methodology will be dependent on what has been done “upstream” in the forecast workflow: the treatment of a single deterministic hydrologic model input is distinctly different than the treatment of “raw” ensemble input, just as the post-processing treatment for “random draw” ensemble input will often be different from weighted ensembles (e.g., Bayesian Model Averaging).

3.2 Error Distributions Versus Ensemble Time Series

As discussed above, the hydrologic post-processing algorithm development will almost by necessity need to be dependent on the inputs into the process that come from the hydrologic modeling component (e.g., are these inputs deterministic or ensemble in nature? If ensemble, then are they equally likely or are some preferentially weighted? Is it part of the uncertainty in the form of parameterized error distributions?). But just as the processing steps are dependent on the inputs, the post-processing outputs need also be designed to meet the needs of the end-user application.

Consider the forecasting of river flow at a specific location. If one is only interested in the likelihood the flow rate will exceed a specific threshold at a given point in time, then the practitioner may best focus on post-processed outputs of exceedance level probabilities that could be generated from a parameterized error distribution. However, if one is concerned in, say, the likelihood of the number of times the river will exceed the threshold and for how long, over a given interval of time, then post-processing needs to contain temporal covariance information across time steps, something for which ensemble outputs are more designed for. Depending on the choice of algorithm, post-processing may desynchronize spatial and temporal information in the calibration process, requiring a further step to reintroduce this

information back into the final outputs (e.g., by either utilizing climatological covariance information, e.g., Schaake shuffle, Clark et al. 2004; or by utilizing the covariance information contained in the meteorological forcing fields, e.g., Schefzik et al. 2013). As a result, then, before selecting a post-processing algorithm for calibrating hydrologic forecasts, it is important that the practitioner take stock of both the form and assumptions of the inputs into the process, as well as the form and requirements of the applications utilizing the final post-processed output itself.

3.3 Overview of Approaches

There are many ways to categorize the plethora of short-range and seasonal ensemble forecast post-processing approaches; and just as there are many ways to categorize, there will be many hybrid approaches and exceptions. One way is to delineate approaches that rely on statistical/empirical versus dynamical/model-based approaches (with a delineation more conducive to seasonal forecasts). Alternatively, post-processing algorithms can broadly be characterized into regression approaches whose output generates a global parameterized forecast PDF and those that construct total forecast uncertainty by retaining input ensemble identification (a delineation arguable more conducive to short-range ensemble hydrologic forecasting, with its prevalent use of ensemble numerical weather prediction), which we will use below (with its warts and all).

With respect to regression, approaches are some of the earliest post-processing techniques, examples of which in short-range forecasting include “Model Output Statistics” (e.g., Glahn and Lowry 1972; Wilks 2011), logistic regression (Clark et al. 2004; Wilks 2011), nonhomogeneous regression (Gneiting et al. 2005), quantile regression (Bremnes 2004), and more direct approaches that adjust the forecast output PDF moments directly, where the PDF is implied from the input ensemble sampling (e.g., Hamill and Colucci 1997; Buizza et al. 2003; Wood and Schaake 2008).

In the context of seasonal forecasting, regression-based approaches may or may not utilize input ensembles of future states in the regression process but rather initial hydrologic and geophysical parameter states (Garen 1992; Pagano et al. 2009; Tootle et al. 2007, among many others) and may or may utilize parametric approaches (Sankarasubramanian and Lall 2003; Souza Filho and Lall 2003; Opitz-Stapleton et al. 2007). In fact, for some of these approaches, the regression process itself (through/its associated error estimation) can be viewed as providing the complete “forecasting chain” (of short-range ensemble forecasting parlance), implicitly generating “post-processed” outputs.

Approaches in the latter category of input ensemble identification often allow for the preferential weighting of the input ensemble. Some of these approaches in short-range forecasting include ensemble dressing approaches (Hopson 2005; Hopson and Webster 2010), best member approaches (Roulston and Smith 2003), and weighted ensemble dressing (Fortin et al. 2006), with the latter category having overlap with Bayesian Model Averaging (BMA; Hoeting et al. 1999).

Multi-modeling can be considered a separate approach to removing systematic biases in the ensemble, as well as increasing the (more often than not under-dispersive) ensemble spread. In its simplest form of the “poor person’s ensemble” (Rousseau and Chapelet 1985; Atger 1999), in which forecasts derived from independent prediction systems (who may or may not be ensemble-based) are simply combined, multi-modeling can be considered separate from other post-processing approaches in that it is not reliant on quantifiable knowledge of past model performance. Behind this approach is the theory that although individual systems will likely have their own systemic biases, they will have orthogonal components, leading to an unbiased ensemble in (large) aggregate. As ensemble system weighting becomes more utilized in this approach (implying a reliance on past quantifiable individual forecast performance), the multi-modeling approach gains similarity to other post-processing approaches. Early examples of the utility of weighted multi-modeling approaches in operational systems can be found in both meteorological (Krishnamurti et al. 1999) and hydrologic ensemble systems (Hopson 2005; Hopson and Webster 2010; although here only hydrologic modeling error was “multi-modeled”).

4 Post-processing Requirements and Challenges

Besides considerations of the form of the inputs and outputs that affect specific choice of post-processing algorithm selection (discussed above), there are other general requirements that need to be considered in the post-processing development. Many of these considerations concern the pairings of hydrologic forecasting with their associated, coincident observations available for the process, which we term here the “hindcast” data set.

A primary concern in hydrological post-processing deals with sample size: Does the hindcast data set contain enough samples to allow a robust application of the post-processing algorithm of interest (i.e., estimation of post-processing model parameters)? Such issues need to be considered in the forecasting of hydrologic extremes and in the case of complex models with many variables (degrees of freedom). Sample size is a notorious difficulty in applications for seasonal forecasting, with its limited data availability: consider, for example, a zero-lead forecast of “summer season” snowmelt accumulation, where only one event is available for each year in historical records that may be only 20–30 years long. There are very general and simple “rules of thumb” that can be referred to (i.e., “100 samples are adequate”), but extreme caution must be used here, since sufficiency of sample size depends on a number of factors, such as the specifics of the model being “fitted,” the natural system being modeled, the strength of the correlation between the explanatory (“x”) variables and the response (“y”) variable, and the number of the explanatory variables being used (Knofczynski and Mundfrom 2008; Hanley 2016; Ogundimu et al. 2016).

Where sample sizes are a concern, and for the application of multiple explanatory variables into the post-processing algorithm, generally, cross-validation must be used, (Alternatively, the more computationally efficient Akaike Information Criterion

(AIC), or Bayesian Information Criterion (BIC), or Generalized Cross Validation (GCV) cost functions may also be used, but which require distributional assumptions.) optionally under a stepwise forward- or backward-selection framework (Granger and Newbold 1986), to ensure that “over-fitting” of the post-processing statistical model does not occur. Further considerations need to be considered in the application of cross-validation to ensure it is applied correctly: Are the validation samples independent and also representative of the modeled system’s critical regimes and range of variability? Are the samples independent and identically distributed: Has the hydrologic forecast model remained constant or the quality of the observations changed over the whole hindcast data set? Complicating the issue is the potential impact of climate change or other long-term system trends, which degrade the assumption of stationarity, and is more problematic for purely statistical models that rely on the iid assumption (independent and *identically* distributed) than more physically tied approaches that might represent the physics of long-term drift.

5 Overview of Section Contents

Many of the points discussed in this introduction are discussed in greater detail but in the context of short-range ensemble forecast post-processing (see the chapter ► “[Short-range Ensemble Forecast Post-processing](#)”). While the reader will note areas of algorithmic overlap (e.g., Bayesian Model Averaging), many specific details of their implementation at the different time scales will be distinct, with much of the separation being due to the seasonal-range restrictions on hindcast data set size, skill of numerical weather-climate prediction forcing, and greater potential for disparate regimes in hydrology as a result of predictability influences from initial hydrologic (watershed) moisture states.

References

- F. Atger, The skill of ensemble prediction systems. *Mon. Weather Rev.* **127**(9), 1941–1953 (1999)
- J. Bremnes, Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Weather Rev.* **132**, 338–347 (2004)
- R. Buizza, D.S. Richardson, T.N. Palmer, Benefits of increased resolution in the ECMWF ensemble system and comparison with poor-man’s ensembles. *Q. J. R. Meteorol. Soc.* **129**(589), 1269–1288 (2003)
- M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262 (2004)
- V. Fortin, A.-C. Favre, M. Saïd, Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. *Q. J. R. Meteorol. Soc.* **132**(617), 1349–1369 (2006)
- D.C. Garen, Improved techniques in regression-based streamflow volume forecasting. *J. Water Resour. Plan. Manage.* **118**(6), 654–670 (1992)

- H.R. Glahn, D.A. Lowry, The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**, 1203–1211 (1972)
- T. Gneiting, A.-E. Raftery, A.-H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**(5), 1098–1118 (2005)
- C.W.J. Granger, P. Newbold, *Forecasting Economic Time Series* (Academic, Orlando, 1986)
- T.M. Hamill, S.J. Colucci, Verification of Eta-RSM short-range ensemble forecasts. *Mon. Weather Rev.* **125**(6), 1312–1327 (1997)
- J.A. Hanley, Simple and multiple linear regression: sample size considerations. *J. Clin. Epidemiol.* **79**, 112–119 (2016). <https://doi.org/10.1016/j.jclinepi.2016.05.014>
- J. Hoeting, D. Madigan, A. Raftery, C. Volinsky, Bayesian model averaging: a tutorial (with discussion). *Stat. Sci.* **14**(4), 382–417 (1999). Correction: vol. 15, pp. 193–195
- T.M. Hopson, *Operational Flood-Forecasting for Bangladesh*. Ph.D. thesis, University of Colorado, 2005, 225pp
- T.M. Hopson, P.J. Webster, A 1–10-day ensemble forecasting scheme for the major river basins of Bangladesh: forecasting severe floods of 2003–07. *J. Hydrometeorol.* **11**(3), 618–641 (2010)
- F. Hoss, P.S. Fischbeck, Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables. *Hydrol. Earth Syst. Sci.* **19**, 3969–3990 (2015). <https://doi.org/10.5194/hess-19-3969-2015>
- G.T. Knoftczynski, D. Mundfrom, Sample sizes when using multiple linear regression for prediction. *Educ. Psychol. Meas.* **68**(3), 431–442 (2008). <https://doi.org/10.1177/0013164407310131>
- T.N. Krishnamurti, C.M. Kishtawal, T.E. LaRow, D.R. Bachiochi, Z. Zhang, C.E. Williford, ... S. Surendran, Improved weather and seasonal climate forecasts from multimodel superensemble. *Science* **285**(5433), 1548–1550 (1999)
- E.O. Ogundimu, D.G. Altman, G.S. Collins, Adequate sample size for developing prediction models is not simply related to events per variable. *J. Clin. Epidemiol.* **76**, 175–182 (2016). <https://doi.org/10.1016/j.jclinepi.2016.02.031>
- S. Opitz-Stapleton, S. Gangopadhyay, B. Rajagopalan, Generating streamflow forecasts for the Yakima River Basin using large-scale climate predictors. *J. Hydrol.* **341**(3–4), 131–143 (2007). <https://doi.org/10.1016/j.jhydrol.2007.03.024>
- T.C. Pagano, D.C. Garen, T.R. Perkins, P.A. Pasteris, Daily updating of operational statistical seasonal water supply forecasts for the Western U.S. *J. Am. Water Resour. Assoc.* **45**(3), 767–778 (2009). <https://doi.org/10.1111/j.1752-1688.2009.00321.x>
- T.N. Palmer, Predicting uncertainty in forecasts of weather and climate. *Rep. Prog. Phys.* **63**(2), 71 (2000)
- S.K. Regonda, B. Rajagopalan, M. Clark, E. Zagona, A multimodel ensemble forecast framework: application to spring seasonal flows in the Gunnison River Basin. *Water Resour. Res.* **42**(9), 1–14 (2006). <https://doi.org/10.1029/2005WR004653>
- M.-S. Roulston, L.-A. Smith, Combining dynamical and statistical ensembles. *Tellus* **55A**(1), 16–30 (2003)
- D. Rousseau, P. Chapelet, A test of the Monte-Carlo method using the WMO/CAS Intercomparison Project data, in *Report of the Second Session of the CAS Working Group on Short-and Medium-Range Weather Prediction Research*. WMO/TD 91, PSMP Rep. Series 18 (1985), 114pp
- A. Sankarasubramanian, U. Lall, Flood quantiles in a changing climate: seasonal forecasts and causal relations. *Water Resour. Res.* **39**(5), 1134 (2003). <https://doi.org/10.1029/2002WR001593>
- R. Schefzik, T.L. Thorarinsdottir, T. Gneiting, Uncertainty quantification in complex simulation models using ensemble copula coupling. *Statistical Science* **48**, 616–640 (2013)
- F.A. Souza Filho, U. Lall, Seasonal to interannual ensemble streamflow forecasts for Ceará, Brazil: applications of a multivariate, semiparametric algorithm. *Water Resour. Res.* **39**(11), 1307 (2003). <https://doi.org/10.1029/2002WR001373>
- G.A. Tootle, A.K. Singh, T.C. Piechota, I. Farham, Long lead-time forecasting of U.S. streamflow using partial least squares regression. *J. Hydrol. Eng.* **12**, 442–451 (2007)

- J.S. Verkade, J.D. Brown, F. Davids, P. Reggiani, A.H. Weerts, Estimating predictive hydrological uncertainty by dressing deterministic and ensemble forecasts; a comparison, with application to Meuse and Rhine. *J. Hydrol.* **555**, 257–277 (2017). <https://doi.org/10.1016/j.jhydrol.2017.10.024>
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences* (Academic/Elsevier, 2011)
- A.W. Wood, D.P. Lettenmaier, An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophys. Res. Lett.* **35**, L14401 (2008). <https://doi.org/10.1029/2008GL034648>
- A.W. Wood, J.C. Schaake, Correcting errors in streamflow forecast ensemble mean and spread. *J. Hydrometeorol.* **9**, 132–148 (2008)



Short-Range Ensemble Forecast Post-processing

Marie-Amélie Boucher, Emmanuel Roulin, and Vincent Fortin

Contents

1	Introduction	796
2	Pre- or Post-processing?	797
3	Analogs	799
3.1	A General Analog Method	799
3.2	Analog Method for Streamflow Data	801
4	Regression-Based Methods	802
4.1	Quantile Regression	802
4.2	Nhomogeneous Regression	805
4.3	Generalized Linear Models and General Linear Models	806
4.4	Logistic Regression	806
5	Ensemble Dressing	809
5.1	The Best Member Method	810
5.2	Weighted Ensemble Dressing	811
6	Bayesian Model Averaging	812
7	Which One to Choose?	813
8	Summary	815
	References	816

Abstract

Short-term hydrological ensemble forecasts do not usually account for the uncertainty in the initial conditions. Consequently, raw forecasts are often biased and under-dispersed and must be post-processed. Both precipitation and streamflow

M.-A. Boucher (✉)

Civil Engineering Department, Université de Sherbrooke, Sherbrooke, QC, Canada

e-mail: Marie-Amelie.Boucher@USherbrooke.ca

E. Roulin

Institut Royal Météorologique de Belgique, Bruxelles, Belgium

V. Fortin

Environment and Climate Change Canada, Dorval, QC, Canada

forecasts for short lead-time depart from the Gaussian distribution, and this important characteristic limits the choice of possible post-processing approaches. Post-processing is performed by calibrating a statistical model using a training dataset containing past forecasts and the corresponding observations. This chapter covers the most common post-processing approaches for short-term hydrological forecasts. They are divided into four categories: analog methods, regressions, kernel dressing, and Bayesian Model Averaging. The vast majority of post-processing methods can be categorized as regression-based. A selection of the most commonly encountered ones in hydrology is presented: quantile regression, nonhomogeneous regression, and logistic regression. Any post-processing approach brings benefits and drawbacks, which are discussed at the end of this chapter. However, according to the few existing comparative studies, no single method is appropriate for all forecasting situation. Therefore, the reader should make his or her own mind regarding which one to choose, according to his or her own specific needs and limitations.

Keywords

Analogs · Kernel dressing · Bayesian Model Averaging · Logistic regression · Quantile regression · Generalized linear models · Nonhomogeneous regression · Preprocessing · Post-processing · Short-range ensemble forecasts

1 Introduction

Statistical post-processing is relevant for hydrometeorological forecasts of all lead times. Most often, the uncertainty in the initial conditions is ignored in hydrologic modelling. This results in biased and under-dispersed raw forecasts in the early time steps, where the initial conditions and effects of hydrologic model structural uncertainty are most important (Zhao et al. 2011). Therefore, post-processing of short-term ensemble forecasts is of utmost importance to ensure the end user of reliable and sharp predictive distribution. In fact, according to Broecker and Smith (2008) the process of fitting a probability density function to the raw ensemble members, even when it does not involve any correction for bias or spread, is still post-processing of the ensemble.

Post-processing is performed by calibrating a statistical model using a training dataset containing past forecasts (and/or past hydrological simulations) and the corresponding observations. The variables forecasted with the numerical weather prediction or the hydrologic model which are selected for the statistical model are the *predictors*, and the variables for which a corrected prediction is sought are the *predictands*.

Both precipitation and streamflow have distributions which depart from the Gaussian and so do the distributions of their forecast errors. Precipitation is characterized by a probability of no precipitation which is positive and a distribution of the quantitative precipitation given that precipitation occurred which is highly skewed. This difficulty is overcome either by a power transformation of the variable in order

to mitigate the skewness or by explicitly modelling precipitation as a mixture of a probability of precipitation and a distribution with appropriate shape. Streamflow is also skewed toward positive values, and simulation or forecast errors are heteroscedastic. There are a variety of techniques to deal with these features like the simple logarithmic transformation, the power transformation, the Box-Cox transformation, the normal quantile transform (NQT), or the log-sinh transformation.

The purpose of post-processing, whether for long-, medium-, or short-range ensemble forecasts, is to address different sources of uncertainties, like correcting the bias, improving the consistency of the ensembles, or assessing the total uncertainty. The estimates of those uncertainties are based on past hydrological simulations and observations, hydrological hindcasts, or reforecasts. It is also very important that a post-processing method should preserve any existing skill in the raw ensemble.

There exist a plethora of methods for post-processing short-term ensemble forecasts. Some of them are more common, some are very new, and some are variants of others. It is not possible to cover such a wide topic exhaustively, so choices had to be made, and only selected methods are described in this chapter. First, only the most well-known and established methods were selected. Analog methods, various regressions, kernel dressing, and Bayesian Model Averaging easily satisfy that requirement. Then, as some of the subcategories themselves comprise many variants, an effort was made to present only the most common or classic ones, rather than to try and find the latest hot of the press method. For this reason, the analog method is described in its most classical form, and only two variants of kernel dressing are presented although there exist many more. Lastly, although there also exist many possibilities to statistically post-process deterministic forecasts into deterministic forecasts, this chapter emphasizes post-processing approaches for raw ensembles. The rationale for this choice is rather partisan. Meteorological ensemble forecasts are now widely available to hydrologists and undergo constant quality improvements. The authors feel that ensemble products should now be used preferably to deterministic products whenever possible, even if ensembles can be statistically built from deterministic forecasts.

This chapter first begins with a question: Is it preferable to preprocess meteorological forecasts or to post-process hydrological forecasts? The presentation of analog methods follows in Sect. 3 and then regression methods in Sect. 4. This regroups the oldest forms of post-processing. Sect. 5 presents the more recent (but very intuitive) kernel dressing approach, and Sect. 6 describes the popular Bayesian Model Averaging.

2 Pre- or Post-processing?

Ensemble streamflow forecasting can take many forms, depending on which sources of uncertainty are to be accounted for in the process. However, one of the most frequent approaches is to feed a hydrologic model with meteorological ensemble forecasts, either from one particular atmospheric model or from a combination of many such models. This methodology accounts for the uncertainty related to the

atmospheric conditions and also for the uncertainty related to the particular choice of structure for the atmospheric model in the case where many of them are involved.

Raw outputs from atmospheric models are often found to be biased and unreliable. Therefore, it might be advisable to preprocess them before using them for hydrologic modelling. After all, if the inputs to the model are wrong, the outputs will also be wrong for sure, hence the popular saying “garbage in, garbage out.” But the hydrologic model can also add biases to the forecast. Consequently, chances are that even if the meteorological forecasts were made bias-free and perfectly reliable, the final outputs (streamflow forecasts) would not be and might need post-processing anyway. In addition, post-processing streamflow ensembles is typically much easier than preprocessing meteorological ensembles. Preprocessing temperature and precipitation forecasts separately for each lead time and each grid point results in an inadequate representation of their space-time covariability (see, for instance, Verkade et al. 2007). Since this space-time covariability of the meteorological forcing is essential for adequate hydrologic modelling, it must absolutely be accounted for, and this considerably complicates the task of preprocessing temperature and precipitation forecasts. One possible solution is to preprocess each variable separately and then to correct for space-time correlation. A possibility for such correction is the “Schaake shuffle” (Clark et al. 2004), a process by which the rank of each ensemble member for both precipitation and temperature is simultaneously reordered using historical meteorological observations for the same start date as the current ensemble forecast. However, as pointed out by Verkade et al. (2007), the Schaake shuffle does not capture the space-time covariability conditional on the actual state of the atmosphere, and the resulting streamflow ensemble forecasts could still benefit from post-processing themselves. Verkade et al. (2007) also remark that preprocessing temperature ensembles alone results in greater improvement than preprocessing the precipitation ensembles alone.

Scheffzík et al. (2013) note that, frequently, statistical post-processing techniques apply to each weather variable at each location and each lead time individually and may fail to take cross-variable, spatial, and temporal correlations properly into account. They point out that NWP models rely on discretizations of the equations that govern the physics of the atmosphere and this multivariate dependence structure tends to be reasonably well represented in the raw ensemble system. Applications like flood management depend on physically realistic probabilistic forecasts of spatiotemporal weather trajectories with much higher dimensions than can be tackled with parametric models. They propose a general procedure which they call ensemble copula coupling (ECC). First, apply the post-processing to the raw ensemble to obtain calibrated and sharp marginal predictive distributions for each weather variable, location, and lead time. Then, draw a discrete sample from each univariate, post-processed predictive distribution. Finally, arrange the sampled values in the rank order structure of the raw ensembles. They show that the ECC approach can be considered an empirical copula technique.

According to a comparative study by Kang et al. (2010), post-processing the ensemble streamflow forecasts is more effective than preprocessing meteorological inputs in terms of quality improvement. This is also in agreement with the findings

by Verkade et al. (2007), who concluded that preprocessing the precipitation and temperature forecasts resulted only in modest improvement in the streamflow forecasts' quality. Finally, it is worth noting that the National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) developed an end-to-end Hydrologic Ensemble Forecast Service (HEFS) which comprises both a preprocessor for meteorological ensemble forecasts and a post-processor for streamflow ensemble forecasts. This very complete system is thoroughly described in Demargne et al. (2014).

3 **Analogs**

Analogs are among the oldest ensemble generation techniques. Early discussion about analogs for weather forecasting can be traced back to Lorenz (1969) and later in Glahn and Lowry (1972), Carter et al. (1989), and Van Den Dool (1994). According to Lorenz (1969), *analog* is a certain state of the atmosphere which resembles another state very closely.

Analogs can be useful either to post-process raw ensemble forecasts or as an ensemble generation technique starting from deterministic forecasts. In this sense, it is an intuitively appealing solution in cases where only deterministic forecasts are available or when the only “real” ensemble forecasts available are judged unsatisfactory.

The basic idea is to use reforecasts and the corresponding observations to find values that are very similar (analog) to the current raw forecast. The observations corresponding to each analog forecast then constitute the post-processed ensemble forecast. It is also possible to fit a probability distribution to the series of past analog observations in order to get a full predictive distribution instead of just members. A variant consists in using the series of past errors between reforecasts and observations to dress either a deterministic forecast or each member of the raw ensemble.

The success of analog methods depends heavily on the availability of a sufficiently long set of reforecasts and on the assumption that the climate remains stable throughout the period covered by the reforecast data. The length of the reforecast dataset is especially important for accurate estimation of the probability of rare or unusual events: it has to be sufficiently large so as to contain a large amount of similar unusual situations. Therefore, post-processing with analogs is inappropriate if only small sample datasets are available. Unfortunately, reforecasts are computationally expensive, so in reality, the length of the reforecast dataset can be very limited, when reforecasts are available at all.

3.1 A General Analog Method

The goal of ensemble forecasting is to find the best possible estimate of the probability density function of a variable y given the ensemble forecast \mathbf{X} for a particular time and location:

$$f(y, |\mathbf{X}) \quad (1)$$

If the climate can reasonably be considered as stable and if a sufficiently long set of reforecasts are available, then it is possible to find a set of past forecasts that are nearly identical (analog) to the current forecast. The observed states for those analogs should correspond to all plausible outcomes of the current forecast, since they should represent the same atmospheric conditions. Therefore, Eq. 1 can be replaced by the distribution formed of the observations corresponding to the analog reforecasts.

In the most general case, both forecasts and observations can be multi-dimensional. Typical meteorological ensemble forecasting systems comprise between 15 and 50 members, for a multitude of variables at the same time, for different levels in the atmosphere on a grid covering the entire earth. For short-term hydrological forecasting, a regional atmospheric model is usually involved. This regional model, however, needs to be constrained at its boundaries by the outputs from the global model. Because of the very large dimensionality of the problem, simplifying assumptions are necessary. Otherwise, analogs would often be impossible to identify. It is possible, for instance, to limit the spatial domain of research to a specific area of interest rather than considering the complete state of the atmosphere over the planet. It could also be advisable to search for analogs of the ensemble mean rather than for each ensemble member. In fact, Hamill and Whitaker (2006) showed that using analogs for the ensemble mean led to post-processed ensembles that were superior in terms of performance than those obtained by matching each ensemble member with its own analogs. Finally, even though the complete state of the atmosphere is best described using all the variables at once, if one is only interested in precipitation, then it is possible to concentrate on finding analogs for precipitation forecasts regardless of the value of the other atmospheric variables.

Figure 1 reproduced from Hamill and Whitaker (2006) describes the method for a deterministic forecast (or for the ensemble mean) and corresponding observation. The observation (left y-axis) is plotted against the corresponding past reforecast for different categories of events. Here 12 groups of analog forecasts are used with a bin width of 0.5. It means that, for instance, reforecasted values between -3 and -2.5 are considered analog, and they belong to the same group (bin).

From Fig. 1, it is possible to compute the probability that the observed value be superior to a certain threshold q . Suppose that we are interested in forecasting the probability that the observed value will be greater than 0. The probability $p(y > q)$ can be computed from

$$p(y > q) = \frac{1}{N} \sum_{i=1}^N I[y^{obs|r}(i), q] \quad (2)$$

where N is the number of analogs in each bin, $y^{obs|r}(i)$ are the past observations corresponding to each analog reforecast (r) in bin i , and $I[y^{obs|r}(i), q]$ is the identity function, equal to 1 if $y^{obs|r}(i) > q$ and 0 otherwise. This amounts to counting the number of observations above the threshold in each analog category (bin). The

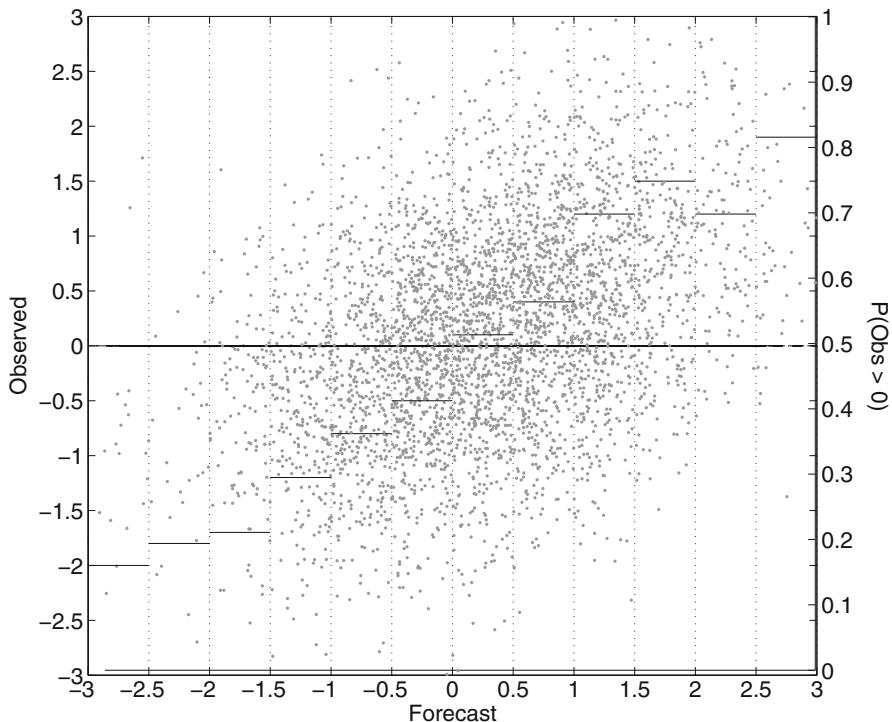


Fig. 1 Illustration of the analog method. (Modified from Hamill and Whitaker (2006))

horizontal bars in Fig. 1 show the value of $p(y > q)$ for each bin, and the right-hand y -axis shows the corresponding probabilities.

If it is desired to correct each forecast member separately (instead of using the ensemble mean to find analogs) and we are interested in a wide range of different thresholds q , then the process illustrated in Fig. 1 and Eq. 2 must be repeated for each member and each probability threshold. This emphasizes the importance of the previously mentioned simplifying assumptions.

Finally, it is worth noting that if the observation (or “analyzed weather”) has a higher spatial resolution than the raw forecasts, then the analog approach also satisfies the purpose of downscaling since the ensemble is formed of past observations rather than forecasts.

3.2 Analog Method for Streamflow Data

Typically, in the framework of streamflow forecasting, analog methods are applied to precipitation forecasts which are then passed on to a hydrologic model. It is also possible, however, to apply the analog method directly to streamflow forecasts.

In fact, “the regime-dependance of hydrologic forecast errors (e.g., differing in the rising or receding hydrograph limbs) provide a strong rationale for investigating analog-based forecast calibration approaches” (Wood 2012). When using other post-processing approaches for streamflow, one often has to separate the database in order to account for the variety of different regimes and calibrate post-processing parameters accordingly. In this view, an analog method can be simpler than many other post-processing approaches for streamflow.

4 Regression-Based Methods

A linear regression between each raw ensemble member and the observation can be applied to obtain the post-processed ensemble. With analog-based approaches, it is among the oldest post-processing methods and certainly the most classical definition of “Model Output Statistics” (MOS; Glahn and Lowry 1972; Wilks 2011). Regressions can be applied either to meteorological forcings to the hydrologic model (precipitation and temperature) or to the ensemble of streamflow forecast itself.

If $\mathbf{X} = X_1, \dots, X_K$ denote an ensemble of individually distinguishable forecast members for a univariate quantity y , such as precipitation or streamflow, this quantity can be expressed as a multiple linear regression between the ensemble members and the observation, involving regression coefficients b_0 and $b_1 \dots b_k$ that must be calibrated.

$$y = b_0 + b_1 X_1 + \dots + b_k X_K + \varepsilon \quad (3)$$

ε is a random error term that averages to zero. Consequently, the predictive variance is independent of the ensemble spread, so the resulting uncertainty assessment is static.

4.1 Quantile Regression

Quantile regression was developed by two econometricians (Koenker and Bassett 1978). Generally speaking, the method aims at estimating a relation between predictor variables and a response variable, for all portions of the probability distribution of the response variable. The best way to understand the method is by comparing it to a very usual linear regression which aims at estimating a relation f between a response variable y and a predictor variable X :

$$y = f(X) \quad (4)$$

For a finite sample of the predictive and response variables, the problem of estimating y using a linear regression and a deterministic forecast X is given by

$$\hat{y} = b_0 + b(X) \quad (5)$$

where \hat{y} is an estimate of y and b_0 and b are, respectively, the intercept and the slope of the regression. The residuals are the differences between the observation y and the corresponding estimation \hat{y} , as given by Eq. 6:

$$e_t = y_t - \hat{y}_t \quad (6)$$

The regression parameters b_0 and b can be estimated with the least squares method, which amounts to minimizing the value of the sum of squared residuals:

$$\min \sum_{t=1}^T (y_t - (b_0 + bX_t))^2 = \min \sum_{t=1}^T (y_t - \hat{y})^2 = \min \sum_{t=1}^T e_t^2 \quad (7)$$

The regression line obtained by estimating a and b in this fashion always goes through the mean (\bar{X}, \bar{y}) , so it defines the *expected* or *mean* value of y conditional on X .

In a similar fashion, the median can be defined as the solution to a problem of minimizing the sum of *absolute* residuals. The minimization of the sum of absolute residuals implies that there should be the same number of residuals above and below the regression line, which corresponds well to the definition of the median. Quantile regression is simply a generalization to quantiles other than the median.

For the post-processing of ensemble forecasts, the τ quantile of case t can be related to K predictors \mathbf{X}_t . Equation 5 becomes

$$q_\tau(\mathbf{X}_t; \mathbf{b}) = b_0 + \sum_{k=1}^K b_k X_{tk} \quad (8)$$

where the parameters $\mathbf{b} = b_0, b_1, \dots, b_K$ are estimated by minimizing the least absolute deviation (LAD) function:

$$\mathbf{b} = \arg \min_{\mathbf{b}} \sum_{t=1}^T \omega_\tau(y_t - q_\tau(\mathbf{X}_t; \mathbf{b})), \quad (9)$$

where y_t is the observed value in a calibration dataset of size T .

It is possible to define regressions for any quantile by assigning different weights to positive and negative residuals using the check function ω_τ :

$$\omega_\tau(u) = \begin{cases} \tau u & \text{if } u \geq 0 \\ (\tau - 1)u & \text{if } u < 0 \end{cases} \quad (10)$$

Bremnes (2004) suggests a two-stage procedure for precipitation. First, the probability of precipitation is forecasted using a probit regression. (The probit regression is, as the logistic regression, an example of generalized linear model (see Sect. 4.3). Here, the linear model is related to the response variable (precipitation or no precipitation) via the inverse Gaussian distribution.) Second, conditional on the

occurrence of precipitation, a range of quantiles are obtained with separate regression equations. The final probabilities are obtained through the multiplicative law of probability. Friederichs and Hense (2007) present an alternative approach where precipitation is represented by a censored variable with zero precipitation as the censoring line. They use the LAD function as a proper score for the quantile forecasts.

Equation 8 can be applied directly to deterministic forecasts to post-process them into predictive distributions (e.g., see Weerts et al. 2011). The method can also be applied to raw uncalibrated ensemble forecasts in various ways. Bremnes (2004) compared four different approaches for quantile regression of short-range ensembles of precipitation forecasts. The first one was identical to Weerts et al. (2011), using the deterministic output from a high resolution model, while the second and third approaches involved ensemble forecasts. In the second approach, the author applied quantile regression to each ensemble member separately. For the third and fourth approaches, instead of applying Eq. 8 directly to ensemble members, they used two types of statistics computed on raw ensemble members: either a series of raw quantiles (5th, 10th, 25th, 50th, 75th, 90th, and 95th) or the minimum and the maximum value of the raw ensemble. According to the results of Bremnes (2004), the most promising approach appears to be the use of raw quantiles as the X in Eq. 8.

As for software, the best option is probably the *quantreg* package for R (Koenker 2018), which is freely available. Popular commercial statistics/econometric softwares also include a quantile regression package (SAS, Stata, Shazam, Limdep, etc.).

The approach of applying the linear MOS post-processing technique developed for single forecasts to the ensemble directly is discarded because of the tendency to converge to the climatological mean. Vannitsem (2009) also notes that the assumption in MOS that the predictors (the model observables) are error-free is unrealistic. By appropriately defining a cost function and assuming the errors are Gaussian, he obtains parameters that allow the corrected forecasts to keep the variability of the reference variable during the whole forecast. Vannitsem (2009) introduces this new scheme as error-in-variables MOS (EVMOS). This new post-processing technique is well suited for ensemble predictions since the variability among the members is not removed.

The variance inflation method has been first proposed in the context of seasonal forecasting to adjust the forecasts so that, for a reliable ensemble, the climatological variance of the forecasts is the same as of the truth (e.g., Wood and Schaake 2008). Johnson and Bowler (2009) show that this method satisfies another requirement for a reliable ensemble that the ensemble spread is, on average, representative of the uncertainty in the mean. With this method, the correlation of the ensemble members with the ensemble mean is the same as the correlation of the truth with the ensemble mean and that the mean square error (MSE) of the ensemble mean is minimized. Applied to ensemble members (Johnson and Bowler 2009), the method adjusts both the ensemble mean, \bar{f}_t and the perturbation to the mean, e_t^i giving new members:

$$g_t^i = \alpha \bar{f}_t + \beta e_t^i \quad (11)$$

With

$$\alpha = \rho_{xf} \frac{\sigma_x}{\sigma_f} \quad (12a)$$

$$\beta = \sqrt{\left(1 - \rho_{xf}^2\right)} \frac{\sigma_x}{\sigma_e} \quad (12b)$$

where ρ_{xf} is the correlation of the observation with the ensemble mean, σ_x is the standard deviation of the observations, σ_f is the standard deviation of the ensemble mean, and σ_e is the average ensemble standard deviation. A generalization to many predictors is proposed in Van Schaeybroeck and Vannitsem (2014).

The EVMOS and inflation methods have been adapted for the post-processing of ensemble predictions of streamflow by Roulin and Vannitsem (2014).

4.2 Nonhomogeneous Regression

The nonhomogeneous Gaussian regression (NGR) was proposed by Gneiting et al. (2005). This approach is an improvement over the classical linear regression, in which the residuals are considered to be normally distributed and their variance is a function of the ensemble's variance, so ε in Eq. 3 becomes $\varepsilon(t)$. This reflects the fact that ensemble forecasts with a large spread exhibit a large variability in their errors (spread-skill relationship), and this characteristic should be preserved during the post-processing. The corrected ensemble members are given by Eq. 13:

$$\hat{y}(t) = a(t) + b(t)\bar{X}(t) + \varepsilon(t) \quad (13)$$

with the residuals given by

$$\varepsilon(t) \sim N\left[0, c + d\sigma_{ens,t}^2\right] \quad (14)$$

where $\bar{X}(t)$ and $\sigma_{ens,t}^2$ are, respectively, the ensemble mean and variance. The regression parameters a , b , c , and d are estimated by minimizing the Continuous Ranked Probability Score (CRPS, see chapter ▶ “Verification Metrics for Hydrological Ensemble Forecasts”, Sect. 2) for the calibration portion of the data.

Scheuerer (2014) adapted the nonhomogeneous Gaussian regression to the characteristics of precipitation. He specifies the generalized extreme value (GEV) distribution to be a suitable model. GEV is characterized by a location, a scale, and a shape parameter. He considers the GEV to be left-censored at zero, i.e., all mass below zero is assigned to exactly zero. Then, he links the parameters to the Model Output Statistics. As predictors, he uses the ensemble mean, the fraction of zero precipitation members, and the ensemble mean difference. The model is fitted through minimizing the CRPS.

4.3 Generalized Linear Models and General Linear Models

Generalized linear models (Nelder and Wedderburn 1972) represent a very broad category of models, and their application is far from restricted to ensemble forecast post-processing. They are not to be confused with general linear models, despite having the very same acronym. In fact, a general linear models can be viewed as a special case of generalized linear models. Equation 3 represents a general linear model, similar to the one used by Zhao et al. (2011) to post-process streamflow forecasts from three different hydrologic models. Such classic linear model rests on important assumptions. Each component X_1, \dots, X_K is independent and normally distributed. Moreover, each component can have a different mean, but they all have the same variance. The relationship between the random and systematic components is specified via the identity function.

Clearly, in hydrometeorological forecasting, the assumption of normality is often violated. In addition, there is no guarantee that the ensemble members (the X_1, \dots, X_K) have the same variance. Consequently, one has to transform the variables beforehand so that these requirements are met. Generalized linear models, however, assume that the component X_1, \dots, X_K are independent and belong to an exponential family of distributions. The normal (Gaussian) distribution, binomial, Poisson, and Gamma are examples of exponential families. Also, the relationship between the random and systematic components is specified by a differentiable and monotonic link function g , such that $E[y] = \mu = g^{-1}\eta$, where η is the linear combination of the predictors given by $\eta = b_0 + b_1X_1, \dots, b_KX_K$. In subsequent work, GLMs have been extended to multivariate exponential families, to certain non-exponential families, and even to situations where the distribution of y is not completely specified.

Rather than provide the reader with a description of the different variants and possibilities for generalized linear models, it was chosen to explain the general definition, mostly to clarify the distinction between generalized linear models and general linear models which can both be used for post-processing ensemble forecasts. The following section explains with greater detail the logistic regression, which is a special case of generalized linear models and is especially relevant for hydrologists as it has been used with great success in many studies.

4.4 Logistic Regression

As previously mentioned, statistical methods to derive probabilistic quantitative precipitation forecast from the outputs of numerical weather prediction models were first based on linear relationships between selected predictors (Glahn and Lowry 1972). The approach of ordinary multiple regression when the predictand is binary is called regression estimation of event probabilities. When the values estimated with the regression estimation of event probabilities are lower than zero or larger than unity, they are truncated to these values or included in broader categories. The logistic regression provides an alternative to the regression estimation of event probabilities with the advantage that probability estimates are constrained between

zero and unity. For instance, the probability that the observed precipitation y does not exceed a threshold q is related to the predictors \mathbf{X} :

$$p = \Pr(y \leq q) = \frac{\exp(f(\mathbf{X}))}{1 + \exp(f(\mathbf{X}))} = \frac{1}{1 + \exp(-f(\mathbf{X}))}, \quad (15)$$

with

$$f(\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K, \quad (16)$$

where $\boldsymbol{\beta}$ is the vector of $K + 1$ parameters to fit. (Alternatively, the logistic regression may read: $\ln(p/(1 - p)) = \text{logit}(p) = f(\mathbf{x})$). The logistic regression is a regression model from the generalized linear model framework in which the linear part is related to the response (here the probability) through the logit link.) These $K + 1$ parameters can be estimated on a calibration dataset containing T observations of y_t and the corresponding predictors \mathbf{X}_t , using the method of maximum likelihood (Wilks 2011). Binary events follow a Bernoulli distribution. The likelihood of the t th observation is

$$l(\boldsymbol{\beta}) = \begin{cases} p = \frac{\exp(f(\boldsymbol{\beta}, \mathbf{X}_t))}{1 + \exp(f(\boldsymbol{\beta}, \mathbf{X}_t))} & \text{if } y_t \leq q, \\ (1 - p) = \frac{1}{1 + \exp(f(\boldsymbol{\beta}, \mathbf{X}_t))} & \text{if } y_t > q. \end{cases} \quad (17)$$

Setting $o_t = 1$ if the event occurred and $o_t = 0$ otherwise, the log-likelihood to be maximized is

$$L(\boldsymbol{\beta}) = \sum_{t=1}^N \{o_t f(\boldsymbol{\beta}, \mathbf{X}_t) - \ln[1 + \exp(f(\boldsymbol{\beta}, \mathbf{X}_t))]\} \quad (18)$$

The usefulness for the post-processing precipitation from numerical weather prediction outputs has been demonstrated, for instance, by Applequist et al. (2002). Wilks (2009) noted that fitting a logistic regression for a set of thresholds implies that a large number of parameters have to be estimated, that probabilities at intermediate thresholds have to be interpolated, and that separate equations for different thresholds may lead to probability forecasts which are inconsistent with each other. Therefore, he introduced the extended logistic regression by which the logistic regression is estimated once for all thresholds by including the threshold q itself as additional predictor:

$$p(q) = \frac{1}{1 + \exp(-f(\mathbf{X}) - g(q))} \quad (19)$$

where $g(q)$ is a non-decreasing function of the threshold. In his analyses, Wilks (2009) used

$$f(\mathbf{X}) = \beta_0 + \beta_1 \sqrt{\bar{X}} \quad (20a)$$

$$g(q) = \beta_2 \sqrt{q} \quad (20b)$$

where \bar{X} is the ensemble mean. In the context of extended logistic regression, other powers of the threshold (1/4) and of the ensemble mean (1/3, 1/4) have been selected by different authors. The mean of the power transformed ensemble members was used by Schmeits and Kok (2010) and Roulin and Vannitsem (2012), and the product of the power of the ensemble mean and the power of the ensemble variance was included in the set of predictors by Hamill (2012).

Using an experimental reforecast dataset from ECMWF, Schmeits and Kok (2010) have shown that the extended logistic regression compares well with the Bayesian Model Averaging method (BMA, see Sect. 6). The extended logistic regression allows to evaluate the whole distribution. This capacity was exploited by Roulin and Vannitsem (2012) to post-process ECMWF ensemble predictions of precipitation based on the information of operational reforecasts. Since the size of the reforecast ensembles is small ($K = 5$) compared to the size of the raw ensembles ($K = 51$), a correction is applied to the ensemble mean in order to avoid biases in the regression parameters. The temporal structure of individual members is preserved by first ranking the 51 members and assigning to each a probability. The corresponding quantiles are then extracted from the post-processed marginal predictive cumulative distribution function and finally reordered with respect to the raw ensemble values.

Several extensions to the extended logistic regression have been proposed. Ben (2013) introduces interaction terms as the product of the thresholds and the primary predictors. More flexibility is provided, but for each implementation, it has to be checked that the regression functions do not converge within physically relevant ranges. The use of interaction terms is found to significantly improve the sharpness of the calibrated probabilities for high precipitation thresholds.

Messner et al. (2014b) note that ensemble spread used as ordinary predictor of the extended logistic regression only affects the location but not the variance of the predictive distribution. They propose an approach where the ensemble spread is used directly to predict the dispersion of the predictive distribution. They show additional skill of this heteroscedastic variant relative to the standard extended logistic regression for their case study on wind speed with ECMWF ensemble predictions. In Messner et al. (2014a), they compare different variants of the logistic regression including the heteroscedastic censored logistic regression in which the parameters of the logistic distribution are estimated by maximizing a log-likelihood function based on the standard logistic distribution function and not on its cumulative distribution like in Eq. 19. Therefore this method is using all information in the training dataset instead of selected category probabilities. The *crch* for R (Messner and Zeileis 2013) can be used for the heteroscedastic extended and censored logistic regression.

The extended logistic regression is used by Fundel and Zappa (2011) to calibrate hydrological ensemble forecasts. (These ensembles were obtained by forcing their hydrologic model with data from the operational reforecast datasets

for the ECMWF ensemble predictions.) They selected the square root of the (streamflow) ensemble mean, the fourth root of the ensemble spread, and the streamflow threshold as predictors of the regression. This post-processing is shown to significantly improve the reliability term of the Brier skill score (see chapter ??, section ??) for the 80th quantile.

5 Ensemble Dressing

Ensemble (or kernel) dressing is intuitive, is easy to implement, and requires short computation time. While it is described here as a post-processing tool for ensemble forecasts, it is more broadly known as a nonparametric distribution-fitting tool and often called “kernel smoothing” in this context.

The basic idea is to fit probability distribution function (the kernel) around each raw ensemble member and next forming a mixture distribution by summing all the kernels to obtain the post-processed probability density function (*pdf*). The post-processed ensemble members can be drawn from this post-processed *pdf*. Each kernel is defined by its shape and by its bandwidth. As it will be described later, it is also possible to weight the kernels to reflect their respective credibility in the case where ensemble members are differentiable. The shape of the kernel depends on the choice of the particular *pdf* that is chosen for the kernels. The most popular choice is the normal kernel, but almost any *pdf* would be possible. The only important restriction for forecasts of hydrological variables such as streamflow or volume is that the kernel must have a strictly positive support. However, it is generally easy to implement this restriction even for normal kernels, for instance, by using a logarithmic transform of the raw members before the dressing and then transforming back the dressed ensemble. According to Wand and Jones (1995), the choice of a particular *pdf* for the kernel is not critical for the success of the method. The determination of the bandwidth is a much more important factor for the success of the method, and the difference between most ensemble dressing post-processing methods lies in the way the bandwidth(s) is or are estimated.

Ensemble dressing is best suited to address under-dispersion problems. Although some ensemble dressing methods can address unconditional bias up to some point (see Sect. 5.2), they aim mostly at correcting the spread of the ensemble. Therefore, the raw ensemble should be bias corrected first. Finally, ensemble dressing has one major drawback: it does not preserve the temporal correlation of the forecast series. Although the processed ensemble envelope (the upper and lower limits of the distribution) accurately follows the fluctuations of the observed streamflow, the individual post-processed members exhibit random fluctuations instead of following a natural temporal pattern with temporal correlation from day to day. For applications where the shape of the hydrograph for each member is important, ensemble dressing methods might not be appropriate.

The most well-known ensemble dressing method is the best member method proposed by Roulston and Smith (2003). It is therefore first described and followed by an alternative method as an example on how this basic method can be improved.

5.1 The Best Member Method

The name of the best member method is self-explicating: the bandwidth of the kernel is estimated using the errors made by the ensemble members that are closest to the corresponding observations for each time step. If the members of the raw ensemble are not differentiable, no member is the best member for all time steps. The best member (closest to the observation) may therefore differ from one time step to the other.

Using a calibration dataset for which both the raw ensemble forecasts and the corresponding observations are known, the first step in implementing the best member method consists in computing the absolute difference between each ensemble member ($X_{t,k}$) and the observation (y_t), for each time step t . The index $k = 1 \dots K$ refers to the number of members. The smallest absolute error for a particular time step constitutes the “best member’s error” (ξ) for this time step, as described by Eq. 21:

$$\xi_t = X_t^* - y_t = \min |X_{t,k} - y_t| \quad (21)$$

where X_t^* is the best member. Once the best member errors are computed for all time steps in the calibration dataset, the variance of those errors is calculated and is considered the bandwidth σ_K of the kernels for the dressing. Typically, a Gaussian kernel is assumed, in which case each kernel \mathcal{K} is described by Eq. 22:

$$\kappa(X - X_i, \sigma_K) = \frac{1}{\sqrt{2\pi}\sigma_K} \exp\left\{-\frac{(X - X_i)^2}{2\sigma_K^2}\right\} \quad (22)$$

The post-processed *pdf* is then obtained by summing all the kernels in a probability density mixture, as given by Eq. 23:

$$F(y_t | X_{t,1}, \dots, X_{t,K}) = \frac{1}{N} \sum_{i=1}^N \mathcal{K}(X - X_i, \sigma_K) \quad (23)$$

with $F(y_t | X_{t,1}, \dots, X_{t,K})$ as the post-processed probabilistic forecasts based on K raw ensemble members.

Equivalently, the post-processed ensemble can be obtained more directly in a nonparametric fashion by adding these errors to each raw ensemble member:

$$\hat{\mathbf{y}}_k = \mathbf{X}_k + \boldsymbol{\xi} \quad k = 1, \dots, K \quad (24)$$

where $\hat{\mathbf{y}}_k = (\hat{y}_{1,k}, \dots, \hat{y}_{T,k})'$ is the post-processed ensemble and $\mathbf{X}_k = (X_{1,k}, \dots, X_{T,k})'$ is the original (raw) ensemble. $\boldsymbol{\xi} = (\xi_1, \dots, \xi_T)'$ is a vector containing all the best member errors (ξ_t). Ensemble dressing in this way adds additional members around each original (raw) ensemble member. The latter are called “dynamical” ensemble members, and the former are known as “statistical” members, since they were not part of the initial raw ensemble but added through post-processing.

5.2 Weighted Ensemble Dressing

The best member method can only correct for under-dispersion and does not correct unconditional bias. From this starting point, more sophisticated methods can be built. For instance, each member of the raw ensemble can be weighted differently according to its rank, and a specific bandwidth can also be estimated for each rank (Fortin et al. 2006). The weighted ensemble dressing method is more generally applicable, since it allows the improvement of both under-dispersed and over-dispersed ensemble forecasts. It also provides a better estimation of extreme event probabilities. As with the best member method described in Sect. 5.1, the best member must be identified at each time step. The rank of the best member among the sorted ensemble is also identified, and errors are stored accordingly. For example, for all time steps t at which the best ensemble member occupies the second rank among ordered ensemble members, the corresponding error will be stored in a vector corresponding to this case. There are thus N best member error vectors for an ensemble containing N members, as described by Eq. 25:

$$\xi_{t,(k)} = \{ |X_t^* - y_t| \mid X_t^* = X_{t,(k)}, t = 1, 2, \dots, T \} \quad (25)$$

where $\xi_{t,(k)}$ is the best member error at time step t , when the best member occupies rank k (out of K) of the ensemble. $X_{t,(k)}$ is the k th sorted ensemble member at time step t . It follows that this dressing method employs a different bandwidth for each member of the ensemble. The post-processed predictive distribution is given by

$$\hat{\mathbf{y}}_{t,k} = \mathbf{X}_{(k)} + \omega_{(k)} \xi_{(k)} \quad k = 1, \dots, K \quad (26)$$

where $\mathbf{X}_{(i)} = (X_{1,(i)}, \dots, X_{T,(i)})'$ and $\xi_{(i)} = (\xi_{1,(i)}, \dots, \xi_{T,(i)})'$ denote, respectively, the vector containing the best member for rank (i) and the best member's errors for this rank. The weight $\omega_{(i)}$ represents the probability that the i th sorted ensemble member is the best member. Although $\omega_{(i)}$ could be estimated as the proportion of occurrence when the best member occupies rank i , a more robust estimator is the beta density function given by the following expression:

$$f_B(X|\omega, \tau) = \frac{x^{\omega\tau-1}(1-X)^{\omega-\omega\tau-1}}{B(\omega\tau, \omega - \omega\tau)}, \quad \omega > 0, \quad 0 \leq \tau \leq 1 \quad (27)$$

The beta function is defined by two parameters, $\alpha = \omega\tau$ and $\beta = \omega - \omega\tau$, with τ the expectation of a random variable following a beta distribution. The parameter τ can be constrained to

$$\tau = \frac{1}{K} \sum_{k=1}^K kp_k - \frac{1}{2K} \quad (28)$$

with p_k representing the observed frequency of the best member occupying rank k . Once the value of τ is known, parameter ω is determined by minimizing the difference between the post-processed ensemble variance s_y^2 and the variance of the observations s_y^2 .

6 Bayesian Model Averaging

Bayesian Model Averaging (BMA) is a popular statistical technique for merging outputs from multiple models into a single probabilistic forecast which can take into account the skill, bias, and uncertainty of each individual model (Hoeting et al. 1999). The BMA framework is flexible, and its implementation can be very different depending on the context. In all cases, BMA relies on the hypothesis that among a (finite) set of models available for issuing a forecast, one of them would be sufficient (in other words “correct”) for each valid time. The difficulty is that we do know which one it is. Let y be the observable quantity to forecast, and let $\mathbf{X} = X_1, X_2, \dots, X_K$ be the outputs from models $\mathbf{M} = M_1, M_2, \dots, M_K$, respectively. Let $p(y|\mathbf{X})$ be a probabilistic forecast for y given the current forecasts \mathbf{X} . In practice, this probabilistic forecast would also be conditional on a training dataset (y^T, \mathbf{X}^T) , which is omitted in the following equations. The hypothesis is that if it was possible to know which model M was sufficient for issuing the forecast, the outputs from the other models could be ignored:

$$p(y|M = M_k, \mathbf{X}) = p(y|M = M_k, X_k) = g_k(y|X_k) \quad (29)$$

where $M = M_k$ signifies that model M_k is correct and $g_M(y|X_k)$ is a probabilistic forecast for y built only from the outputs of model M_k . This is a strong hypothesis of conditional independence which might not be verified in practice. It is however very convenient because it makes it possible to obtain the *pdf* of y as a weighted sum of predictive distributions. Indeed, from the law of total probability:

$$p(y|\mathbf{X}) = \sum_{k=1}^K p(y|M = M_k, \mathbf{X})p(M = M_k, \mathbf{X}) \quad (30)$$

Combining 29 and 30,

$$p(y|\mathbf{X}) = \sum_{k=1}^K g_k(y|X_k)p(M = M_k, \mathbf{X}) \quad (31)$$

where $p(M = M_k, \mathbf{X})$ corresponds to the probability that model M_k be the correct model, prior to observing y . Often, the forecast information \mathbf{X} provided by the various models does not bring information for assessing these probabilities, which then become constants that can be interpreted as model weights:

$$p(y|\mathbf{X}) = \sum_{k=1}^K w_k \cdot g_k(y|X_k) \quad (32)$$

where $w_k = p(M = M_k, \mathbf{X})$ is the weight given to model M_k .

A first context in which BMA has been used is when the models being considered vary greatly in structure and skill so that it becomes very important to identify individual model weights w_k and probabilistic forecasts $g_k(y|X_k)$ based on a training

dataset (Raftery et al. 2005; Sloughter et al. 2007; Wilson et al. 2007). This is possibly the best use of BMA, since the model weights can serve as a diagnostic, helping to understand which modelling system is more useful.

BMA can also be used in the context of exchangeable ensemble members, where each member of an ensemble is obtained, for example, by randomly perturbing initial conditions of a numerical model. By definition, both the model weights w_k and the distribution $g_k(y|X_k)$ are then identical for each model M_k . The BMA predictive distribution then simply becomes

$$p(y|\mathbf{X}) = \frac{1}{K} \sum_{k=1}^K g(y|X_k) \quad (33)$$

where $g(y|X_k)$ is analogous to a kernel smoothing function.

The more complex case where an ensemble of forecast is composed of subsets of exchangeable ensemble members is discussed by Fraley et al. (2010).

Implementation of the method is relatively straightforward using the **ensembleBMA** package for *R* (Fraley et al. 2007).

One of the limitations of the BMA method is that, despite its Bayesian origins, it does not make use of prior information other than the training dataset. In many cases, there exists information on which to base a prior estimate $p(y)$ for y . For example, a climatological prior can generally be obtained. Marty et al. (2014) have generalized the BMA method to include a climatological prior in the model, which is then updated when the information \mathbf{X} becomes available. The framework, only applied for now to the case where the ensemble members are exchangeable, is slightly different: a probabilistic model for X_k given y is first obtained from the training sample and combined with a prior distribution for y using Bayes' rule:

$$g(y|X_k) \propto p(X_k|y)p(y) \quad (34)$$

This posterior distribution for y given X_k is then averaged over all ensemble members:

$$p(y|\mathbf{X}) \propto \frac{1}{K} \sum_{k=1}^K p(X_k|y)p(y) \quad (35)$$

7 Which One to Choose?

No one post-processing approach is perfect, and the choice of an appropriate method highly depends on the particularities of the forecasting situation at hand. A few recent studies have compared the advantages and disadvantages of various post-processing methods. However, no such study involved all the methods presented in this chapter at once.

In addition, up to date, most comparative studies for post-processing approaches were performed in the context of atmospheric sciences, and many of them in a synthetic setting. For instance, Williams et al. (2014) concluded that ensemble dressing, Bayesian Model Averaging, and nonhomogeneous regression perform similarly, while logistic regression performs less well. Their experiment, however, was held in the Lorenz 1996 setting. This was also the case with Wilks (2006), who compared numerous post-processing approaches for short-term precipitation and temperature forecasts. They found the three most promising methods to be the logistic regressions, kernel dressing, and nonhomogeneous regression. Bayesian Model Averaging produced poorer forecasts, and the author attributes this result to an overcorrection of under-dispersion for forecasts of the rarer events. However, this method may have advantages over others for post-processing multi-model ensembles. Given that kernel dressing and nonhomogeneous regression require fewer parameters than most other methods, they represent good choices for a user with restrictions regarding computer time. In addition, according to Wilks (2006), these two methods were the best when only short training samples were available. According to the result of a second experiment (Wilks and Hamill 2007) on the three most promising approaches using real reforecast data, logistic regressions and nonhomogeneous regressions performed better for daily temperature than for precipitation, but there was no single best method for all applications.

Both kernel dressing and nonhomogeneous regression assume a Gaussian error distribution, which could be inappropriate for precipitation and streamflow forecasts. However, recent results by Boucher et al. (2015) on gamma-distributed synthetic data show that this assumption of the error distribution might not be as binding as first thought. Pagano et al. (2013) also used kernel dressing to post-process short-term streamflow forecasts for 128 catchments in southern Australia and found that the raw ensembles could be made reliable.

According to Van Schaeybroeck and Vannitsem (2011), the choice of a particular variant of linear regression method does not affect the skill of the post-processed ensemble for short lead time. The choice of particular predictors and their number has a much stronger influence. Using information from reforecasts for ECMWF ensemble predictions, Roulin and Vannitsem (2014) studied the preprocessing with the extended logistic regression and the post-processing with the error in variable MOS and with the variance inflation. Post-processing alone is found to improve the verification scores of the hydrological ensembles better than preprocessing alone with the variance inflation being able to improve the reliability component of the CRPS. In the case of large biases in the precipitation, combining pre- and post-processing allows for further improvements.

Hopefully, the recent “Post-processing hydrological ensemble prediction inter comparison experiment” van Andel et al. (2012) launched within HEPEX in 2011 will be able to provide more guidelines in the future regarding which post-processing method suits which situation best.

8 Summary

Statistical post-processing of short-term hydrological forecasts is often needed in order to quantify the uncertainty in the initial conditions, since they are not represented directly in the modelling process. According to Broecker and Smith (2008), post-processing of ensemble forecasts refers to the process of fitting a probability density function to the raw ensemble members. However, most techniques also aim at removing systematic biases and adjusting the spread of the ensemble. In this chapter, a selection of the most frequently encountered post-processing approaches for short-term hydrological forecasts is presented. These techniques are divided into analog methods (Sect. 3), regression methods (Sect. 4), kernel dressing methods (Sect. 5), and Bayesian Model Averaging (Sect. 6). The key elements from this chapter are listed below as a summary:

- **Analog methods** use reforecasts and the corresponding observations to identify past forecasting situations that are very similar to the current raw forecast. The observations corresponding to each analog forecast then constitute the post-processed ensemble forecast. Alternatively, the series of past errors between reforecasts and observations can be used to dress the raw ensemble member and obtain the post-processed ensemble.
- **Analog methods** are most often used as *preprocessor* rather than *post-processors*, meaning that the meteorological forcings to the hydrologic model are preprocessed beforehand. However, it is possible to use an analog method directly to post-process streamflow forecasts.
- **Quantile regression** aims at estimating a relation between predictor variables and a predictand variable, for all portions of the probability distribution of the predictand variable. To do so, one uses the check function to weight positive and negative residuals from a regression involving the ensemble members and free parameters that must be calibrated. Those parameters for each quantile are calibrated by minimizing the regression residuals for that particular quantile.
- The **nonhomogeneous regression** is an improvement over the classical linear regression, in which the variance of the residuals is a function of the ensemble's variance. This reflects the fact that ensemble forecasts with a large spread exhibit a large variability in their errors (spread-skill relationship). The regression parameters a , b , c , and d can be estimated by minimizing the CRPS.
- The **logistic regression** is well suited for probabilistic forecasts of binary events as probability estimates are constrained between zero and unity (e.g., probability of precipitation).
- The **extended logistic regression** allows to estimate the whole distribution which is supposed to follow – usually after a power transformation of the variable – a logistic distribution. In assuming a distribution function, this method has similarities with the nonhomogeneous regression.

- In **kernel dressing**, a probability distribution function (the kernel) is fitted around each raw ensemble member. Each kernel is defined by its shape and by its bandwidth (spread). All the kernels are summed to form a mixture distribution which becomes the post-processed probability density function (*pdf*) and from which the post-processed ensemble members can be drawn.
- **Bayesian Model Averaging** is a statistical technique for merging outputs from multiple models into a single probabilistic forecast which can take into account the skill, bias, and uncertainty of each individual model. The BMA relies on the hypothesis that among a (finite) set of models, one is sufficient (or “correct”) for each valid time. Therefore, each model is assigned a weight, and the post-processed predictive distribution is the weighted sum of raw predictive distributions from each individual model.

References

- S. Applequist, G.E. Gahrs, R.L. Pfeffer, X.-F. Niu, Comparison of methodologies for probabilistic quantitative precipitation forecasting. *Weather Forecast.* **17**(4), 783–799 (2002)
- Z. Ben Bouallègue, Calibrated short-range ensemble precipitation forecasts using extended logistic regression with interaction terms. *Weather Forecast.* **28**(2), 515–524 (2013)
- M.-A. Boucher, L. Perreault, F. Anctil, A.-C. Favre, Exploratory analysis of statistical post-processing methods for hydrological ensemble forecasts, a comparative study. *Hydrol. Process.* **29**, 1141–1155 (2015)
- J.B. Bremnes, Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. *Mon. Weather Rev.* **132**, 338–347 (2004)
- J. Broecker, L.A. Smith, From ensemble forecasts to predictive distribution functions. *Tellus Ser A* **60**(4), 663–678 (2008)
- G.M. Carter, J.P. Dallavalle, H.R. Glahn, Statistical forecasts based on the National Meteorological Center’s numerical weather prediction system. *Weather Forecast.* **4**, 401–412 (1989)
- M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The Schaake shuffle: a method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**(1), 243–262 (2004)
- J. Demargne, L. Wu, S.K. Regonda, J.D. Brown, H. Lee, M.X. He, D.J. Seo, R. Hartman, H.D. Herr, M. Fresch, J. Schaake, Y.J. Zhu, The science of NOAA’s operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* **95**(1), 79–98 (2014)
- V. Fortin, A.-C. Favre, M. Said, Probabilistic forecasting from ensemble prediction systems: improving upon the best-member method by using a different weight and dressing kernel for each member. *Q. J. R. Meteorol. Soc.* **132**(617), 1349–1369 (2006)
- C. Fraley, A.E. Raftery, J.M. Sloughter, T. Gneiting, *ensembleBMA: An R Package for Probabilistic Forecasting Using Ensembles and Bayesian Model Averaging*. Technical Report 516 (Department of Statistics, University of Washington, 2007)
- C. Fraley, A.E. Raftery, T. Gneiting, Calibrating multi-model forecast ensembles with exchangeable and missing members using Bayesian Model Averaging. *Mon. Weather Rev.* **138**, 190–202 (2010)
- P. Friederichs, A. Hense, Statistical downscaling of extreme precipitation events using censored quantile regression. *Mon. Weather Rev.* **135**(6), 2365–2378 (2007)
- F. Fundel, M. Zappa, Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. *Water Resour. Res.* **47**(9), p 9520 (2011)
- H.R. Glahn, D.A. Lowry, The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteorol.* **11**, 1203–1211 (1972)

- T. Gneiting, A.-E. Raftery, A.-H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**(5), 1098–1118 (2005)
- T.M. Hamill, Verification of TIGGE multimodel and ECMWF reforecast-calibrated probabilistic precipitation forecasts over the contiguous United States. *Mon. Weather Rev.* **140**, 2232–2252 (2012)
- T.M. Hamill, J.S. Whitaker, Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Weather Rev.* **134**(11), 3209–3229 (2006)
- J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, Bayesian Model Averaging: a tutorial (with discussion). *Stat. Sci.* **14**(4), 382–417 (1999). Correction: vol. 15, pp. 193–195. The corrected version is available at <http://www.stat.washington.edu/www/research/online/hoeting1999.pdf>
- C. Johnson, N. Bowler, On the reliability and calibration of ensemble forecasts. *Mon. Weather Rev.* **137**, 1717–1720 (2009)
- T.H. Kang, Y.O. Kim, I.P. Hong, Comparison of pre- and post-processors for ensemble streamflow prediction. *Atmos. Sci. Lett.* **11**(2), 153–159 (2010)
- R. Koenker, Quantile regression in R: a vignette (2018). Retrieved from <https://cran.r-project.org/web/packages/quantreg/vignettes/rq.pdf>
- R. Koenker, G. Bassett, Regression quantiles. *Econometrica* **46**(1), 33–50 (1978)
- E.N. Lorenz, Atmospheric predictability as revealed by naturally occurring analogues. *J. Atmos. Sci.* **26**(4), 636–646 (1969)
- R. Marty, V. Fortin, H. Kuswanto, A.-C. Favre, E. Parent, Combining the Bayesian processor of output with Bayesian Model Averaging for reliable ensemble forecasting. *J. R. Stat. Soc. C Appl. Stat.* **64**(1), 75–92 (2014)
- J.W. Messner, A. Zeileis, *crch: Censored Regression with Conditional Heteroscedasticity* (2013). R package version 0.1-0
- J.W. Messner, G.J. Mayr, D.S. Wilks, A. Zeileis, Extending extended logistic regression: extended vs. separate vs. ordered vs. censored. *Mon. Weather Rev.* **142**(8), 3003–3014 (2014a)
- J.W. Messner, G.J. Mayr, A. Zeileis, D.S. Wilks, Heteroscedastic extended logistic regression for post-processing of ensemble guidance. *Mon. Weather Rev.* **142**(1), 448–456 (2014b)
- J.A. Nelder, R.W.M. Wedderburn, Generalized linear models. *J. R. Stat. Soc. Ser. A Gen.* **135**(3), 370–384 (1972)
- T.C. Pagan, D.L. Shrestha, Q.J. Wang, D. Robertson, P. Hapuarachchi, Ensemble dressing for hydrological applications. *Hydrol. Process.* **27**(1), 106–116 (2013)
- A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian Model Averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174 (2005)
- E. Roulin, S. Vannitsem, Post-processing of ensemble precipitation predictions with extended logistic regression based on hindcasts. *Mon. Weather Rev.* **140**, 874–888 (2012)
- E. Roulin, S. Vannitsem, Post-processing of medium range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrol. Process.* **29**(6), 1434–1449 (2014)
- M.-S. Roulston, L.-A. Smith, Combining dynamical and statistical ensembles. *Tellus* **55A**(1), 16–30 (2003)
- R. Schefzik, T.L. Thorarinsdottir, T. Gneiting, Uncertainty quantification in complex simulation models using ensemble copula coupling. *Stat. Sci.* **28**(4), 616–640 (2013)
- M. Scheuerer, Probabilistic quantitative precipitation forecasting using ensemble model output statistics. *Q. J. R. Meteorol. Soc.* **140**(680), 1086–1096 (2014)
- M.J. Schmeits, K.J. Kok, A comparison between raw ensemble output, (modified) Bayesian Model Averaging, and extended logistic regression using ECMWF ensemble precipitation reforecasts. *Mon. Weather Rev.* **138**, 4199–4211 (2010)
- J.M. Sloughter, A.E. Raftery, T. Gneiting, Probabilistic quantitative precipitation forecasting using Bayesian Model Averaging. *Mon. Weather Rev.* **135**, 3209–3220 (2007)
- S.J. van Andel, A.H. Weerts, J. Schaake, K. Bogner, Post-processing hydrological ensemble predictions intercomparison experiment. *Hydrol. Process.* **27**(1), 158–161 (2012)

- H.M. Van Den Dool, Searching for analogues, how long must we wait? *Tellus A* **46**, 314–324 (1994)
- B. Van Schaeybroeck, S. Vannitsem, Post-processing through linear regression. *Nonlinear Process. Geophys.* **18**, 147–160 (2011)
- B. Van Schaeybroeck, S. Vannitsem, Ensemble post-processing using member-by-member approaches: theoretical aspects. *Q. J. R. Meteorol. Soc.* **141**(688), 807–818 Part: A Published: APR 2015. Early View (2014)
- S. Vannitsem, A unified linear model output statistics scheme for both deterministic and ensemble forecasts. *Q. J. R. Meteorol. Soc.* **135**(644), 1801–1815 (2009)
- J.S. Verkade, J.D. Brown, P. Reggiani, A.H. Weerts, Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *Mon. Weather Rev.* **135**(6), 2379–2390 (2007)
- M.-P. Wand, M.-C. Jones, *Kernel Smoothing* (Chapman and Hall, London, 1995)
- A.H. Weerts, H.C. Winsemius, J.S. Verkade, Estimation of predictive hydrological uncertainty using quantile regression: example from the national flood forecasting system (England and Wales). *Hydrol. Earth Syst. Sci.* **15**, 255–265 (2011)
- D.S. Wilks, Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteorol. Appl.* **13**(3), 243–256 (2006)
- D.S. Wilks, Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteorol. Appl.* **16**(3), 361–368 (2009)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences* (Academic/Elsevier, 2011)
- D.S. Wilks, T.M. Hamill, Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Weather Rev.* **135**(6), 2379–2390 (2007)
- R.M. Williams, C.A.T. Ferro, F. Kwasniok, A comparison of ensemble post-processing methods for extreme events. *Q. J. R. Meteorol. Soc.* **140**, 1112–1120 (2014)
- L.J. Wilson, S. Beauregard, A.E. Raftery, R. Verret, Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging (with discussion). *Mon. Weather Rev.* **135**, 1364–1385 (2007). Discussion pages 4226–4236
- A.W. Wood, Dynamical-statistical approaches for hydrological ensemble prediction, in *Science Symposium Proceedings*, Melbourne, Australia, 2012 (CSIRO: Water for a Healthy Country National Research Flagship, 2012)
- A.W. Wood, J.C. Schaake, Correcting errors in streamflow forecast ensemble mean and spread. *J. Hydrometeorol.* **9**, 132–148 (2008)
- L. Zhao, Q. Duan, J. Schaake, A. Ye, J. Xia, A hydrologic post-processor for ensemble streamflow predictions. *Adv. Geosci.* **29**, 51–59 (2011)



Seasonal Ensemble Forecast Post-processing

Andy Wood, A. Sankarasubramanian, and Pablo Mendoza

Contents

1	Introduction	820
2	Seasonal Streamflow Forecast Approaches and Characteristics	821
2.1	Seasonal Hydrologic Predictability	821
2.2	Statistical/Empirical Forecasting	822
2.3	Dynamical (Model-Based) Forecasting	823
2.4	Hybrid and Multi-model Forecasting	825
3	Post-processing of Single-Source Seasonal Ensemble Forecasts	825
3.1	Bias Correction Approaches	827
3.2	Seasonal Forecast Calibration Through Regression	827
3.3	ESP Trace Weighting	829
3.4	Sub-seasonal Disaggregation of Seasonal Runoff Volume Forecasts	830
3.5	Merging of Forecast Ensembles for Different Forecast Periods	831
4	Combination of Multiple Forecasts	832
4.1	Weighted Resampling	833
4.2	Bayesian Joint Probability (BJP)	834
4.3	Bayesian Model Averaging (BMA)	835
4.4	Quantile Model Averaging (QMA)	836
4.5	Bayesian Hierarchical Modeling	837
4.6	Challenges in the Design of Multi-model Systems	838
5	Discussion	839
	References	841

A. Wood (✉)

National Center for Atmospheric Research, Boulder, CO, USA

e-mail: andywood@ucar.edu

A. Sankarasubramanian

Department of Civil Construction and Environmental Engineering, North Carolina State University, Raleigh, NC, USA

e-mail: sankar_arumugam@ncsu.edu

P. Mendoza

Advanced Mining Technology Center (AMTC), Universidad de Chile, Santiago de Chile, Chile

e-mail: pablo.mendoza@amtc.uchile.cl

Abstract

In many parts of the world, water resources systems manage sub-seasonal to seasonal (S2S) variability in climate and runoff in part through the use of operational streamflow forecasts, supplemented by predictions of climate and other hydrologic variables. S2S hydrologic forecasts are produced through both statistical and dynamical (model-based) approaches, and separate S2S forecasts may be combined in multi-model frameworks to increase their skill. Statistical post-processing can be used to enhance the skill and reliability of model-based S2S predictions, and to reduce bias, as well as to merge forecasts from multiple approaches. This chapter describes seasonal hydrologic forecast approaches and products, and presents common techniques used in both the post-processing of single ensemble forecast series as well as the combination of multiple forecasts. Also discussed are the sources of S2S hydrological predictability and particular challenges and opportunities related to post-processing seasonal hydrologic predictions, for which the sample sizes of past simulations, observations and predictions are relatively more limited than in the context of short to medium range prediction.

Keywords

Post-processing · Seasonal forecast · Multi-model combination · Bias-correction · Ensemble forecast · Watershed model · Statistical forecasting · Predictability · Hydrologic variability · Climate

1 Introduction

In many parts of the world, water resources systems experience sub-seasonal to seasonal variability that is actively managed in part through the use of operational seasonal streamflow forecasts, supplemented by seasonal predictions of climate and other hydrologic variables such as snowpack or potential evaporation, which influences water demand, as well as other situational variables (Raff et al. 2013). The practice of seasonal streamflow forecasting is nearly a century old (Helms et al. 2008) with some of the earliest forecasts taking the form of spring snowmelt runoff estimates, using graphical techniques to relate manually measured snow depths in the headwaters of reservoir drainage areas to the impending inflow volumes. Operational statistical seasonal streamflow prediction continues today (e.g., Pagano et al. 2014) with increasingly sophisticated regression and machine-learning techniques being investigated to harness the relationships between potential predictors of watershed variability and future runoff (e.g., Moradkhani and Meier 2010; Rosenberg et al. 2011). In general, the skill of operational seasonal predictions ranges from very low, at times when predictor-predict and relationships are weak, to very high, such as when the dominant contribution to future runoff is already contained in the stored moisture (snow or soil) in the watershed. Because the longer lead times of seasonal-scale predictions bring relatively lower skill than is expected for short-range weather and flood predictions, seasonal predictions have for decades been augmented by estimates of forecast uncertainty (i.e., error bounds), which help support risk-based decision making.

The development of computer-based hydrologic modeling in the 1970s led to the rapid rise of operational seasonal streamflow predictions being generated by running conceptual simulation models of a basin snowpack and soil water balance. One of the earliest such systems was deployed by the US National Weather Service (NWS) to produce an ensemble forecast of reservoir inflows (Day 1985; Wood et al. 2016b), defining a technique that came to be known as “Ensemble Streamflow Prediction” (ESP) and to be widely used by forecasting centers in numerous countries. Seasonal ESP forecasts quantify uncertainty by using a deterministic simulation of watershed model states at the start of the forecast, termed initial hydrologic conditions (IHCs), to initialize an ensemble of simulations for the prediction period that are driven by sequences of forecast meteorology. In the traditional ESP approach, these sequences are drawn from historical observations of weather during the forecast period; more recently, a number of forecast centers derive them from the outputs of NWP and climate forecasting models (e.g., Crochemore et al. 2016).

Several characteristics of seasonal streamflow forecasts are important to users (e.g., water managers) and stakeholders. Not surprisingly, users desire the highest possible forecast skill for the median forecast (as measured through various metrics such as the coefficient of determination, R^2), but they also expect that the median forecast will be unbiased and that the forecast spread will provide statistically reliable estimates of uncertainty. Statistical post-processing can be used to enhance or achieve these characteristics, either through application to ensemble forecasts from a single approach or in merging forecasts from multiple approaches (as in a multi-model forecast system).

This chapter focuses primarily on the post-processing of dynamical seasonal hydrologic forecasts, including post-processing applications for merging multi-model forecasts. Section 2 of the chapter describes common seasonal hydrologic forecast approaches and products, as well as sources of prediction skill, uncertainty, and error. Section 3 discusses the post-processing of single ensemble forecast series, while Sect. 4 describes examples and techniques for the post-processing combination of multiple forecasts. Section 5 summarizes the particular challenges and opportunities related to seasonal hydrologic prediction. A wide variety of approaches appear in the literature, most of which are either sufficiently standard (e.g., linear regression) or sufficiently complex that it would be ineffective to reproduce their algorithmic details in this chapter. The methods are instead described and couched in references that will allow readers to pursue any particular technique at a greater depth if needed.

2 Seasonal Streamflow Forecast Approaches and Characteristics

2.1 Seasonal Hydrologic Predictability

At seasonal lead times, hydrologic predictions – whether dynamical or statistical – derive predictability from two major sources (Wood and Lettenmaier 2008; Greuell et al. 2016; Wood et al. 2016a; Koster and Mahanama 2012). At lead times of

months to a few seasons, it is common for initial watershed moisture and to a lesser extent energy (collectively, IHCs) to exert a clear if not dominant influence on hydrologic forecasts – a result of the inertia and persistence found in hydrologic systems. Where groundwater contributes a significant fraction of river flow, IHCs can even affect forecast outcomes for lead times of over a year (Harrigan et al. 2017). The second major source of skill is weather and climate during the forecast period, or future meteorological forcings (FMFs), which drive the evolution of the IHCs and associated runoff. These two determinants of hydrologic predictions are directly analogous to the initial value and boundary condition components of the weather and climate forecasting problem. Relative to the atmosphere, hydrologic systems are more strongly influenced by initial values at equivalent lead times, though the strength of IHCs and FMFs varies greatly depending on watershed conditions and characteristics.

The dual forecast skill influences imply that the post-processing of hydrological forecasts, to a greater degree than meteorological forecasts, can benefit from conditioning that accounts for the current state or “regime” of a watershed as well as, in some cases, the state of the atmosphere. Geographic regions exhibiting strong climate seasonality may also exhibit distinct hydrologic regimes. For example, forecasts in mid-to-northern latitude regions behave differently depending on whether a watershed is snow-covered or not, and regions that experience dry/warm and wet/cold seasons may also show different forecast characteristics in these distinct hydroclimate regimes. Hydrologic post-processing methods are likely to perform best when they factor in significant regime dependencies of forecast error, related both to IHCs and FMFs. A common conditioning factor for short-range forecasts, for example, is the observed or forecast streamflow, which can be used to separate baseflow or recession from high or active flow regimes (e.g., Hoss and Fischbeck 2015; Seo et al. 2006).

2.2 Statistical/Empirical Forecasting

Statistical forecasting approaches employ purely data-driven methods that rely on empirical relationships between seasonal streamflow volumes, in situ watershed observations, and large-scale climate variable observations. A number of statistical approaches can be found in the literature, exhibiting different degrees of complexity. A common example of a seasonal hydrologic forecast in the western United States is the issuance of spring snowmelt runoff volume for a predictand period of several months (e.g., April–July). The operational prediction models used by two US federal agencies are formed using principal components regression (PCR) to derive linear regression predictors from in situ watershed observations of snow water equivalent (measured using snow pillows), accumulated precipitation prior to the forecast date, antecedent streamflow, and occasionally climate indices (e.g., Garen 1992; Slater et al. 2017). PCR was adopted to circumvent the problem of predictor multicollinearity when using multiple intercorrelated station observations from within or near a single watershed.

This basic statistical approach has since been refined and reformulated to use other statistical methods, including Z-score regression (Pagano et al. 2009), partial least squares regression (PLSR) (Tootle et al. 2007), linear discriminant analysis (Piechota and Chiew 1998; Piechota et al. 2001), multiple linear regression (e.g., Berg and Mulroy 2006; Hidalgo-Muñoz et al. 2015), canonical correlation analysis (e.g., Salas et al. 2011), independent component analysis (e.g., Westra et al. 2008), semi-parametric techniques (e.g., Sankarasubramanian and Lall 2003; Souza Filho and Lall 2003), and nonparametric regression (e.g., Opitz-Stapleton et al. 2007). Other research has shown that watershed variables simulated by hydrologic models (such as SWE and soil moisture) can augment or even replace direct watershed observations as predictors in statistical frameworks (Rosenberg et al. 2011, 2013; Robertson et al. 2013) – indicating the potential for hybrid dynamical-statistical prediction approaches. Another vein of inquiry has focused on incorporating predictors related to the climate system – i.e., attempting to harness the second source of predictability (Grantz et al. 2005; Wang et al. 2009; Moradkhani and Meier 2010).

Statistical methods are briefly described here because they remain prominent in seasonal volume prediction operations. In contrast to dynamical (model-based) seasonal forecasts, however, the purely statistical seasonal predictions described above tend to have low and high statistical reliability, though not necessarily higher predictive skill (e.g., explained variance). This is a result of the explicit optimization of the statistical forecast model parameters given the seasonal prediction objective, though it can be undermined by parameter estimation uncertainty in the case of small sample sizes. For this reason, the post-processing approaches described in this chapter presented primarily in the context of applications to forecasts based on dynamical models an approach summarized in the next section.

2.3 **Dynamical (Model-Based) Forecasting**

Dynamical hydrology models simulate watershed processes including snow accumulation and melt, rainfall and melt partitioning, and baseflow and runoff generation. In the 1970s, seasonal streamflow forecasting using computer models began with low-dimensional (“simple”) conceptual watershed models run on mainframe computers (Linsley and Crawford 1974; Burnash et al. 1973), and the ESP technique was introduced (Day 1985; Wood et al. 2016a). Hydrologic forecasting models today rely on either the legacy conceptual schemes or on more explicit, complex process-resolving physical parameterizations (i.e., algorithms) to predict future streamflow. They are typically embedded within a forecast system in which many repetitive data processing tasks are semiautomated, but expert teams of forecasters may monitor or adjust various elements of the forecast workflow (including the models) in real time.

The ESP approach is a template for most seasonal model-based forecasts. ESP runs a deterministic simulation of a watershed’s hydrologic state up to the forecast start date and then evolves the initial states and streamflow into the future (i.e., forecast period) by running the model with an ensemble of future meteorological

model input sequences. These sequences may be drawn from historical observations during the prediction period or by deriving conditional input sequences based on operational climate forecasts or other information about climate state. An example of ESP ensemble trace forecast product is shown in Fig. 1. A key motivation for the ESP approach is that it accounts for the impact of FMF uncertainty on the evolution of the presumably better-known watershed conditions (IHCs). At times, the FMF uncertainty is the major source of future runoff uncertainty, but numerous studies have also shown that IHCs provide a dominant influence on the runoff forecast signal. In such cases, ESP's deterministic estimate of the forecast ensemble's IHC signal does not incorporate modeling uncertainties and systematic errors, yielding overconfident (underspersive) and biased forecasts (Wood and Schaake 2008).

Hydrologic model uncertainty can arise from errors associated with parameter estimation, meteorological input forcings, or model physics and structure (e.g., Wagener et al. 2003; Beven 1993; Sorooshian et al. 1993). Three sources of model bias and/or error when used for seasonal forecasting include (a) the optimization of

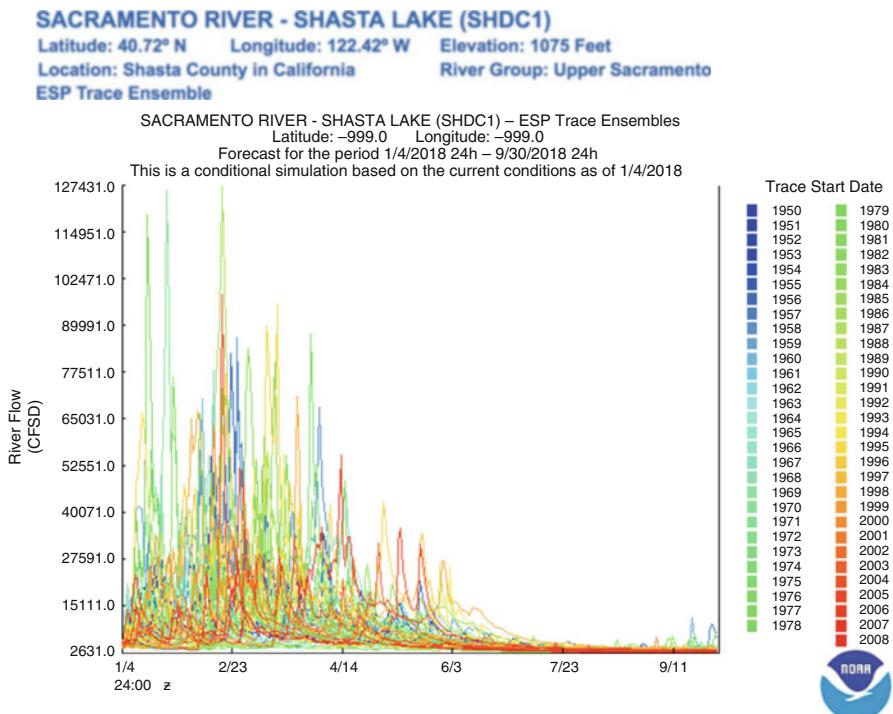


Fig. 1 Example of traditional ESP ensemble forecast traces in a forecast initialized Jan 3, 2017, for Shasta Reservoir, CA, inflows. (Accessed online from http://www.cnrfc.noaa.gov/send_espTrace.cgi?location=SHDC1&accumtype=mean&interval=day&disttype=empirical&S_month=01&S_day=04&S_year=2018&E_month=09&E_day=30&E_year=2018&plottype=traces&outtype=Generate+a+Plot&tabletype=forecastinfo on 2017-01-03)

model parameters for modeling objectives other than the seasonal prediction period flow (e.g., for daily flow values throughout the year); (b) systematic errors and biases in the real-time meteorological input analyses, which often rely on degraded observational networks and otherwise differ from model calibration forcings; and (c) potential errors arising from forecaster real-time modifications to model states. Consequently, when model-based ensemble streamflow forecasts are used either as a single-source prediction or within a multi-model forecast combination, the resulting systematic biases in the forecast ensemble mean and spread can be improved through post-processing (e.g., Seo et al. 2006).

2.4 Hybrid and Multi-model Forecasting

Although operational agencies use statistical and dynamical methods individually to generate forecast products, recent work has shown skill advantages to approaches that combine statistical and dynamical techniques. In practice, forecasters have traditionally integrated multiple sources of information, often using expert judgment. For example, Fig. 2 shows a seasonal forecast series of reservoir inflow volume during the snowmelt period in which daily inflow ranges predicted by ESP are plotted together with an “official” forecast made once per month by integrating the ESP results with a purely statistical forecast. Also shown is the progression of the observed accumulated runoff through the period to provide the verifying end-of-period observation.

The statistical post-processing of a model-based forecast may also be considered a hybrid prediction technique. One general forecasting approach is the strategy of statistically combining multiple predictions either from different dynamical models or from dynamical and statistical models. This approach often takes a hierarchical form, with a first level of prediction analysis being the formation of individual forecasts (deterministic or probabilistic), followed by a second level of analysis being their merger. These approaches are discussed further in Sect. 3.

3 Post-processing of Single-Source Seasonal Ensemble Forecasts

Sub-seasonal to seasonal forecasts from a single forecast source may be post-processed not only to correct for systematic errors but also to add information or skill through techniques such as merging with long-range ensembles with shorter-range predictions from a separate system or through disaggregation to provide sub-period shaping to predicted seasonal volumes. This section describes a range of approaches that are used in this context. As discussed in chapter ▶ “[Motivation and Overview of Hydrological Ensemble Post-processing](#)”, important considerations that affect the applicability of each post-processing approach include whether consistent retrospective model runs and hindcast series exist and their sufficiency for training a post-processing method.

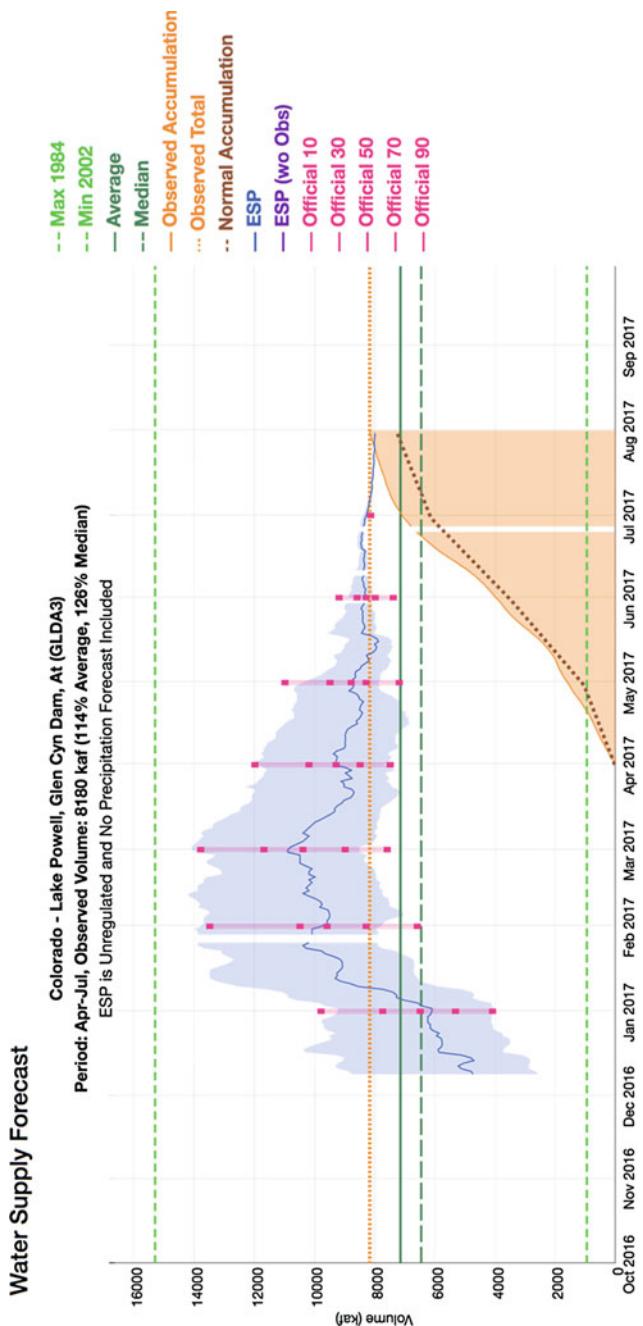


Fig. 2 Example of a seasonal probabilistic series of runoff volume forecasts for Lake Powell reservoir inflow during the April–July snowmelt period, produced operationally by the US NWS Colorado Basin River Forecast Center. Daily ESP runoff forecast ranges are shown versus a multi-model forecast (the “official”) formed from ESP and a regression-based approach. (Accessed online from http://www.cbrfc.noaa.gov/wsup/graph/front/esplot_dg.html?year=2017&id=GLDA3 on 2017-01-03)

3.1 Bias Correction Approaches

Seasonal forecasts can suffer from systematic bias for a variety of reasons. Bias can arise from modeling error as described in Sect. 2.2, and it can also result from biases in the meteorological forecasts. For example, in a traditional ESP, the historical meteorological ensemble sample may represent a past climatology that is no longer valid to describe climate uncertainty for the current forecast, as in the case of a watershed in which there has been a strong warming trend or pronounced decadal climate variability (Lehner et al. 2017).

Model bias alone can be removed using results from only a retrospective simulation of sufficient length to represent a basin's variability. One first calculates a mean scalar error factor (i.e., $Q_{\text{obs}}/Q_{\text{sim}}$) from the past simulation and then multiplies the ensemble forecast member values by this factor. One can account for seasonal or process-related variations in this mean bias through calculating factors for different seasons or conditions (i.e., snow accumulation period, snowmelt period, dry summer low-flow season). If a model simulation exhibits different biases in different parts of the flow distribution (e.g., a negative bias at high flows and a positive bias at low flows), quantile mapping (QM; Wood and Schaake 2008; Hashino et al. 2007; also called “cdf matching”) can be used to post-process simulated flow values. The percentile of a simulated flow $F_s(Q)$, where F_s is the cumulative distribution function (CDF) value of the simulated flow Q_s , extracts a corresponding conditional forecast flow Q_{bc} from the inverse CDF of the observations, F_o^{-1} :

$$Q_{bc} = F_o^{-1}(F_s(Q_s)) \quad (1)$$

Empirical or fitted CDFs may be used, but in the former case it may be necessary to use a fitted distribution where either the simulated flows exceed the bounds of the simulated climatology or their percentiles exceed the range of the observed inverse CDF. The correction forces the mean and variance of the simulation outputs to match those of the observed climatology. The observed and simulated CDFs are typically estimated from the same time period, often of a number of years of simulation to provide a large enough sample to estimate stable CDFs.

If a hindcast series of sufficient length is available (i.e., providing a sample of more than 30 hindcasts), these two bias correction techniques can also be used to post-process total forecast error (e.g., both model and meteorological input forecast error, assuming the error of observations is minimal). The correction methods are applied similarly, with one CDF being calculated for the series of hindcasts and one for the verifying observations (e.g., Lucatero et al. 2017).

3.2 Seasonal Forecast Calibration Through Regression

Bias correction addresses only systematic biases without guaranteeing statistically reliable outcomes (i.e., in which forecast spread is commensurate with forecast skill).

Zhao et al. (2017) argue (in the context of seasonal precipitation forecasts from climate models) that bias correction alone is only partially sufficient for post-processing, in that the errors in forecast signal and spread are only adjusted to correct for climatological biases. If a hindcast series and matching observations are available, their information about the pairwise correspondence of forecasts and observations (hence the skill) can be used to create calibrated, unbiased conditional forecast ensembles. Where bias correction applies the same adjustments to forecasts regardless of lead time and skill, regression techniques account for systematic differences in skill that might exist for forecasts at different lead times or made in different seasons of the year. The spread and signal of an unskillful forecast will damp toward a naive climatology when calibrated, yielding forecasts that default to being no worse than climatology (Zhao et al. 2017; Wood and Schaake 2008).

The forms of regression described in Handbook chapter “Seasonal Ensemble Forecast Post-processing” for post-processing short-range ensembles are also applicable to seasonal ensemble predictions, with one major caveat. The chief hurdle, as discussed in both Handbook chapter ► [“Motivation and Overview of Hydrological Ensemble Post-processing”](#) and Sect. 3.1, is that the sample size of available seasonal hindcasts may not be sufficient to estimate the post-processing model parameters – particularly if the signal to noise ratio in the hindcasts is low (leading to noisy or unstable parameter estimates). For this reason, parametrically parsimonious approaches (i.e., having just a few parameters) are preferable to reduce the risk of overfitting the post-processing model, and cross-validation and verification of the approach are critical. For low-skilled forecasts, difficulty improving upon one characteristic (e.g., bias) may degrade another (e.g., correlation skill) due to the impact of parameter estimation uncertainties.

A relatively simple method for forecast calibration is multiple linear regression, in which the parameters of the post-processing equation model are estimated by regressing the verifying observations against parts of the forecast ensemble (e.g., the median and at times the spread), based on a hindcast series. If the forecast components and observations depart significantly from normality, they should each be normalized before estimating regression parameters, and the post-processing procedure applied in normal space, after which the conditional forecast is back-transformed to streamflow space. Common transforms for streamflow are the square and cube root, the normal quantile transform, and the Box-Cox transform. The log normal distribution can also be used, but is considered fairly aggressive, risking instability for high values upon the reverse exponentiation. The resulting forecast model error can be used to estimate the quantiles of the forecast uncertainty distribution. Note that when estimating conditional forecast quantiles, the error deviates are added to the predicted mean in normal space, before being transformed to flow space. Ideally, the error is estimated from a series of jackknifed or cross-validated hindcasts (as discussed earlier), in which the error from out-of-sample hindcasts defines the forecast model uncertainty rather than the standard error of the regression training.

Some stakeholder applications (such as running an ensemble-based reservoir operations model) require the forecast in the form of an ensemble of streamflow sequences rather than a probability distribution of future inflow volumes. In some cases, the raw ESP members can be scaled to reproduce the conditional,

post-processed distribution of time-aggregated flow volumes. Such a trace generation approach is likely to be preferable to a conditional resampling of historical flows when the ESP forecast ensemble reflects a strong contribution of IHCs, which particularly affect the traces at shorter lead times. The post-processing regression may use only the model forecasted streamflow or may leverage additional predictors (such as related to climate or ancillary watershed predictors thought to add information). Commonly available statistical software platforms (such as the free R software for statistical computing) offer robust routines for developing linear regression models.

Unlike linear regression, other forms of post-processing such as quantile regression (QR; see chapter “Seasonal Ensemble Forecast Post-processing” for details) are notable for allowing one to estimate the conditional forecast quantiles directly from the forecast ensemble rather than derive them from the predictive error distribution of the regression of the ensemble median. To date, however, QR has been applied most frequently in the forecasting context to shorter-range predictions that offer larger training samples.

3.3 ESP Trace Weighting

Ensemble “trace weighting” is one of the earliest used approaches for post-processing ESP forecasts, and the first usage was motivated by the desire to incorporating a future climate prediction signal into ESPs in which the ensemble forecast meteorology is sampled from historical year meteorological sequences. Hamlet and Lettenmaier (1999), among others, showed that using the current El Nino-Southern Oscillation (ENSO) index to select ensemble members (i.e., “traces”) of the ESP ensemble for which the meteorological forcings had been drawn from years experiencing a similar ENSO index category (i.e., La Nina, El Nino) to the current climate, while discarding other members, improved seasonal streamflow prediction skill for rivers in the Pacific Northwest. This basic approach, which assigned an equal weight to all matching category members and a weight of zero to nonmatching members, has also been used with other climate-related covariates (such as geopotential height patterns) and indices.

The simple category selection technique was later expanded by Werner et al. (2004) to allow a local (rather than categorical) weighting of the ESP members based on past member “similarity” to current conditions. Similarity can be defined by any hydroclimate factor deemed relevant or likely to add skill, though Werner et al. (2004) also used climate indices. This trace-weighting scheme, which can also be thought as an analog-based weighting, requires two parameters and involves the following steps:

1. Select predictors that will be used to assign the weight of the ESP members (e.g., climate index, climate model forecast), forming a predictor vector for the current forecast target year (x_t) for each of the n years from which the historical meteorological input sequences in the ESP were drawn.

2. Compute Euclidean distances (d_i) between the predictor vector values for the forecast year and the predictor vector values in each of the historical years:

$$d_i = \| x_t - x_i \| \quad i = 1, n \text{ years} \quad (2)$$

3. Select k ensemble members out of the total ensemble n that will receive a non-zero weight using the parameter α :

$$k = \text{NINT}(n/\alpha) \quad (3)$$

where NINT refers to the nearest integer operator.

4. Assign non-zero weights w_i to the k ensemble members having the lowest distances from (i.e., that are most similar to) the current forecast according to

$$w_i = (1 - d_i/d_k)^\lambda \text{ for } i = 1, k \text{ members} \quad (4)$$

where ensemble member k has the largest distance in the remaining ensemble and parameter λ influences the shape of the weighting function.

The weights are then normalized, and forecast quantiles (e.g., non-exceedance percentiles such as the forecast median) can be calculated from the weighted cumulative distribution function (CDF).

Other published studies on trace-weighting approaches, including a Bayesian approach to estimating climate index-based weights proposed by Bradley et al. (2015), further demonstrate that ESP trace weighting can improve forecast skill where informative covariates can be found (Beckers et al. 2016). The technique is popular because it is often more straightforward to implement for existing operational forecasting centers than methods that require the pre-generation of conditional ensemble input forcings for an ESP (e.g., Verdin et al. 2015; Wood and Lettenmaier 2006). Mendoza et al. (2017), however, note that the trace weighting can only reshape the distribution of an ESP forecast within its original distributional bounds. Thus, the ability of trace weighting to correct large biases in ESP forecasts (i.e., to substantially shift the ESP distribution) may be limited, and it may be most effective when the ESP forecast has at least moderate skill in its uncorrected form.

3.4 Sub-seasonal Disaggregation of Seasonal Runoff Volume Forecasts

Water managers who use seasonal runoff forecasts often want to know more about the sub-seasonal shape (e.g., weekly, monthly) of the streamflow hydrograph the forecast period. Probabilistic volume forecasts describe total period amounts only, while ensemble trace-based forecasts offer many versions of the possible runoff

timing, varying by ensemble member, with little guidance toward favoring one versus another. Sub-seasonal prediction has received less attention than seasonal forecasting in part because it requires long-lead knowledge of the sub-seasonal timing of future precipitation, for which forecast skill is low. Consequently, operational hydrologic sub-seasonal products are rare.

Since the 1960s, conditional streamflow disaggregation schemes have been existed to support synthetic trace generation used in water resources infrastructure design (e.g., Valencia and Schakke Jr 1973). More recently, simpler semi-parametric techniques have been used to select sub-seasonal patterns for application to seasonal forecasts, including the use of analogs. An example is the application of K-nearest neighbors (KNN; Rajagopalan and Lall 1999) method to identify the best matching analog year(s) from historical periods matching the current forecast period and then to apply the sub-seasonal proportions from the analog(s) to the current forecast volume. The premise of this approach is that vector of current year features (such as total forecast runoff volume, antecedent streamflow, and/or an index of climate) can be found that is informative about forecast shape. In practice, such approaches can struggle to outperform a disaggregation based on climatological sub-seasonal shape, but in snowmelt-driven systems, forecasted runoff magnitude can be weakly indicative of runoff peak timing that is later than average (and vice versa for a relatively low magnitude forecast).

3.5 Merging of Forecast Ensembles for Different Forecast Periods

Seamless hydrologic prediction has been touted as a desirable operational goal, but many forecast systems are optimized to produce separate forecast products matching different requirements at different lead times. It is not uncommon, for example, for a center to generate a deterministic or small-ensemble short-range product based on the need for rapid updates and the availability of a certain NWP input source, with a separate longer-range ensemble forecast that is updated less frequently. The merging of sequential ensembles from different sources involves a post-processing step to linking shorter-range with longer-range ensemble members when there is no obvious association between the two members.

In this case, the Schaake shuffle (Clark et al. 2004) technique or one of the more recent variants (e.g., Shefzik 2016) can be used. The Schaake shuffle relies on samples of observed historical sequences to provide a joint ensemble rank structure that can be used to link multiple ensembles. It has been described as an empirical copula in part because of similarities between the monotonically increasing rank structures linking the ensembles and the monotonically increasing parametric copula function linking the marginal distributions. The ensembles that are linked can be variables separated in time (as discussed here), space, or both.

The procedure for linking two forecast flow ensembles in time is as follows.

1. A set of flow sequences for the calendar period encompassing the time span of the two forecast ensembles is sampled from the historical observed flow record for

- the forecast location. The number of flow sequences equals the number of ensemble members in the larger of the two ensembles, if they are not equal in size.
2. Meaningful index values are chosen to represent each member in each of the two ensembles. The index values should be selected to capture the maximum covariance between the flows during the two forecast time periods. For example, if the short-range ensemble is 1 week long, and the seasonal ensemble to which it will be merged is 3 months long, the short-range ensemble may be represented by the full week mean flow, and the seasonal ensemble may be represented by the mean flow in the first week, few weeks, or month, rather than the entire mean flow of the longer seasonal period.
 3. The index values from each ensemble are sorted (ranked) by magnitude, and then the members of the first and second ensembles are concatenated together according to the ranks drawn in Step 1.

The resulting seamless ensemble now contains a rank structure between the two periods that is consistent with observed historical patterns while containing the values of the original un-merged ensembles.

4 Combination of Multiple Forecasts

Following applied research in atmospheric sciences and meteorology showing that the combination of predictions from multiple models could lead to more skillful predictions than those from a single model (e.g., Krishnamurti et al. 2000), hydrologic forecasters have also explored the benefits of multi-model prediction. Currently, multi-model weather and climate prediction systems or initiatives exist for research and operations at seasonal lead times (e.g., the US National Multi-Model Ensemble; Kirtman et al. 2014). Multiple model forecast systems are more rare in hydrology than in meteorology, primarily because regional to national streamflow forecasting agencies have traditionally focused only on their own service areas, using a single model (in contrast to meteorological forecast models that commonly have a global extent). Early multi-model efforts in hydrology applied Bayesian model averaging (BMA; Raftery et al. 2005) to merge predictions from several models, yielding simulation and/or forecast skill improvements (e.g., Duan et al. 2007; Najafi and Moradkhani 2015; Bohn et al. 2010; Rajagopalan et al. 2002; Georgakakos et al. 2004; Devineni et al. 2008; Devineni and Sankarasubramanian 2010a, b).

Multi-model approaches are typically a combination of forecasts from different models – either deterministic (i.e., the “poor man’s ensemble”) or probabilistic. In hydrology, probabilistic forecasts may be generated either through statistical or dynamical model methods; thus the multi-model forecasts may combine both types. A common argument in favor of multi-model frameworks is that they harness “compensatory effects” that can reduce the spread coming from individual model errors. Weigel et al. (2008) assessed when multi-model forecasts can perform better than the best single model. Using a synthetic modeling experiment, Wiegel et al.

(2008) demonstrated that the multi-model forecasts reduce the overconfidence of individual model forecasts by pooling the ensembles from multiple models or by optimally weighting the forecast probabilities from individual models. Thus, multi-model climate forecasts produce improved forecast reliability by reducing the aggressiveness/overconfidence (e.g., 90% probability of above-normal precipitation) of individual models. If none of the individual models produce overconfident forecasts, then optimally combined multi-model forecasts approach the skill of the best individual model (Weigel et al. 2008). Because climate forecasts from GCMs often don't produce well-calibrated (i.e., reliable) forecasts, multi-model forecasts often perform better than individual model forecasts (Devineni and Sankarasubramanian 2010a, b; Wang et al. 2013).

Combination methods for the multiple forecasts are a form of statistical forecast post-processing. The subsections below describe several multi-model combination techniques that have been applied to seasonal streamflow forecasting. Practical considerations that may help scientists and practitioners in the design and operational implementation of the methods are also summarized.

4.1 Weighted Resampling

A basic challenge of all multi-model forecast combination approaches is the merging of probabilistic information in the form of probability density functions (PDFs) or ensembles to achieve a superior probabilistic forecast. Solutions to this challenge range from empirical/nonparametric to semi-parametric to parametric (requiring analytical solutions to manipulate statistical distributions).

Weighted resampling is a nonparametric technique for combining seasonal forecasts (e.g., Regonda et al. 2006; Bracken et al. 2010; Mendoza et al. 2014) that is straightforward to implement but requires a consistent hindcast series from which to derive performance weights for each of the forecast candidates. Weighted resampling proceeds as follows.

1. Select a performance metric (e.g., root-mean-squared error (RMSE) of the forecast median; continuous ranked probability skill score of a full ensemble) or suite of metrics used in combination as a basis for assigning a weight to each candidate. Average the metrics for each of the candidate hindcasts.
2. Translate each average metric into a weight that gives preference to the best-performing candidate forecast, and normalize the weights to sum to 1. Examples of weighting functions that could be used with a metric that is negatively oriented (e.g., perfect score is 0 and higher scores indicate poorer performance) are the multiplicative inverse of the metric (e.g., $1/RMSE$) and the Gaussian error function with mean zero and a prescribed variance. The form of the weight function is important because it determines how equally candidate forecasts will be combined across a gradient of skill.
3. Create the post-processed weighted forecast by randomly resampling each candidate forecast with a frequency matching its weight. If the candidate forecasts are

ensembles, the ensemble members are resampled with replacement by transforming uniform random numbers into ensemble member ranks. If the forecasts are PDFs, the uniform random numbers are used to select quantile values through the associated inverse CDF. The samples taken from all candidate forecasts are then pooled to make the new forecast distribution.

If the candidate forecasts have low or unstable skill, it may be useful to add climatology as a candidate forecast to improve the reliability of the forecast combination. Care must be taken to ensure that the candidate weights are stable, which will be discussed further in Sect. 4.6.

With the rise of machine learning, a great breadth of parametric methods for what is called kernel combination have been introduced and can be researched online (e.g., https://en.wikipedia.org/wiki/Multiple_kernel_learning). Most of these have not been applied to seasonal hydrologic probabilistic forecasting, and their discussion is beyond the scope of this chapter, but indicates that there are no shortages of statistical solutions that may be potentially exploited for the multiple forecast combination challenge.

4.2 Bayesian Joint Probability (BJP)

A technique used in a variety of sub-seasonal to seasonal post-processing applications for both precipitation and streamflow is the Bayesian joint probability (BJP) approach (e.g., Wang et al. 2009; Schepen et al. 2016, and other references contained therein). In the multi-model forecast combination context, BJP can merge any number of candidate forecasts, though it has also been used in a bivariate implementation as a post-processor for a single forecast series. BJP involves transforming the predictands to a normal distribution; then forming a multivariate normal distribution that is used to generate a conditional distribution for the predictand, given new values of the predictors; and then back-transforming the predictand. The conditional prediction equations resulting from BJP are the same as in using linear regression for post-processing (see, e.g., Wood and Schaake 2008), but BJP adopts a Bayesian approach to parameter estimation (using a Markov chain Monte Carlo, MCMC) rather than a frequentist approach using, e.g., maximum likelihood estimation via ordinary least squares to identify parameters. As a result, the BJP yields posterior (i.e., post-processed) forecast distributions that explicitly include parameter estimation uncertainty, potentially yielding more reliable and accurate outcomes.

Using the bivariate implementation as an example, BJP involves the following key steps:

1. Transform forecast \mathbf{x} and verify observation \mathbf{y} to normal space variables \mathbf{g} and \mathbf{h} , respectively.
2. Estimate the parameters Θ of the bivariate normal distribution as an ensemble using MCMC, including the means (μ_g and μ_h), variances (σ_g^2 and σ_h^2), and

cross-correlation ρ_{gh} . Prior distributions may involve as many as 1000 random samples.

3. Derive each conditional forecast \mathbf{h}_{new} given raw forecast \mathbf{g}_{new} using the bivariate normal formulation:

$$h_{\text{new}} \mid g_{\text{new}}, \Theta \sim N\left(\mu_h + \rho_{gh} \frac{\sigma_h^2}{\sigma_g^2} (g_{\text{new}} - \mu_g), \sigma_h^2 (1 - \rho_{gh}^2)\right) \quad (5)$$

4. Back-transform \mathbf{h}_{new} to provide conditional, post-processed forecast series. Ensembles are formed by sampling different estimates of Θ .

Full details of this procedure are given in Schepen et al. (2018). The multivariate normal concept is also at the core of the Meteorological Ensemble Forecast Processor (MEFP) used in the NWS Hydrologic Ensemble Forecast Service (HEFS) to downscale precipitation and temperature. Most BJP applications for skewed, truncated random variables such as precipitation and streamflow have used the log-sinh or Yeo-Johnson transformations, whereas the MEFP applies a normal quantile transform (NQT).

4.3 Bayesian Model Averaging (BMA)

The principle of BMA (Raftery et al. 2005) is that given an ensemble forecast with M members, each ensemble member f_i ($i = 1, 2, \dots, M$) is associated with a conditional PDF $h_i(y|f_i)$, which can be interpreted as the PDF of the variable y given f_i . Thus, the BMA predictive model is:

$$p(y|f_1, \dots, f_m) = \sum_{i=1}^M w_i h_i(y|f_i) \quad (6)$$

where the BMA weight w_i is the posterior probability of forecast i and is obtained based on its relative performance during the training period. Therefore, the weights w_i s are nonnegative and add up to 1, i.e., $\sum_{i=1}^M w_i = 1$ (Raftery et al. 2005).

The weights for the models considered can be estimated, for example, maximizing the likelihood function, for which the R package *ensembleBMA* (<https://cran.r-project.org/web/packages/ensembleBMA/ensembleBMA.pdf>) has an expectation-maximization (EM) algorithm (Dempster et al. 1977). Prior information (i.e., initial weights) for each model can be specified based on performance throughout the training period. Finally, the BMA forecast ensemble can be obtained by sampling a fraction of members from each model equal to the weight w_i .

Studies have found that using climatology as one of the candidate models often improves the multi-model forecast skill (e.g., Rajagopalan et al. 2002) in BMA. Since BMA weights are obtained from a calibration period, considering climatological ensembles further helps in reducing the overconfidence of individual model forecasts (Weigel et al. 2008). Estimation of weights for model combination has also

considered predictor conditions (e.g., ENSO condition) if prior knowledge is available about model performance. In this context, model weights could be allowed to vary depending on the predictor conditions, thereby giving higher weights could be given for a model contingent on the climatic conditions. However, Weigel et al. (2008) caution against dynamic weighting approach for seasonal forecasting due to limited sample size available for weight estimation.

4.4 Quantile Model Averaging (QMA)

The QMA forecast is obtained from the weighted average of forecast quantiles from all models:

$$y_{QMA} = \sum_{i=1}^M w_i y_i(F) \quad (7)$$

where M is the number of models and w_i is the model weight for model i such that $\sum_{i=1}^M w_i = 1$. The QMA forecast value is a weighted average of forecast variable values (quantiles) from all models, given a cumulative probability (quantile fraction) F . This can be done by sorting the forecast ensemble members from each model and averaging equally ranked ensemble members across all models (Schepen and Wang 2015).

A notable difference between BMA and QMA is that the latter produces smoother and consistently unimodal distributions compared to potentially bimodal BMA outputs (Fig. 3). Schepen and Wang (2015) recently compared BMA and QMA for seasonal streamflow forecasting in Australia, finding that nearly identical skill

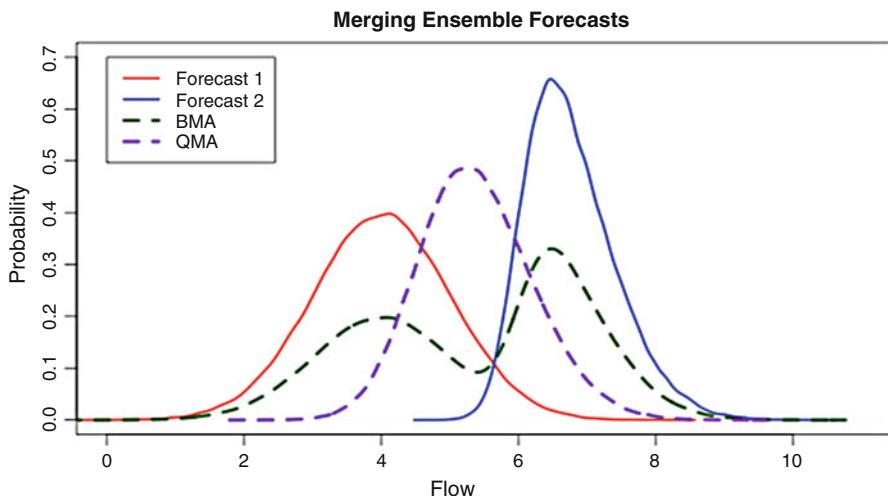


Fig. 3 Synthetic example of merging two probabilistic forecasts using BMA and an equally weighted QMA. (Based on Schepen and Wang 2015)

results can be obtained with BMA and QMA and that very similar performance can be achieved either by calibrating QMA weights or by using BMA weights within a QMA framework.

4.5 Bayesian Hierarchical Modeling

Bayesian hierarchical modeling (BHM) (e.g., Chen et al. 2014) offers flexibility in developing forecasting models by considering predictors and predictands that may be available at different spatiotemporal scales or across other different categories of information. For instance, climate information (e.g., SSTs or gridded precipitation forecasts) available over large spatial scales may have similar response on the hydroclimatology of the region, but individual site-specific land surface conditions (e.g., SWEs) and basin attributes (e.g., drainage area) could modulate the regional response. Under this situation, individual regression models developed independently for each site (i.e., no space-time pooling) may ignore detectable covariability across predictors and the predictands. Another approach is to develop one single regression model (i.e., full pooling) for the entire region by pooling all the predictors and predictands across the region, at the risk of suffering from significant at-site bias in the prediction models. Hierarchical models, aka multilevel models, however, pursue a partial pooling of information (e.g., temporal covariability of predictors) for forecasting the predictands (e.g., streamflow/precipitation) by estimating the regression variable intercepts and slopes (coefficients) across the region (Devineni et al. 2013). Given BHM estimation procedure is fully probabilistic, the forecast attributes are naturally available as a conditional distribution (e.g., Chen et al. 2014).

Chen et al. (2014) applied BHM for forecasting seasonal streamflow and precipitation using SSTs over the Huai River basin in China and evaluated the forecasting skill under leave-ten-out cross-validation. The primary challenge in employing BHM is the high-dimensional aspects related to parameter estimation and in evaluating model convergence, but statistical packages available in R and STAN provide various tools to address such challenges. Renard (2011), Lima and Lall (2010), and Renard et al. (2013) offer perspectives on BHM application for hydroclimatic forecasting. The main advantage with BHM approach is that it generalizes the traditional regression by explicitly considering the cross-correlation structure across the predictands and by considering predictors available at multiple levels (e.g., gridded and at site) for regional hydroclimatic applications. It also could be employed for converting gridded estimates to point estimates and vice versa for developing regional forecast products.

Typically, the BHM is formulated similar to the traditional regression model, but the predictors available under different spatiotemporal scales are related. Given the interest in forecasting the predictand, $y_{i,t}$ (e.g., log of the streamflow), at i th location for the time step t using the predictors, $x_{i,t}$ (e.g., log of the streamflow over the previous time step), available at the same location i , and another set of predictors, $z_{j|i,j,t}$ (e.g., gridded precipitation forecasts), over a larger grid point, j , influencing the predictand at i th location, then a partial pooling model can be formulated as

$$y_{i,t} \mid \alpha_i, \beta_i, \gamma_{j[i]} = \alpha_i + \beta_i x_{i,t} + \gamma_{j[i]} z_{j[i],t} + \varepsilon_{i,t} \quad (8)$$

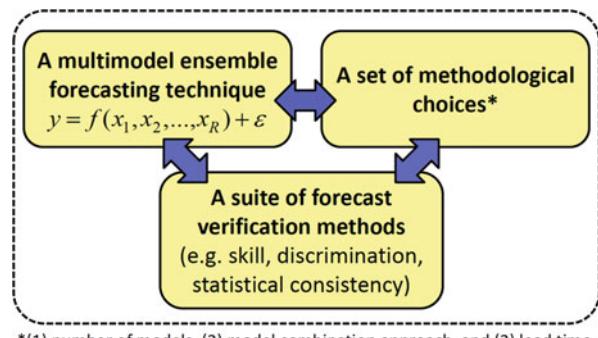
where the parameters α , β , and γ are specified as prior distributions and the residual $\varepsilon_{i,t}$ is modeled as multivariate normal distribution. For additional details on modeling the priors, see Chen et al. (2014) or Devineni et al. (2013). Here the predictors' influence on the predictand is captured at two levels – with the level 1 captured by site-specific predictors and level 2 captured by the large-scale climatic conditions – in developing the forecast. Since the parameter estimation is typically performed across the region using MCMC techniques, the prior distribution of parameters and residual terms also preserve cross-covariance structure across the predictand. In case the multilevel structure collapses with all the predictors available at the same location as that of the predictand, then the BHM results in a structure similar to BJP (described in Sect. 4.4). Thus, the hierarchical structure provides the flexibility in incorporating the predictors available at different spatiotemporal scales in estimating the forecast pdf.

4.6 Challenges in the Design of Multi-model Systems

Although many authors have reported the benefits of using multi-model combination approaches instead of the “best” single model in forecasting applications, it often far from a trivial challenge to select a suite of models to use and a method for combining their outputs. Hagedorn et al. (2005) showed that the suite of performance measures chosen to compare the single best model with several multi-model configurations can influence the ranking of the best approach. This implies that the choice of model forecast membership in a multi-model ensemble will depend on the performance or quality attributes that the modelers and water managers seek to achieve.

Expanding this line of argument, Mendoza et al. (2014) examined the impact of design choices on the performance of multi-model forecast configurations (Fig. 4). These included decisions about forecast quality attributes and weighting methods and the number of models to include, when facing different forecast skill (predictability) situations, which are proxied in the study by forecast lead time. The analysis

Fig. 4 Elements in the multi-model design experiment used by Mendoza et al. (2014) for understanding objective model selection and combination approaches



*(1) number of models, (2) model combination approach, and (3) lead time

demonstrated that such implementation choices can have a large impact on the identification of an optimal approach. For example, using probabilistic verification criteria may lead to different choices regarding the number of models or the multi-model combination method versus using deterministic metrics.

5 Discussion

Seasonal forecasting differs from short- to medium-range forecasting in that consistent retrospective hindcasts are less frequently available, and when they are, the size of the sample provided for training statistical post-processing methods is orders of magnitude smaller. For example, 3 years of daily-updated past short-range forecasts offer a nominal sample of over 1000 records of forecast performance for training the post-processing and the possibility of estimating a sufficient number of parameters to account for different predictability regimes. In contrast, 30 years of hindcasts of a particular seasonal prediction, in which predictability regimes may shift dramatically based on hydroclimatic seasonality, may only offer a sample size of 30, making it difficult to estimate more than a few parameters in a statistical post-processing model. Many state of the science operational inputs to seasonal forecasts (such as the NMME or other global climate forecast resources) only extend back to the beginning of the satellite era (around 1980), limiting the potential to create substantially longer hindcast series. As a result, post-processing parameters are often estimated with great uncertainty, which makes cross-validation of the post-processed a critical step to judge whether the post-processing has added significant marginal skill benefits relative to the raw forecasts. Another implication of the reduced ability to specify models is that simpler models, even those that merely bias-correct rather than calibrate the raw forecasts, may be more supportable given the data available; this is especially true in low-skill situations.

It also affects the ability to confidently estimate multi-model forecast combination parameters, despite the availability of adequate statistical techniques to do so. Mendoza et al. (2017), for instance, found that the default approach of equally weighting forecast ensembles can outperform more complex multi-model combination methods (e.g., BMA, QMA), supporting previous findings by Najafi and Moradkhani (2015). They also highlighted sample size as a serious limitation for the training process and the need for rigorous cross-validation to avoid overconfident solutions. Figure 5 illustrates this small-sample effect on statistical prediction model parameter instability using an example of an empirical seasonal inflow forecast based on climate predictors. The predictors selected and their weights for each of the cross-validation training samples (with 3 years of validation records left out during each training) vary significantly. This variation, in an operational application, adds uncertainty (noise) to the forecast, which may not be recognized if cross-validation steps have not been diligently applied.

Post-processing of any type of hydrological forecast, either short or long range, depends on assumptions of stationarity in climate, weather patterns, and hydrologic response. For several decades, evidence has grown that stationarity is a poor

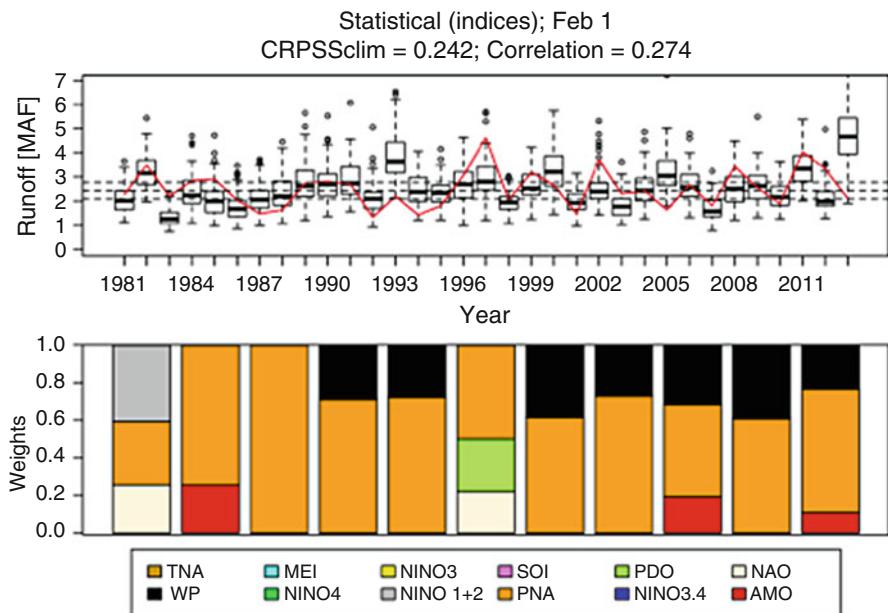


Fig. 5 Illustration of sample-driven instability in the selection and weighting of predictors in a cross-validation (leave-3-years-out) of seasonal streamflow predictions for Dworshak Reservoir, Idaho. Forecasts are made on February 1 for April–July period inflows based only on climate indices (listed at bottom). (Top) Hindcast series with observations (red line). (Bottom) Selected predictors and their weights for each test period using a Bayesian inference criteria during forward model selection. See Mendoza et al. (2017) for more details on the experiments

assumption about climate, given either anthropologically forced trends in the climate system or pronounced low-frequency variability in some regions that imparts apparent trends to the observed historical records. In addition, much of the Earth's land surface area is undergoing large-scale alteration due to human development activities (such as forest clearing, agriculture, and urban expansion), as well as due to natural causes, such as wildfire or vegetation die-off resulting from large-scale beetle infestations. Given the limitations of existing meteorological and hydrological monitoring networks and observations to quantify the impacts of such changes, it is common to ignore many potential forms of non-stationarity in designing and implementing both forecasting and post-processing approaches – particularly if hydroclimate and land surface alterations are not deemed severe. Consequently, such phenomena can contribute systematic errors to post-processed forecasts, leading to greater uncertainty than is expected.

This chapter provides background on common approaches to seasonal streamflow forecasting and discusses briefly our underlying of predictability in the seasonal hydrologic context. It also summarizes a variety of post-processing techniques that have been applied to seasonal streamflow predictions and the inherent challenge of specifying post-processing models given the sample size limitations of

seasonal forecasting. It is not possible to recommend any particular method for a given application, but the forecast designer is strongly encouraged to select methods through the diligent application of cross-validation and benchmarking and to prefer simpler methods where the marginal benefits of more complex, parameter-intensive methods cannot be confidently demonstrated.

References

- J. Beckers, A. Weerts, E. Tijdeman, E. Welles, ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction. *Hydrol. Earth Syst. Sci.* **20**, 3277–3287 (2016). <https://doi.org/10.5194/hess-20-3277-2016>
- A.A. Berg, K.A. Mulroy, Streamflow predictability in the Saskatchewan/Nelson River basin given macroscale estimates of the initial soil moisture status. *Hydrol. Sci. J.* **51**(4), 642–654 (2006). <https://doi.org/10.1623/hysj.51.4.6422006>
- K.J. Beven, Prophecy, reality and uncertainty in distributed hydrological modelling. *Adv. Wat. Resour.* **16**, 41–51 (1993)
- T.J. Bohn, M.Y. Sonessa, D.P. Lettenmaier, Seasonal hydrologic forecasting: do multimodel ensemble averages always yield improvements in forecast skill? *J. Hydrometeorol.* **11**(6), 1358–1372 (2010)
- C. Bracken, B. Rajagopalan, J. Prairie, A multisite seasonal ensemble streamflow forecasting technique. *Water Resour. Res.* **46**, W03532 (2010). <https://doi.org/10.1029/2009WR007965>
- A.A. Bradley, M. Habib, S.S. Schwartz, Climate index weighting of ensemble streamflow forecasts using a simple Bayesian approach. *Water Resour. Res.* **51**, 7382–7400 (2015). <https://doi.org/10.1002/2014WR016811>
- R.J.C. Burnash, R.L. Fernal, R.A. McGuire, *A Generalized Streamflow Simulation System – Conceptual Modeling for Digital Computers* (U.S. Department of Commerce National Weather Service and State of California Department of Water Resources, Sacramento, 1973)
- X. Chen, Z. Hao, N. Devineni, U. Lall, Climate information based streamflow and rainfall forecasts for Huai River basin using hierarchical Bayesian modeling. *Hydrol. Earth Syst. Sci.* **18**, 1539–1548 (2014). <https://doi.org/10.5194/hess-18-1539-2014>
- M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The schaake shuffle: A method for reconstructing space-time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262 (2004). [https://doi.org/10.1175/1525-7541\(2004\)005](https://doi.org/10.1175/1525-7541(2004)005)
- L. Crochemore, M.-H. Ramos, F. Pappenberger, Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.* **20**, 3601–3618 (2016). <https://doi.org/10.5194/hess-20-3601-2016>
- G. Day, Extended streamflow forecasting using NWSRFS. *J. Water. Res. Plan. Manag.* **111**(2), 157–170 (1985). [https://doi.org/10.1061/\(ASCE\)0733-9496\(1985\)111:2\(157\)](https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157))
- A. Dempster, N. Laird, D. Rubin, Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc.* **39**, 1–38 (1977)
- N. Devineni, A. Sankarasubramanian, Improving the prediction of winter precipitation and temperature over the continental United States: role of the ENSO state in developing multimodel combinations. *Mon. Weather Rev.* **138**(6), 2447–2468 (2010a). <https://doi.org/10.1175/2009MWR3112.1>
- N. Devineni, A. Sankarasubramanian, Improved categorical winter precipitation forecasts through multimodel combinations of coupled GCMs. *Geophys. Res. Lett.* **37**, L24704 (2010b). <https://doi.org/10.1029/2010GL044989>
- N. Devineni, A. Sankarasubramanian, S. Ghosh, Multimodel ensembles of streamflow forecasts: role of predictor state in developing optimal combinations. *Water Resour. Res.* **44**, W09404 (2008). <https://doi.org/10.1029/2006WR005855>

- N. Devineni, U. Lall, N. Pederson, E. Cook, A tree ring based reconstruction of Delaware River basin streamflow using hierarchical Bayesian regression. *J. Clim.* **26**, 4357–4374 (2013). <https://doi.org/10.1175/JCLI-D-11-00675.1>
- Q. Duan, N.K. Ajami, X. Gao, S. Sorooshian, Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**(5), 1371–1386 (2007). <https://doi.org/10.1016/J.ADVWATRES.2006.11.014>
- D.C. Garen, Improved techniques in regression-based streamflow volume forecasting. *J. Water Resour. Plan. Manag.* **118**, 654–670 (1992). [https://doi.org/10.1061/\(ASCE\)0733-9496](https://doi.org/10.1061/(ASCE)0733-9496)
- K.P. Georgakakos, D.-J. Seo, H. Gupta, J. Schaake, M.B. Butts, Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.* **298**(1–4), 222–241 (2004). <https://doi.org/10.1016/j.jhydrol.2004.03.037>
- K. Grantz, B. Rajagopalan, M. Clark, E. Zagona, A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resour. Res.* **41**, W10410 (2005). <https://doi.org/10.1029/2004WR003467>
- W. Greuell, W.H.P. Franssen, R.W.A. Hutjes, Seasonal streamflow forecasts for Europe – II. Explanation of the skill. *Hydrol. Earth Syst. Sci. Discuss.* (2016). <https://doi.org/10.5194/hess-2016-604>. in review
- R. Hagedorn, F. Doblas-Reyes, T. Palmer, The rationale behind the success of multimodel ensembles in seasonal forecasting I. Basic concept. *Tellus. Ser.A* **57**, 219–233 (2005)
- A.F. Hamlet, D.P. Lettenmaier, Columbia River streamflow forecasting based on ENSO and PDO climate signals. *J. Water Resour. Plan. Manag.* **125**(6), 333–341 (1999)
- S. Harrigan, C. Prudhomme, S. Parry, K. Smith, M. Tanguy, Benchmarking ensemble streamflow prediction skill in the UK. *Hydrol. Earth Syst. Sci. Discuss.* (2017). <https://doi.org/10.5194/hess-2017-449>. in review
- T. Hashino, A.A. Bradley, S.S. Schwartz, Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.* **11**, 939–950 (2007)
- D. Helms, S.E. Phillips, P.F. Reich, *The History of Snow Survey and Water Supply Forecasting*. Natl. Bull. 290-9-6 (Natural Resources Conservation Service, U.S. Department of Agriculture, Washington, DC, 2008)
- J.M. Hidalgo-Muñoz, S.R. Gámiz-Fortis, Y. Castro-Díez, D. Argüeso, M.J. Esteban-Parra, Long-range seasonal streamflow forecasting over the Iberian Peninsula using large-scale atmospheric and oceanic information. *Water Resour. Res.* **51**(5), 3543–3567 (2015). <https://doi.org/10.1002/2014WR016826>
- F. Hoss, P.S. Fischbeck, Performance and robustness of probabilistic river forecasts computed with quantile regression based on multiple independent variables. *Hydrol. Earth Syst. Sci.* **19**, 3969–3990 (2015). <https://doi.org/10.5194/hess-19-3969-2015>
- B.P. Kirtman, D. Min, J.M. Infant, J.L. Kinter, D.A. Paolino, Q. Zhang, H. van den Dool, S. Saha, M.P. Mendez, E. Becker, P. Peng, P. Tripp, J. Huang, D.G. DeWitt, M.K. Tippett, A.G. Barnston, S. Li, A. Rosati, S.D. Schubert, M. Rienecker, M. Suarez, Z.E. Li, J. Marshak, Y. Lim, J. Tribbia, K. Pegion, W.J. Merryfield, B. Denis, E.F. Wood, The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; Phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.* **95**, 585–601 (2014). <https://doi.org/10.1175/BAMS-D-12-00050.1>
- R.D. Koster, S. Mahanama, Land surface controls on hydroclimatic means and variability. *J. Hydrometeorol.* **13**, 1604–1620 (2012)
- T. Krishnamurti, C. Kishtawal, Z. Zhang, T. LaRow, D. Bachiochi, E. Williford, Multimodel ensemble forecasts for weather and seasonal climate. *J. Clim.* **13**, 4196–4216 (2000)
- F. Lehner, A.W. Wood, D. Llewellyn, D.B. Blatchford, A.G. Goodbody, F. Pappenberger, Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the U.S. southwest. *Geophys. Res. Lett.* **44**, 12,208 (2017). <https://doi.org/10.1002/2017GL076043>
- C.H. Lima, U. Lall, Spatial scaling in a changing climate: a hierarchical bayesian model for non-stationary multi-site annual maximum and monthly streamflow. *J. Hydrol.* **383**(3), 307–318 (2010)
- R. Linsley, N. Crawford, Continuous simulation models in urban hydrology. *Geophys. Res. Lett.* **1**, 59–62 (1974). <https://doi.org/10.1029/GL001i001p00059>
- D. Lucatero, H. Madsen, J.C. Refsgaard, J. Kidmose, K.H. Jensen, Seasonal streamflow forecasts in the Ahlergaard catchment Denmark: effect of preprocessing and postprocessing on skill and statistical consistency. *Hydrol. Earth Syst. Sci. Discuss.* (2017). <https://doi.org/10.5194/hess-2017-379>. in review

- P.A. Mendoza, B. Rajagopalan, M.P. Clark, G. Cortes, J. McPhee, A robust multimodel framework for ensemble seasonal hydroclimatic forecasts. *Water Resour. Res.* **50**, 6030 (2014). <https://doi.org/10.1002/2014WR015426>
- P.A. Mendoza, A.W. Wood, E.A. Clark, E. Rothwell, M.P. Clark, B. Nijssen, L.D. Brekke, J.R. Arnold, An intercomparison of approaches for improving predictability in operational seasonal streamflow forecasting. *Hydrol. Earth Syst. Sci.* **21**, 3915–3935 (2017)
- H. Moradkhani, M. Meier, Long-lead water supply forecast using large-scale climate predictors and independent component analysis. *J. Hydrol. Eng.* **15**(10), 744–762 (2010). [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000246](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000246)
- M. Najafi, H. Moradkhani, Ensemble combination of seasonal streamflow forecasts. *J. Hydrol. Eng.* **21**(1), 04015043 (2015). [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0001250](https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250)
- S. Opitz-Stapleton, S. Gangopadhyay, B. Rajagopalan, Generating streamflow forecasts for the Yakima River Basin using large-scale climate predictors. *J. Hydrol.* **341**(3–4), 131–143 (2007). <https://doi.org/10.1016/j.jhydrol.2007.03.024>
- T.C. Pagano, D.C. Garen, T.R. Perkins, P.A. Pasteris, Daily updating of operational statistical seasonal water supply forecasts for the Western U.S. *J. Am. Water Resour. Assoc.* **45**(3), 767–778 (2009). <https://doi.org/10.1111/j.1752-1688.2009.00321.x>
- T. Pagano, A.W. Wood, K. Werner, R. Tama-Sweet, Western U.S. water supply forecasting: a tradition evolves. *Eos. Trans. AGU* **95**(3), 28 (2014)
- T. Piechota, F. Chiew, Seasonal streamflow forecasting in eastern Australia and the El Niño–southern oscillation. *Water Resour. Res.* **34**(11), 3035–3044 (1998)
- T.C. Piechota, F.H.S. Chiew, J.A. Dracup, T.A. McMahon, Development of exceedance probability streamflow forecast. *J. Hydrol. Eng.* **6**(1), 20–28 (2001)
- D. Raff, L. Brekke, K.V. Werner, A. Wood, K. White, *Short-Term Water Management Decisions: User Needs for Improved Climate, Weather, and Hydrologic Information*. Technical Report CWTS-2013-1 (Bureau of Reclamation U.S. Army Corps of Engineers and National Oceanic and Atmospheric Administration, Denver, USA, 2013)
- A.E. Raftery, T. Gneiting, F. Balabdaoui, M. Polakowski, Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**, 1155–1174 (2005)
- B. Rajagopalan, U. Lall, A k-nearest-neighbor simulator for daily precipitation and other weather variables. *Water Resour. Res.* **35**(10), 3089–3101 (1999). <https://doi.org/10.1029/1999WR900028>
- B. Rajagopalan, U. Lall, S. Zebiak, Optimal categorical climate forecasts through multiple GCM ensemble combination and regularization. *Mon. Weather Rev.* **130**(7), 1792–1811 (2002)
- S.K. Regonda, B. Rajagopalan, M. Clark, E. Zagona, A multi-model ensemble forecast framework: Application to spring seasonal flows in the Gunnison River Basin. *Water Resour. Res.* **42**, W09404 (2006). <https://doi.org/10.1029/2005WR004653>
- B. Renard, A Bayesian hierarchical approach to regional frequency analysis. *Water Resour. Res.* **47**, W11513 (2011). <https://doi.org/10.1029/2010WR010089>
- B. Renard, X. Sun, M. Lang, Bayesian methods for non-stationary extreme value analysis, in *Extremes in a Changing Climate* (Springer Netherlands, 2013), pp. 39–95
- D.E. Robertson, P. Pokhrel, Q.J. Wang, Improving statistical forecasts of seasonal streamflows using hydrological model output. *Hydrol. Earth Syst. Sci.* **17**, 579–593 (2013). <https://doi.org/10.5194/hess-17-579-2013>
- E.A. Rosenberg, A.W. Wood, A.C. Steinemann, Statistical applications of physically based hydrologic models to seasonal streamflow forecasts. *Water Resour. Res.* **47**, W00H14 (2011). <https://doi.org/10.1029/2010WR010101>
- E.A. Rosenberg, A.W. Wood, A.C. Steinemann, Informing hydrometric network design for statistical seasonal streamflow forecasts. *J. Hydrometeorol.* **14**, 1587–1604 (2013). <https://doi.org/10.1175/JHM-D-12-0136.1>
- J.D. Salas, C. Fu, B. Rajagopalan, Long-range forecasting of Colorado streamflows based on hydrologic atmospheric and oceanic data. *J. Hydrol. Eng.* **16**(6), 508–520 (2011). [https://doi.org/10.1061/\(ASCE\)HE.1943-5584.0000343](https://doi.org/10.1061/(ASCE)HE.1943-5584.0000343)
- A. Sankarasubramanian, U. Lall, Flood quantiles in a changing climate: seasonal forecasts and causal relations. *Water Resour. Res.* **39**(5), 1134 (2003). <https://doi.org/10.1029/2002WR001593>

- R. Schefzik, A similarity-based implementation of the Schaake shuffle. *Mon. Weather Rev.* **144**, 1909–1921 (2016). <https://doi.org/10.1175/MWR-D-15-0227.1>
- A. Schepen, Q.J. Wang, Model averaging methods to merge operational statistical and dynamic seasonal streamflow forecasts in Australia. *Water Resour. Res.* **51**, 1797 (2015). <https://doi.org/10.1002/2014WR016163>
- A. Schepen, Q.J. Wang, Y. Everingham, Calibration, bridging, and merging to improve GCM seasonal temperature forecasts in Australia. *Mon. Wea. Rev.* **144**, 2421–2441 (2016). <https://doi.org/10.1175/MWR-D-15-0384.1>
- A. Schepen, T. Zhao, Q.J. Wang, D.E. Robertson, A Bayesian modelling method for post-processing daily sub-seasonal to seasonal rainfall forecasts from global climate models and evaluation for 12 Australian catchments. *Hydrol. Earth Syst. Sci.* **22**, 1615–1628 (2018). <https://doi.org/10.5194/hess-22-1615-2018>
- D.-J. Seo, H. Herr, J. Schaake, A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discuss.* **3**, 1987–2035 (2006)
- L.J. Slater, G. Villarini, A.A. Bradley, et al., *Clim. Dyn.* (2017). <https://doi.org/10.1007/s00382-017-3794-7>
- S. Sorooshian, Q. Duan, V.K. Gupta, Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture accounting model. *Water Resour. Res.* **29**, 1185–1194 (1993)
- F.A. Souza Filho, U. Lall, Seasonal to interannual ensemble streamflow forecasts for Ceara Brazil: applications of a multivariate semiparametric algorithm. *Water Resour. Res.* **39**(11), 1307 (2003). <https://doi.org/10.1029/2002WR001373>
- G.A. Tootle, A.K. Singh, T.C. Piechota, I. Farnham, Long lead-time forecasting of U.S. streamflow using partial least squares regression. *J. Hydrol. Eng.* **12**, 442–451 (2007)
- R.D. Valencia, J.C. Schakke Jr., Disaggregation processes in stochastic hydrology. *Water Resour. Res.* **9**(3), 580–585 (1973). <https://doi.org/10.1029/WR009i003p00580>
- A. Verdin, B. Rajagopalan, W. Kleiber, G. Podestá, F. Bert, A conditional stochastic weather generator for seasonal to multi-decadal simulations. *J. Hydrol.* (2015). <https://doi.org/10.1016/j.jhydrol.2015.12.036>
- T. Wagener, N. McIntyre, M.J. Lees, H.S. Wheater, H.V. Gupta, Towards reduced uncertainty in conceptual rainfall-runoff modelling: Dynamic identifiability analysis, *Hydrol. Processes.* **17**(2), 455–476 (2003)
- Q.J. Wang, D.E. Robertson, F.H.S. Chiew, A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.* **45**(5), 1–18 (2009). <https://doi.org/10.1029/2008WR007355>
- H. Wang, A. Sankarasubramanian, R.S. Ranjithan, Integration of climate and weather information for improving 15-day-ahead accumulated precipitation forecasts. *J. Hydrometeorol.* **14**(1), 186–202 (2013)
- A.P. Weigel, M.A. Liniger, C. Appenzeller, Can multi-model combination really enhance the prediction skill of probabilistic ensemble forecasts? *Q. J. R. Meteorol. Soc.* **134**(630), 241–260 (2008)
- K. Werner, D. Brandon, M. Clark, S. Gangopadhyay, Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *J. Hydrometeor.* **5**, 1076–1090 (2004). <https://doi.org/10.1175/JHM-381.1>
- S. Westra, A. Sharma, C. Brown, U. Lall, Multivariate streamflow forecasting using independent component analysis. *Water Resour. Res.* **44**(2), 1–11 (2008). <https://doi.org/10.1029/2007WR006104>
- A.W. Wood, D.P. Lettenmaier, A new approach for seasonal hydrologic forecasting in the western U.S. *Bull. Amer. Met. Soc.* **87**(12), 1699–1712 (2006). <https://doi.org/10.1175/BAMS-87-12-1699>
- A.W. Wood, D.P. Lettenmaier, An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophys. Res. Lett.* **35**, L14401 (2008). <https://doi.org/10.1029/2008GL034648>

- A.W. Wood, J.C. Schaake, Correcting errors in streamflow forecast ensemble mean and spread. *J. Hydrometeorol.* **9**, 132–148 (2008)
- A.W. Wood, T. Hopson, A. Newman, L. Brekke, J. Arnold, M. Clark, Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeorol.* **17**, 651–668 (2016a). <https://doi.org/10.1175/JHM-D-14-0213.1>
- A.W. Wood, T. Pagano, M. Roos, Tracing the origins of ESP HEPEX historical hydrology series edition 1 (online at: <https://hepex.irstea.fr/tracing-the-origins-of-esp/>) (2016b)
- T. Zhao, J.C. Bennett, Q.J. Wang, A. Schepen, A.W. Wood, D.E. Robertson, M. Ramos, How suitable is quantile mapping for post processing GCM precipitation forecasts? *J. Clim.* **30**, 3185–3196 (2017). <https://doi.org/10.1175/JCLI-D-16-0652.1>

Part VIII

Verification of Hydrometeorological Ensemble Forecasts



Attributes of Forecast Quality

A. Allen Bradley, Julie Demargne, and Kristie J. Franz

Contents

1	Introduction	850
2	Ensemble Verification Process	851
3	Mathematical Formulation of the Verification Process	853
4	Aspects of Forecast Quality	855
5	Common Measures of Forecast Quality with Distribution Moments	858
5.1	Calibration-Refinement Decomposition	858
5.2	Likelihood-Base Rate Decomposition	860
5.3	Skill Score and Its Decompositions	862
6	Forecast Quality of Hypothetical Ensemble Forecasts	864
6.1	Illustration of Forecast Quality for a Single-Valued Forecast	865
6.2	Illustration of Forecast Quality for a Probability Forecast for a Discrete Event	869
6.3	Illustration of Forecast Quality for an Ensemble Probability Distribution Forecast	877
7	Practical Considerations	885
7.1	Diagnostic Verification and Ensemble Forecasts	885
7.2	Absolute Versus Relative Measures	886
7.3	Choosing a Verification Data Sample	887
7.4	Sampling Uncertainties for Small Data Samples	888
8	Conclusions	889
	References	890

A. A. Bradley (✉)

IIHR–Hydroscience and Engineering, The University of Iowa, Iowa City, IA, USA

e-mail: allen-bradley@uiowa.edu

J. Demargne (✉)

HYDRIS Hydrologie, Saint Mathieu de Tréviers, France

e-mail: julie.demargne@hydris-hydrologie.fr; julie@demargne.com

K. J. Franz (✉)

Department of Geological and Atmospheric Sciences, Iowa State University, Ames, IA, USA

e-mail: kfranz@iastate.edu

Abstract

Forecast verification is a process used to assess the quality of hydrometeorological ensemble forecasts. This chapter describes the many aspects of forecast quality using a distributions-oriented approach. Using the joint distribution of forecasts and observations, or one of its factorizations into a conditional and marginal distribution, the aspects of forecast quality are defined. Hypothetical ensemble forecasts are then used to illustrate aspects of forecast quality. The hypothetical ensemble forecasts are used to construct single-valued forecasts, probability forecasts for an event, and ensemble probability distribution forecasts. Their forecast quality is then diagnosed using visual comparisons and numerical comparisons of forecast quality measures. The examples illustrate that a single aspect of forecast quality is insufficient and that many aspects are needed to understand the nature of the forecasts. Some practical considerations in the application of the framework to ensemble forecast verification are discussed.

Keywords

Forecast verification · Ensemble forecasts · Deterministic forecasts · Probabilistic forecasts · Distributions-oriented approach

1 Introduction

Everyone wants forecasts to be good – from the forecasters who issue them to those who use them to make decisions. In part, what makes a forecast “good” depends on one’s point of view and purpose with respect to the forecast (Murphy 1993). Murphy (1993) describes three types of forecast “goodness”: consistency (Type 1), quality (Type 2), and value (Type 3). Consistency means that the forecast corresponds to a forecaster’s best judgment and true state of knowledge. This is often assumed as a forecast system that should provide the best and most appropriate information to forecast users. Quality is the ability of the forecast to predict an event well according to some objective criteria (Murphy 1993). Value refers to the benefits realized by the forecast users when the forecast is used to inform their decision making. Evaluating consistency and value inherently requires consideration of the forecaster and user, and is dependent upon the situation under which the forecast is made or used. Quality, in contrast, is assessed objectively by comparing predicted events with the corresponding observations (Brier and Allen 1951; Murphy and Winkler 1987; Wilks 2011; Jolliffe and Stephenson 2011). Forecast verification is the process of assessing the quality of forecasts (Murphy and Winkler 1987).

Fundamentally, forecast verification compares forecasts with observations. The comparison can be as simple as a plot of a sample of forecasts and observations. But more often it produces one or more indices or scores from which the forecast–observation pair or pairs can be interpreted (Brier and Allen 1951). Forecast verification is essential to any forecasting system (Murphy and Winkler 1987). It is the means by which the performance of forecasting methods can be documented and

compared to each other or against standards. Forecast verification provides guidance for improving forecast systems. It is a tool for analyzing forecasts under different conditions (e.g., low vs. high flows, specific seasons) and assessing how suitable forecasts might be for different applications (e.g., droughts or floods) (Brier and Allen 1951; Murphy and Winkler 1987; Murphy 1993; Welles et al. 2007).

Assessing forecast quality involves the statistical examination of the joint distribution of forecasts and observations, with higher quality forecasts displaying a greater degree of correspondence with the observations (Murphy and Winkler 1987). There are many scores that can be used to quantify different aspects of forecast quality. A variety of different scores should be used because a single score cannot describe all the properties of the relationship between forecasts and observation (Murphy 1993, 1997). Additionally, no one set of scores is suitable to all applications. Depending on the forecast type, the forecast application, and the expertise of the user, a set of scores can be chosen to answer questions of interest and give the user information upon which he/she can make decisions (Joliffe and Stephensen 2011).

The chapter is organized as follows. Section 2 describes the ensemble forecast verification process and defines the common forms of forecasts (deterministic and probabilistic) and observations (discrete and continuous). Section 3 reviews the distributions-oriented approach for forecast verification, and the role of the joint distribution of forecasts and observations in the verification processes. Section 4 defines the various aspects of forecast quality using the distributions-oriented framework. Section 5 defines common measures of forecast quality using the moments of the joint distribution. Section 6 illustrates aspects of forecast quality for alternative hypothetical ensemble forecasts. The ensemble forecasts are evaluated in three ways: as single-valued forecasts using the ensemble mean, as probability forecasts for a discrete event occurrence, and as ensemble probability distribution forecasts of the outcome. The forecast quality for each is diagnosed using visual comparisons and numerical comparisons of forecast quality measures. Section 7 discusses some practical considerations in the application of the framework to ensemble forecast verification. Section 8 concludes the chapter.

2 Ensemble Verification Process

Forecasts come in many different forms. In a broad sense, a forecast can be either *deterministic* or *probabilistic*. Deterministic forecasts consist of unqualified statements that a single outcome will occur (single-valued forecast with no statement of uncertainty) (Wilks 2011). Probabilistic forecasts express the degree of uncertainty that an outcome or set of possible outcomes will occur. The observation corresponding to the forecast can be either *discrete* or *continuous* (Potts 2011). An observation is discrete when the outcome being predicted can have only a limited set of possible values (e.g., a flood/no flood outcome has a discrete observation). An observation is continuous when the outcome can have an infinite number of possible values (e.g., a spring flood volume outcome can have a continuous observation).

Ensemble forecasts consist of multiple estimates of some future event. Most commonly, hydrologic ensembles are a set of streamflow time series. A forecast is derived from the ensemble by taking a summary measure, subsampling, thresholding, categorizing, or assigning a probability distribution to the forecast data (Potts, 2011). One of the most common approaches is to choose the mean or median of the ensemble, thereby converting an ensemble prediction into a deterministic forecast. However, this removes all measures of uncertainty from the forecast and potentially useful information contained in the ensemble (Franz et al. 2003; Demargne et al. 2010). Dichotomous forecasts are another type of deterministic forecasts that make statements about whether an event will happen or won't happen. Dichotomous forecasts can be derived from the ensemble through thresholding, such as flood or no flood. Another approach to thresholding ensembles is to make statements about the possibility of a flood occurring or not occurring by determining the frequency with which the ensemble members predict either category, i.e., a probabilistic forecast.

Probabilistic forecasts provide explicit statements of uncertainty regarding how well the future hydrology is known (Krzysztofowicz 2001; Wilks 2011). Probability distribution forecasts are generated by assuming an appropriate distribution for the ensemble members. A forecast based on the empirical cumulative distribution is a direct measurement of the sample, or ensemble. A forecast based on a parametric distribution uses a characteristic shape with specified distribution parameters that describe the magnitude and variation of the variable of interest (Wilks 2011). In this case, the distribution parameters represent properties of the population, although they are estimated from the sample.

There are a variety of ways to express the continuous probability distribution function of the outcome being predicted. Probability distribution forecasts are often displayed as cumulative frequency distribution functions or exceedance probability forecasts, which describe the probability that the event will be less than a given value. Although hydrologic ensembles and observations are in essence discrete variables with a finite number of significant figures, in practice, these data are often treated as continuous variables and the forecasts are displayed as a smooth cumulative distribution function. An empirical distribution, which is a step function with probability jumps of $1/n$ at each of the n data points, can be transformed to a smooth line through kernel smoothing (Wilks 2011).

Often it is necessary or convenient to convert the ensemble traces to a set of discrete variables through categorizing or defining event thresholds. Categorical predictands are discrete predictands that can take on only one of a finite set of values (Potts 2011). Multicategory probabilistic forecasts can be created by approximating the continuous probability distribution function with quantiles, probability categories, or event thresholds (Krzysztofowicz 2001; Potts 2011). For example, Franz et al. (2003) evaluated seasonal water supply ensemble hindcasts using five forecast categories with threshold values based on the historical streamflow record, specifically 0–10%, >10–30%, >30–70%, >70–90%, and > 90% of the climatology. Ensemble members are assigned to bins or categories, and the frequency of the ensemble members in each bin is used to compute a probabilistic forecast.

The forecast verification process starts with the collection of well-defined forecasts and observations. As the preceding discussion illustrates, ensemble forecast information can be used in different ways and by different users, so one could choose alternative forms of a forecast for verification (Potts 2011). The simplest types of forecasts constructed from ensemble predictions are ones where the forecast is a *single number* (either deterministic or probabilistic) and the observation is a *single number* (either discrete or continuous). One example of an elemental forecast would be a deterministic flow volume forecast; the mean ensemble flow volume is used as a single-valued forecast, and the observation is the future flow volume. Another example would be a probabilistic flood event forecast; the forecast is the probability of a flood occurrence (a number from 0 to 1), and the observation is whether the flood occurred or not (represented by a discrete value of 1 if a flood occurs, or 0 if it does not). Verification of elemental forecasts like these is straightforward. However, given the information available in an ensemble forecast, other (more complex) forms are frequently constructed as well. For instance, probability forecasts for multiple categories constructed from ensembles, the forecast is in the form of probability distribution function for the discrete category outcomes. For probability forecasts of a continuous outcome, the forecast is in the form of a probability distribution of a continuous variable. Verification of forecasts in these forms (where the forecast itself is no longer a single number) is much more challenging.

Regardless of the form of the ensemble forecast chosen for verification, once a sufficiently large sample of well-defined forecasts and observations has been collected, the verification sample can be used to explore the relationship between the forecast-observation pairs. Attributes of this relationship describe different aspects of the quality of the forecasts. In the next section, a general framework is presented for characterizing forecast quality based on the statistical relationship between forecasts and observation. The framework was developed for the elemental ensemble forecast verification problem – the case where the forecast is a single number (either deterministic or probabilistic) and the observation is a single number (either discrete or continuous). However, the aspects of forecast quality developed for the elemental case also have meaning for more complex probability distribution forecasts. As will be seen in the ensemble forecasting examples shown in Sect. 6, the framework can be extended to describe aspects of forecast quality for probability distribution forecasts from ensemble predictions.

3 Mathematical Formulation of the Verification Process

Murphy and Winkler (1987) introduced the *distributions-oriented (DO) approach* – also called diagnostic verification (Potts 2011) – as a general framework for forecast verification. The approach explicitly defines a stochastic process in which: (1) the forecast and observation are treated as random variables, and (2) each forecast-observation pair is assumed to be independent of all other pairs and identically distributed.

Consider the case where the forecast is a single number (either deterministic or probabilistic) and the observation is a single number (either discrete or continuous). Let F be the random variable denoting the forecast and X the random variable denoting the observation of the underlying variable of interest. Let f and x denote the numerical values of these two variables. The relationship between the forecast and the observation is thus defined by the *joint probability distribution function* (pdf) of the forecast variable and the corresponding observed variable:

$$p_{FX}(f, x) = \Pr(F = f, X = x). \quad (1)$$

In other words, for any forecast of this type – a single-valued deterministic forecast of a continuous flow variable, or a probability forecast (between 0 and 1) of a discrete event outcome (either 1 or 0) – the joint pdf contains all the relevant information about their relationship.

The *distributions-oriented approach* focuses on describing the characteristics of the joint distribution of forecasts and observations; it has been developed and used extensively for both probabilistic and deterministic meteorological forecast verification (see Bradley et al. 2003 for references). The *measures-oriented verification approach*, in contrast, does not focus explicitly on the joint distribution. It reduces the verification information to a small number of performance measures that describe specific aspects of the overall quality of forecasts (such as accuracy or skill).

Regarding the DO verification approach, the different aspects of forecast quality can be defined using the joint probability distribution or its two factorizations into marginal and conditional distributions (Murphy and Winkler 1987). By conditioning on the forecast f , the *calibration-refinement factorization* of the joint distribution is:

$$p_{FX}(f, x) = p_{X|F}(x | f) \times p_F(f). \quad (2)$$

By conditioning on the observation x , the *likelihood-base rate factorization* of the joint distribution is:

$$p_{FX}(f, x) = p_{F|X}(f | x) \times p_X(x). \quad (3)$$

The distributions $p_{X|F}(x | f)$ and $p_{F|X}(f | x)$ are the conditional distribution of the observations given the forecast f and the conditional distribution of the forecasts given the observation x , respectively. The distributions $p_F(f)$ and $p_X(x)$ are the marginal (unconditional) probability distributions of the forecasts and observations, respectively. The marginal distribution $p_F(f)$ is sometimes called the *predictive* distribution or *refinement* distribution (Wilks 2011). Forecast refinement refers to the dispersion of the forecast distribution $p_F(f)$. A forecast that only takes on a small range of values (e.g., it does not stray far from climatology) has a small spread of its refinement distribution $p_F(f)$; conversely, a forecast that takes on a wide range of values (e.g., it can stray far from climatology) has a large spread, and the forecast F has the potential to discern a broad range of conditions. The marginal distribution $p_X(x)$ is sometimes called the *base rate* or *uncertainty* distribution and characterizes

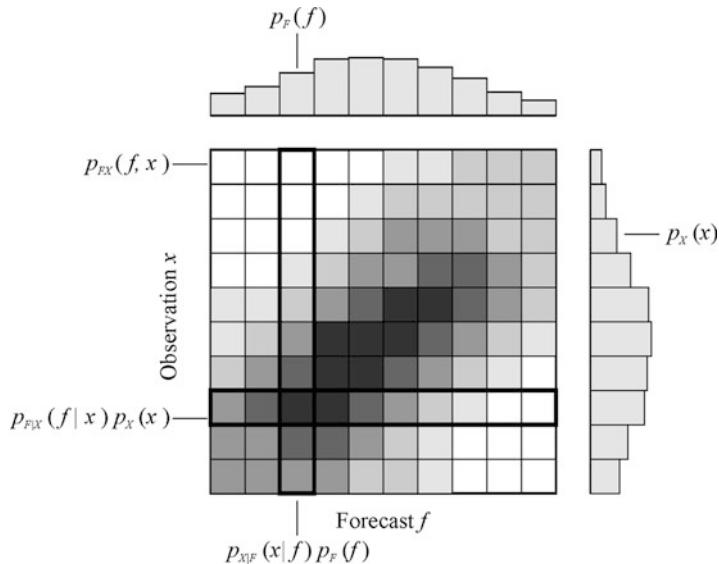


Fig. 1 Conceptual illustration of the joint distribution and its marginal and conditional distributions for discrete forecasts and observations. The joint distribution $p_{FX}(f, x)$ is represented by the *box*; each square represents the density for a specific discrete forecast f and observation x . The histogram *above the box* represents the marginal (unconditional) distribution of the forecast $p_F(f)$. The histogram *beside the box* represents the marginal (unconditional) distribution of the observation $p_X(x)$. The *highlighted column* within the joint distribution represents the product of the conditional distribution $p_{X|F}(x|f) \times p_F(f)$; hence, each column is related to the conditional distribution of a discrete forecast. The *highlighted row* represents the product of the conditional distribution $p_{F|X}(f|x) \times p_X(x)$; hence, each row is related to the conditional distribution of a discrete observation

the forecasting situation (i.e., the observation X of the quantity to be forecast), not the forecast itself. Figure 1 illustrates the joint distribution and its marginal and conditional conditions for a case where the forecast and observation are discrete random variables.

4 Aspects of Forecast Quality

The different aspects of the forecast quality are defined as follows (Wilks 2011; Jolliffe and Stephenson 2011; Murphy 1997):

Bias (or first order bias, overall bias, unconditional bias) describes the difference between the average forecast and the average observation. For example, if flow forecasts are consistently larger than observed flows, the deterministic forecast has a positive bias. If the forecast probabilities for a flood event are consistently less than the flood event's occurrence frequency, the probability forecast has a negative bias. Bias describes differences between the marginal (unconditional) distribution of

forecasts $p_F(f)$ and observations $p_X(x)$; it is a measure of their central tendency (such as the mean, median, or mode).

Accuracy describes the average difference between the individual forecasts and observations. Forecasts that consistently agree closely with the observed outcomes are accurate. It is common to hear an individual forecast described as “accurate” when its forecast error is small and “inaccurate” when its forecast error is large. But in the context of DO verification, accuracy refers to average characteristics of forecast errors (over many forecasts). Accuracy is a property defined by the entire joint distribution $p_{FX}(f, x)$.

Skill describes the accuracy of a forecast relative to a reference forecast or benchmark. Forecasts that are more accurate than the reference have skill (or are “skillful”). Generally the reference forecast corresponds to a more “naïve” forecast, such as (observed) climatology, persistence, or random chance. In other situations, the output from a baseline forecasting system serves as the reference. For example, at short lead times persistence (a forecast that the current condition will continue) can be accurate and is an important reference for comparison; forecasts must accurately predict short-range changes to be skillful (more accurate than persistence). At longer lead times, climatology (a forecast of the climatological outcome) is an important reference; forecasts must accurately predict long-range deviations from climatology to be skillful. (Note that climatology should be defined from a long record of observations, as it is a more stable estimate than one based on the sample climatology from the verification dataset.)

Association (or correlation) describes the strength of the linear relationship between the forecasts and observations. A forecast has good association if the observations are highly correlated with the forecasts. The correlation coefficient is a common measure of association. Correlation is a property defined by the entire joint distribution $p_{FX}(f, x)$.

Reliability or **type-1 conditional bias** describes how well the forecast agrees with the observed outcome on average when a specific forecast is issued. For a flood probability forecast to be reliable (or conditionally unbiased), a flood should be observed 20% of the time when a forecast probability of 0.2 is issued. A flow forecast is reliable if the average observed flow is 1000 m³/s when a forecast flow of 1000 m³/s has been issued. Measures of reliability evaluate this conditional bias for all possible forecasts issued; forecasts that are reliable for all possible forecasts are said to be well calibrated. Reliability is a property of the conditional distribution $p_{X|F}(x | f)$ and the marginal distribution $p_F(f)$ from the calibration-refinement (CR) factorization.

Resolution describes whether the outcomes are different for different forecasts issued (independent of whether or not the forecasts are reliable). For example, flood events should occur more often when a forecast probability of 0.8 is issued than when a forecast probability of 0.2 is issued. The average observed flow when a forecast of 1000 m³/s is issued should be higher than that when a forecast of 250 m³/s is issued. Measures of resolution evaluate this difference for all possible forecasts issued; when the forecasts can resolve the different outcomes, they are said to have

resolution. Resolution is also a property of the conditional distribution $p_{X|F}(x | f)$ and the marginal distribution $p_F(f)$ from the calibration-refinement (CR) factorization.

Type-2 conditional bias describes how well the observation agrees with the forecast on average when a specific outcome is observed. For all the cases when a flood event occurs, if the average forecast flood probability issued is 75%, the probability forecasts have type-2 conditional bias. For all the cases when a flow of 1000 m³/s is observed, if the average flow forecast issued is 1200 m³/s, the deterministic forecast has type-2 conditional bias. Measures of type-2 conditional bias evaluate this conditional bias for all possible outcomes. Type-2 conditional bias is a property of the conditional distribution $p_{F|X}(f | x)$ and the marginal distribution $p_X(x)$ from the likelihood-base rate (LBR) factorization.

Discrimination describes whether the forecasts are different for different outcomes. If forecast probabilities issued when a flood occurs tend to be higher than those issued when a flood does not occur, the probability forecasts have discrimination. If the flow forecasts issued when high flows occur tend to be higher than those issued when low flows occur, the deterministic forecasts have discrimination. Measures of discrimination evaluate this difference for all possible outcomes. Discrimination is also a property of the conditional distribution $p_{F|X}(f | x)$ and the marginal distribution $p_X(x)$ from the likelihood-base rate (LBR) factorization.

Sharpness describes the degree of variability of the forecasts. For a probability forecast, it indicates the tendency to predict with extreme probabilities (0 or 1). Probability forecast are said to be sharp if they issue probabilities close to 0 or 1; in contrast, a climatology forecast (which issues only the climatological event probability) is “unsharp.” For deterministic forecasts, the forecasts must deviate significantly from the mean forecast to be sharp. Sharpness is closely related to the notion of resolution although it only concerns the forecast (some authors have used the two terms, sharpness and resolution, as synonymous – see the discussion in Jolliffe and Stephenson 2011). A high degree of sharpness is only desirable in the context of other measures. Without reliability, a sharp forecast is misleading; however, for a given level of reliability, a sharp forecast is preferred over an “unsharp” one since it contributes less uncertainty to decision making (Gneiting et al. 2007). Sharpness is a property of the marginal (unconditional) distribution of forecasts $p_F(f)$ alone.

Uncertainty describes the degree of variability in the observations. It is an important aspect in the performance of a forecasting system, but is independent of the forecast. Uncertainty is most simply measured by the variance of the observations. Uncertainty is a property of the marginal (unconditional) distribution of observations $p_X(x)$ alone.

For any given aspect of forecast quality, there are many possible ways to assess it. Some involve graphical representation of statistics from the verification data sample (such as a reliability diagram or a rank histogram). Others involve metrics or measures of forecast quality. In the next section, measures based on the joint distribution of forecasts and observations are presented for the different aspects of forecast quality.

5 Common Measures of Forecast Quality with Distribution Moments

When the forecast F is a random variable (e.g., a single-valued deterministic flow or an event probability between 0 and 1) and the observation X is a random variable (a continuous flow variable or a discrete event outcome of either 1 or 0), then the moments of X and F may be used to characterize the different aspects of the joint distribution. The first moment of X and F are the *expected value* (generally referred as the mean) of the observations, $E[X] = \mu_X$, and the expected value of the forecasts, $E[F] = \mu_F$. One can also characterize the *conditional moments*. For example, the first moment of F conditioned on the observation x is denoted $E[F|X = x] = \mu_{F|X}$. The first moment of X conditioned on the forecast f is denoted $E[X|F = f] = \mu_{X|F}$.

To characterize the deviation of a random variable from its expected value, one can use the *variance*. The variance of the observation is defined as $\sigma_X^2 = E[(X - \mu_X)^2]$. The variance of the forecast is defined as $\sigma_F^2 = E[(F - \mu_F)^2]$.

A measure of bias is the *mean error (ME)*, which is characterized by the first moments of the joint distribution as:

$$ME = \mu_F - \mu_X. \quad (4)$$

For unbiased forecasts, $ME = 0$. A positive mean error denotes over-forecasting whereas a negative mean error denotes under-forecasting.

A measure of accuracy is the *mean square error (MSE)* defined as the expected value of the square differences between the individual pairs of forecasts and observations:

$$MSE = E[(F - X)^2]. \quad (5)$$

Note that, being a quadratic measure, the *MSE* tends to penalize the large differences. For probability forecasts, the *MSE* is equivalent to the Brier Score (Brier 1950).

A measure of the association is the *correlation (ρ)*, defined by the joint distribution as:

$$\rho_{FX} = \frac{E[FX]}{\sigma_F \sigma_X}. \quad (6)$$

5.1 Calibration-Refinement Decomposition

The accuracy of the forecasts depends on other aspects of forecast quality. Viewed in one way, accuracy depends on the reliability and resolution of the forecasts. By conditioning on the forecast, the calibration-refinement (*CR*) decomposition of the *MSE* can be written as:

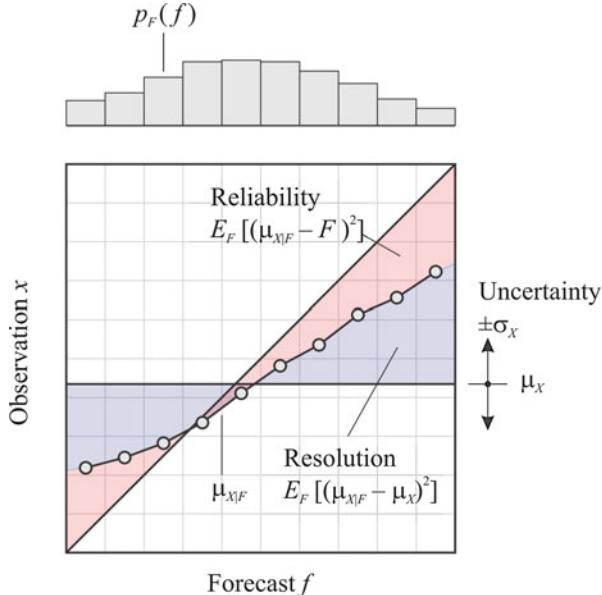


Fig. 2 Conceptual illustration of the calibration-refinement (CR) decomposition and its related aspects of forecast quality. The *box* represents all combinations of discrete forecasts and discrete observations. The *circles* (connected by the *line*) represent the expected value of the observation given the forecast $\mu_{X|F}$ for all possible forecasts. The *horizontal line* indicates the mean of the observations μ_X . The *sloping line* represents $x = f$. The histogram above the *box* represents the marginal distribution of the forecasts $p_F(f)$ (known as the predictive or refinement distribution). The forecast has perfect reliability if all the $\mu_{X|F}$ points fall on the *sloping line*. The reliability component *REL* measures the squared deviations from the *sloping line* weighted by the marginal distribution $p_F(f)$ (indicated by the *red shaded area*). The forecast has resolution since the $\mu_{X|F}$ points increase as the forecast f increases, indicating that the observations are different for different forecasts. The resolution component *RES* measures the squared deviations from the *horizontal line* μ_X weighted by the marginal distribution $p_F(f)$ (indicated by the *blue shaded area*). The uncertainty is the variability of the observations. The uncertainty component *UNC* is represented by the standard deviation of the observations σ_X about the mean

$$\begin{aligned} MSE &= E[(F - X)^2] = \sigma_X^2 + E_F(\mu_{X|F} - F)^2 - E_F(\mu_{X|F} - \mu_X)^2 \\ &= UNC + REL - RES \end{aligned} \quad (7)$$

where E_F is the expected value with respect to the forecast distribution and $\mu_{X|F}$ is the expected value of the observations conditioned on the forecast. Figure 2 illustrates these aspects of forecast quality for discrete forecasts and observations.

The *reliability component (REL)* or type-1 conditional bias measures the conditional bias of the forecasts:

$$REL = E_F[(\mu_{X|F} - F)^2] \quad (8)$$

For perfectly reliable forecasts, the expected value of the observations conditioned on the forecast equals f (i.e., $\mu_{X|F} = f$) and REL is 0.

The *resolution component* (RES) measures the degree to which the average observations given a specific forecast f differs from the unconditional mean (or climatology):

$$RES = E_F \left[(\mu_{X|F} - \mu_X)^2 \right] \quad (9)$$

For high-resolution probability forecasts, when the forecast probability is higher (lower) than the climatological event frequency, the event should be observed more (less) frequently than climatology. For high-resolution deterministic forecasts, when the forecast is higher (lower) than the mean observation, the average outcome should be more (less) than climatology. Forecasts with larger differences have higher resolution. If the conditional expected value of the observations $\mu_{X|F}$ is the same, regardless of the forecast f , the forecasts have no resolution. For a forecast with no resolution, RES is 0. The maximum resolution is obtained for a perfectly reliable forecast (i.e., $\mu_{X|F} = f$) where RES is equal to σ_X^2 .

The (inherent) *uncertainty component* (UNC) of the observations is measured with the variance of the observations:

$$UNC = \sigma_X^2 \quad (10)$$

Therefore, Eq. 7 shows the relationship between accuracy (as measured by the MSE) and the reliability and resolution of the forecasts. If the forecasts are perfectly reliable ($REL = 0$), then biases (conditional and unconditional) do not inflate the forecast errors; whenever there are biases ($REL > 0$), the MSE increases (lower accuracy). The forecasts must have good resolution to be accurate. As the resolution (RES) increases, the MSE decreases (higher accuracy). Note that a climatology forecast is perfectly reliable ($REL = 0$) but has no resolution ($RES = 0$), so the MSE is equal to the inherent uncertainty ($MSE = \sigma_X^2$).

5.2 Likelihood-Base Rate Decomposition

Viewed in another way, forecast accuracy depends on the sharpness, type-2 conditional bias, and discrimination of the forecasts. By conditioning on the observation, the likelihood-base rate (LBR) decomposition of the MSE can be written as:

$$\begin{aligned} MSE &= E[(F - X)^2] = \sigma_F^2 + E_X(\mu_{F|X} - X)^2 - E_X(\mu_{F|X} - \mu_F)^2 \\ &= SHA + T2B - DIS \end{aligned} \quad (11)$$

where E_X is the expected value with respect to the distribution of the observations and $\mu_{F|X}$ is the expected value of the forecasts conditioned on the observation.

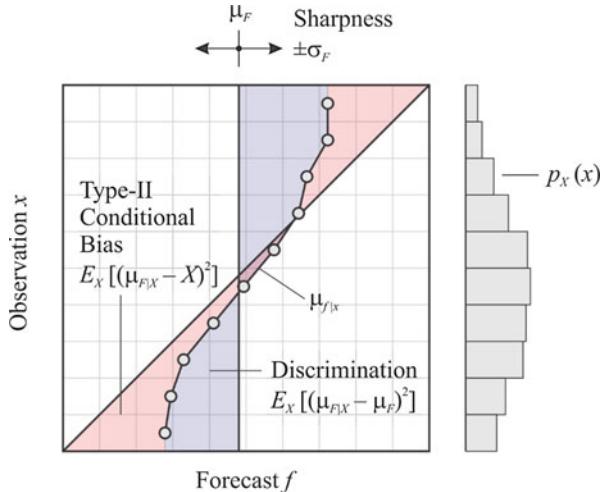


Fig. 3 Conceptual illustration of the likelihood-base rate (*LBR*) decomposition and its related aspects of forecast quality. The *box* represents all combinations of discrete forecasts and discrete observations. The *circles* (connected by the *line*) represent the expected value of the forecast given the observation $\mu_{F|x}$ for all possible outcomes. The *vertical line* indicates the mean of the forecasts μ_F . The *sloping line* represents $x = f$. The histogram to the side of the *box* represents the marginal distribution of the observations $p_X(x)$ (known as the base rate or uncertainty distribution). The forecast has no type-2 conditional bias if all the $\mu_{F|x}$ points fall on the *sloping line*. The type-2 conditional bias component *T2B* measures the squared deviations from the *sloping line* weighted by the marginal distribution $p_X(x)$ (indicated by the *red shaded area*). The forecast has discrimination since the $\mu_{F|x}$ points increase as the observation x increases, indicating that the forecasts are different for different outcomes. The discrimination component *DIS* measures the squared deviations from the *vertical line* μ_F weighted by the marginal distribution $p_X(x)$ (indicated by the *blue shaded area*). The sharpness is the variability of the forecasts. The sharpness component *SHA* is represented by the standard deviation of the forecasts σ_F about the mean

Figure 3 illustrates these aspects of forecast quality for discrete forecasts and observations.

The first term in the *LBR* decomposition is the *sharpness component (SHA)*, which measures the variability of the forecasts:

$$SHA = \sigma_F^2 \quad (12)$$

Forecasts with high sharpness are desirable because a large variance in forecast values has the potential to discern a broad range of outcomes. For probability forecasts, sharpness measures the degree to which forecast probabilities approach 0 and 1. For a climatology forecast (i.e., $f = \mu_X$ in all cases), the sharpness *SHA* is equal to 0.

The *type-2 conditional bias component (T2B)* (Murphy 1997) measures how close the observation agrees with the forecasts when a specific outcome is observed. It describes the bias conditioned on the observation:

$$T2B = E_X \left[(\mu_{F|X} - X)^2 \right] \quad (13)$$

Only a perfect forecast has no type-2 conditional bias. For a climatology forecast, $T2B$ is equal to σ_F^2 .

The *discrimination component (DIS)* measures how different the forecasts are for different outcomes:

$$DIS = E_X \left[(\mu_{F|X} - \mu_F)^2 \right] \quad (14)$$

For a flood probability forecast, the forecasts have discrimination if the average forecast probability $\mu_{F|X}$ when the event occurs ($x = 1$) differs from the average forecast probability when the event does not occur ($x = 0$). For a deterministic flow forecast, the forecasts have discrimination if the average forecast flow $\mu_{F|X}$ differs when a high or low flow is observed. If forecasts have no discrimination, then $\mu_{F|X}$ is the same (and equal to μ_F) regardless of the outcome and DIS is 0; a climatology forecast is one example of a forecast with no discrimination.

Therefore, Eq. 11 shows the relationship between accuracy (as measured by the *MSE*) and the sharpness, type-2 conditional bias, and discrimination of the forecasts. There is a seeming contradiction in Eq. 11, which implies that sharper forecasts would increase the *MSE* (lower accuracy). However, there is a trade-off between sharpness, type-2 conditional bias, and discrimination. At one extreme, if the same forecast is always issued, the forecasts have no sharpness ($SHP = 0$) and no discrimination ($DIS = 0$), and the type-2 conditional bias is maximized. Hence, forecasts must be sharp ($SHP > 0$) in order to have discrimination ($DIS > 0$) and low type-2 conditional bias.

5.3 Skill Score and Its Decompositions

The forecast skill is a measure of the accuracy of the forecast relative to a reference forecast. A skill score shows the improvement of the forecast accuracy relative to the reference (Murphy 1997). Using the *MSE* as a measure of accuracy, and climatology as a reference forecast, the *MSE skill score* is:

$$SS_{MSE} = 1 - \frac{MSE}{\sigma_X^2} \quad (15)$$

where σ_X^2 is the *MSE* of a climatology forecast. Perfect forecasts have a skill score of 1. Forecasts that are more accurate than a climatology forecast have a skill score greater than 0. Forecasts that are less accurate than a climatology forecast have a skill score less than 0.

The aspects of forecast quality that enhance or degrade forecast skill can be diagnosed using one or more decompositions of the skill score. Substituting the

decomposition in Eq. 7 for *MSE*, the calibration-refinement (*CR*) decomposition of the *MSE* skill score is:

$$SS_{CR} = \frac{E_F(\mu_{X|F} - \mu_X)^2}{\sigma_X^2} - \frac{E_F(\mu_{X|F} - F)^2}{\sigma_X^2} = \frac{RES}{\sigma_X^2} - \frac{REL}{\sigma_X^2} = R_{RES} - R_{REL} \quad (16)$$

The *CR* decomposition implies that forecasts are skillful (or more accurate than a climatology forecast) when their resolution is greater than their conditional bias (the reliability). In other words, forecasts with good resolution can be skillful, but poor reliability (conditional bias) will degrade their skill.

Similar to the *CR* decomposition, substituting the decomposition in Eq. 11 for *MSE* leads to the likelihood-base rate (*LBR*) decomposition of the *MSE* skill score:

$$\begin{aligned} SS_{LBR} &= 1 - \frac{\sigma_F^2}{\sigma_X^2} + \frac{E_X(\mu_{F|X} - \mu_F)^2}{\sigma_X^2} - \frac{E_X(\mu_{F|X} - X)^2}{\sigma_X^2} \\ &= 1 - \frac{SHA}{\sigma_X^2} + \frac{DIS}{\sigma_X^2} - \frac{T2B}{\sigma_X^2} \\ &= 1 - R_{SHA} + R_{DIS} - R_{T2B} \end{aligned} \quad (17)$$

The *LBR* decomposition relates the forecast skill to relative measures of the sharpness, discrimination, and type-2 conditional bias. As noted above, forecasts must be sharp in order to have high discrimination and low type-2 conditional bias (see Bradley et al. (2004) also).

Another useful *MSE* skill score decomposition was derived by Murphy (1988), which will be used extensively in the following sections. A basic decomposition of the *MSE* is:

$$MSE = E[(F - X)^2] = (\mu_F - \mu_X)^2 + \sigma_F^2 + \sigma_X^2 - 2\rho_{FX}\sigma_F\sigma_X \quad (18)$$

Substituting the decomposition in Eq. 18 for the *MSE* in Eq. 15, and rearranging the terms, yields:

$$SS = \rho_{FX}^2 - \left[\rho_{FX} - \left(\frac{\sigma_F}{\sigma_X} \right) \right]^2 - \left[\frac{\mu_F - \mu_X}{\sigma_X} \right]^2 = PS - SREL - SME. \quad (19)$$

Murphy (1988) interpreted this *MSE* skill score decomposition with the aid of a linear regression analogy for the conditional mean $\mu_{X|F}$ as a function of f . The first term is the potential skill (*PS*), a measure of association; it can be interpreted as the forecast skill for perfectly calibrated forecasts. The second term is the slope reliability (*SREL*), a linear regression-based measure of the reliability (or miscalibration). The third term is the standardized mean error (*SME*), a measure of the bias. One use of the decomposition is to diagnose the potential skill of the forecast and distinguish between the conditional bias (*SREL*) and unconditional bias (*SME*) that degrades the

forecast skill (Murphy and Winkler 1992; Hashino et al. 2007). In essence, this decomposition is similar to the CR decomposition shown in Eq. 16. With the linear regression analogy, the potential skill (PS) is equivalent to the relative resolution, where the slope reliability ($SREL$) and standardized mean error (SME) are the conditional and unconditional biases that represent the relative reliability. But because this decomposition is based on the moments of forecasts and observations (and not their conditional moments), it is often easier to compute from a verification data sample.

6 Forecast Quality of Hypothetical Ensemble Forecasts

To illustrate aspects of forecast quality for ensemble forecasts, we generated hypothetical ensemble forecasts for four forecasting systems. Each forecasting system makes an ensemble forecast for the same outcome – a continuous forecast variable. We will use a set of 100 forecast-observation pairs for each system for illustration purposes (a visual comparison); a much larger set will be used to derive moments (by Monte Carlo simulation) for a comparison with verification measures. Figure 4 shows a subset of the hypothetical forecasts for six forecast periods, along with the corresponding observed outcomes. For comparison, a climatological ensemble forecast (corresponding to the unconditional distribution of the observations) is also shown.

To create these forecasts, the hypothetical relationship between the forecasts and observations is represented by a statistical model (a bivariate normal distribution). The ensemble forecasting examples were constructed by randomly generating forecast-observations pairs from the model (the same observations were used in all four cases). By changing certain parameters of the model, we can represent forecasts of differing quality. For the high quality forecasts, the forecasts and observations both have zero mean and unit variance, with a correlation of 0.9. For the unconditional bias forecasts, the forecast mean was changed to 0.632. For the conditional bias forecasts, the forecast standard deviation was changed to 0.260. For the poor association forecasts, the correlation was changed to 0.7.

From Fig. 4 it is clear that all four hypothetical forecasts have some properties one would expect from good quality forecasts. All the ensembles have less spread than the climatology forecast. Also, all the ensembles tend to shift away from the climatology forecast and towards the observed outcomes. As we shall see, the first of the four forecasts shown for all six forecast periods (red box plots) is a high quality forecast. The second forecast (blue box plots) has the same ensemble spread as the high quality forecast but is systematically higher. In contrast, the third forecast (green box plots) has much less ensemble spread, but the observations rarely fall within the range of the forecasts. The fourth forecast (light blue box plots) resembles the high quality forecasts but has a larger ensemble spread.

In the following sections, we evaluate the quality of the four hypothetical ensemble forecasts. First we evaluate their forecast quality as single-valued forecasts, using the forecast ensemble mean as a single-valued forecast of the outcome.

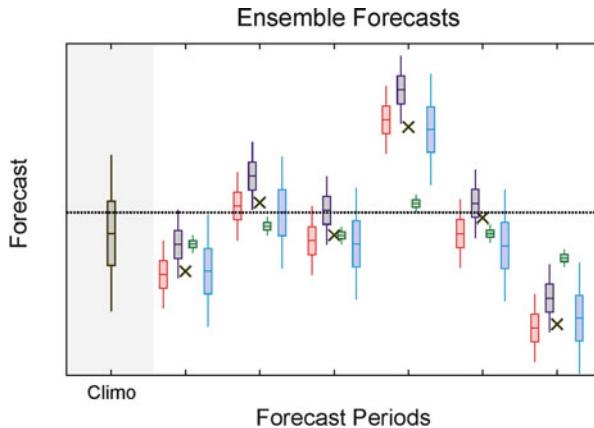


Fig. 4 Hypothetical ensemble forecasts for six forecast periods. Each forecast is represented by a box plot. The box contains 25–75% of the ensemble members; the whiskers extend to the 5% and 95% levels. The median is indicated by the bar within the box. The outcome for each forecast period (the observation) is indicated by the x-symbol. A climatology ensemble forecast, corresponding to the unconditional distribution of the observations, is shown on the left-hand side for comparison. The dashed horizontal line shows the upper tercile of the observation distribution; in a latter section, the ensemble forecasts will be used to make probability forecasts for outcomes above the threshold. For each forecast period, hypothetical ensemble forecasts are issued by the four forecasting systems, representing high quality forecasts (red), forecasts with unconditional biases (blue), forecasts with conditional biases (green), and forecasts with poor association (sky blue)

Next we evaluate their forecast quality as probability forecasts for a discrete event. Finally, we evaluate the forecast quality of the complete ensemble, treating it as a probability distribution forecast of the continuous outcome.

6.1 Illustration of Forecast Quality for a Single-Valued Forecast

First we will examine the use of ensemble forecasts as a single-valued forecast of a continuous forecast variable (e.g., streamflow). To accomplish this, the ensemble forecast must be transformed into a single value. As is often done in ensemble forecasting, the ensemble mean will be used to represent a single-valued forecast of the outcome. To characterize the forecast quality of single-valued forecasts, we examine the joint distribution of the forecast-observation pairs $p_{FX}(f, x)$. In this case, the forecast f is the ensemble mean, and the observation x is the outcome variable. We will first make a visual comparison of the joint distribution, as represented by scatter plots of forecast-observation pairs from a verification data sample; many aspects of forecast quality are readily interpreted from such information. Then we compare the forecasts using verification measures, which quantify different aspects of forecast quality.

6.1.1 Visual Comparison

Figure 5a illustrates single-valued forecasts with high forecast quality. The forecasts are unbiased (the mean of forecasts and observations falls on the one-to-one line) and have strong association (their correlation with the observations is high). Since the observations are closely scattered around the one-to-one line, the forecasts are reliable (the expected outcome is close to the forecast outcome) and have good resolution (when forecasts are higher than average, their observations tend to be higher than average, and vice versa). Furthermore, the forecasts have low type-2 conditional bias (the expected forecast is close to the observation) and high discrimination (when observations are higher than average, their forecasts tend to be higher than average, and vice versa). All in all, the forecasts for this case are very accurate (the forecasts and observations correspond quite well).

Figure 5b illustrates single-valued forecasts that are identical to the high quality case, except that their accuracy is degraded by unconditional bias. The forecasts are systematically higher than the observations, resulting in a shift from the one-to-one line to the right. As a result, the mean of forecasts and observations no longer falls on the one-to-one line. The forecasts still have strong association, good resolution, and high discrimination. However, they are no longer reliable (the expected outcome is lower than the forecast outcome) and have significant type-2 conditional bias (the expected forecast is higher than the observations).

Figure 5c illustrates single-valued forecasts that are identical to the high quality case, except that their accuracy is degraded by conditional bias; the forecasts are much less variable than in the high quality case, resulting in a steeper sloping relationship that no longer follows the one-to-one line. Overall, the forecasts are unbiased; the mean of the forecasts and observations falls on the one-to-one line. Also, they still have strong association and good resolution. However, as in the unconditional bias case, the forecasts are no longer reliable; the expected outcome is higher (lower) than the forecast at higher (lower) values. Furthermore, since the forecasts are less variable, their discrimination is poor (different observations are no longer associated with substantially different forecasts) and they have significant type-2 conditional bias, since the expected forecast is lower (higher) than the observations at higher (lower) values.

Figure 5d illustrates single-valued forecasts that are identical to the high quality case, except that their accuracy is degraded by poor association. The correlation between the forecasts and observations is much lower, resulting in more scatter about the one-to-one line. The forecasts remain unbiased. But because of the greater scatter in the relationship, the resolution and discrimination of the forecasts is weaker.

6.1.2 Comparison with Verification Measures

A variety of measures can be used to quantify aspect of forecast quality. Here we use moments of the joint distribution, including the mean squared error (*MSE*) as a measure of accuracy and its various decompositions, to compare the four sets of single-valued forecasts shown in Fig. 5.

Table 1 shows elements of the basic measures of forecast quality. The high forecast quality case is the most accurate; it has the lowest *MSE*. In contrast, the

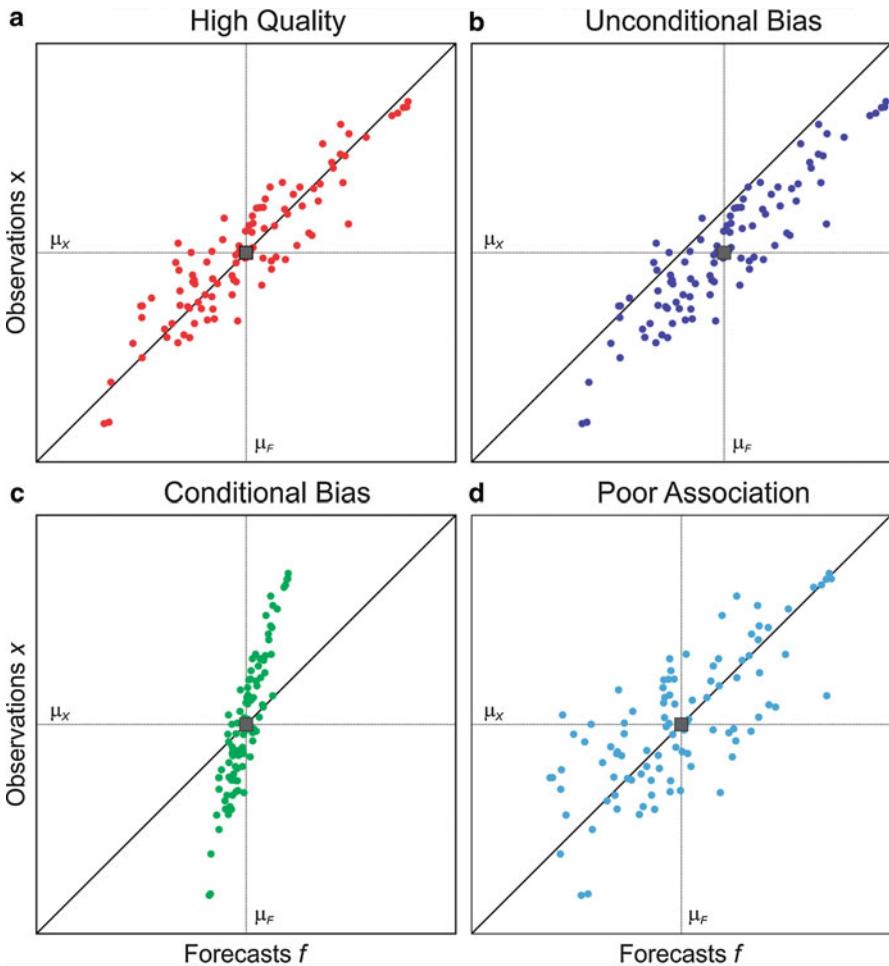


Fig. 5 Comparison of single-valued forecasts and observations for the four hypothetical cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. The panels show 100 forecasts-observation pairs; note that the observations are the same for all four cases. The *slope line* is the one-to-one line. The mean of the forecasts is indicated by the *vertical dashed line*; the mean of the observations is indicated by the *horizontal dash line*. The intersection of the means is indicated by the *box*. The intersection falls on the one-to-one for all but one case, indicating the forecasts are unbiased. The exception is the unconditional bias case (b), where the mean of the forecasts is much larger than the mean of the observations

other three cases are much less accurate; their *MSE* is three times higher. For the unconditional bias case, the reason for lower accuracy is the bias; the mean error (*ME*) is high for this case but equal to 0 for all the others (indicating unbiased forecasts). For the poor association case, the reason for lower accuracy is the association; the correlation is 0.7 for this case, but for all the others the correlation

Table 1 Measures of accuracy, bias, and association of single-valued forecasts for the four cases

Case	Accuracy MSE	Bias ME	Association ρ_{FX}
High quality	0.2	0	0.9
Unconditional bias	0.6	0.632	0.9
Conditional bias	0.6	0	0.9
Poor association	0.6	0	0.7

Table 2 Elements of the calibration-refinement MSE decomposition showing measures of accuracy, uncertainty, reliability, and resolution for the four single-valued forecast cases

Case	Accuracy MSE	Uncertainty σ_X^2	Reliability $E_F \left[(\mu_{X F} - F)^2 \right]$	Resolution $E_F \left[(\mu_{X F} - \mu_F)^2 \right]$
High quality	0.2	1	0.01	0.81
Unconditional bias	0.6	1	0.41	0.81
Conditional bias	0.6	1	0.41	0.81
Poor association	0.6	1	0.09	0.49

is much higher (0.9). For the conditional bias case, the bias and association are the same as in the high quality case; to understand the reason for lower accuracy, we must examine other aspects of the forecast quality.

Table 2 shows elements of the calibration-refinement MSE decomposition from Eq. 7. The high quality case is the most accurate because the forecasts are reliable and have good resolution. For the unconditional bias and conditional bias cases, the forecasts still have good resolution, but are no longer reliable. Note that both these cases have identical results for the calibration-refinement MSE decomposition, indicating that the two types of bias affect the reliability measure by the same magnitude; however, the ME of 0 for the conditional bias case (see Table 1) indicates that its poor reliability is due solely to conditional bias, while the large ME for the unconditional bias case indicates that its poor reliability is due to (unconditional) bias. For the poor association case, the forecasts are less accurate primarily because they have lower resolution.

Table 3 shows elements of the likelihood-base rate MSE decomposition from Eq. 11. The high quality case is the most accurate because the forecasts have low type-2 conditional bias and high discrimination. Although the unconditional bias case does not degrade the discrimination, it does contribute to a larger type-2 conditional bias. On the other hand, conditional bias case makes the forecast much less variable; its sharpness is 0.067, compared to 1 for the other three cases. Less variable forecasts mean they have very low discrimination and high type-2 conditional bias. For the poor association case, the forecasts are less accurate primarily because their discrimination is less.

Table 3 Elements of the likelihood-base rate MSE decomposition showing measures of accuracy, sharpness, type-2 conditional bias, and discrimination for the four single-valued forecast cases

Case	Accuracy MSE	Sharpness σ_F^2	Type-2 bias $E_X \left[(\mu_{F X} - X)^2 \right]$	Discrimination $E_X \left[(\mu_{F X} - \mu_X)^2 \right]$
High quality	0.2	1	0.010	0.81
Unconditional bias	0.6	1	0.410	0.81
Conditional bias	0.6	0.067	0.587	0.05
Poor association	0.6	1	0.090	0.49

Table 4 Elements of the MSE skill score decomposition showing measures of association (potential skill PS), reliability (slope reliability $SREL$), and unconditional bias (standardized mean error SME) for the four single-valued forecast cases. Climatology (μ_X) is the reference forecast

Case	Skill SS	Association PS	Reliability SREL	Bias SME
High quality	0.8	0.81	0.01	0
Unconditional bias	0.4	0.81	0.01	0.40
Conditional bias	0.4	0.81	0.41	0
Poor association	0.4	0.49	0.09	0

Table 4 shows elements of the MSE skill score decomposition from Eq. 19, using climatology (μ_X) as the reference forecast. The high quality case has high skill. In contrast, the skill for the other three cases is half that of the high quality case. The decomposition shows that bias (SME) degrades the forecasts for the unconditional bias case, poor reliability ($SREL$) degrades the forecasts for the conditional bias case, and lowered potential skill (PS) degrades the forecasts for the poor association case.

By the design of these hypothetical ensemble forecasts, the three lower quality cases have the exact same accuracy and skill, as indicated by the MSE and the skill score. However, the nature of the forecast themselves are quite different, as is seen in Fig. 5. Clearly, one single measure – accuracy or skill – is insufficient to characterize the nature of the forecasts. Indeed, one needs to fully consider the various aspects of forecast quality, as described by the joint distribution and factorizations (calibration refinement and likelihood-base rate), to distinguish between the nature of the different forecasts.

6.2 Illustration of Forecast Quality for a Probability Forecast for a Discrete Event

Although single-valued forecasts derived from the ensembles provide insights on the quality of the forecasts, they are an incomplete description of the ensembles. By transforming ensemble forecasts into single-valued forecasts, much of the

information contained in the ensemble is lost. In particular, the ensemble distribution provides information of the probability of different outcomes. In this section, we will use the hypothetical ensemble forecasts to examine probability forecasts for a discrete (event) outcome.

Consider a situation where one would like to forecast the probability of observing a relatively high outcome (e.g., high streamflow). To define a “high outcome,” we use a threshold. If the observed outcome z exceeds the threshold z^* , we say the event occurs. If the observed outcome does not exceed the threshold, the event does not occur. Therefore, the observation x is now a binary event – either it occurs or it does not. Mathematically, the binary observation x is simply a transformation of the continuous outcome variable z :

$$x = \begin{cases} 1 & \text{if } z > z^* \\ 0 & \text{if } z \leq z^* \end{cases}. \quad (20)$$

The forecast f for this event is the probability that the event occurs. In the case of ensemble forecasts, one can use the ensemble forecast probability distribution to define the probability forecast f of the event occurrence. This task is accomplished by estimating the relative frequency of the event occurrence using the ensemble members.

In this section, we will use the hypothetical ensemble forecasts to create probability forecasts for an outcome in the upper tercile of the climatological distribution of observed outcomes. (Note that this upper tercile threshold is illustrated in Fig. 4 by the horizontal dashed line.). By definition, the climatological exceedance of the threshold is one third ($\mu_X = 1/3$); the event occurs one third of the time and does not occur two thirds of the time. The hypothetical ensemble forecasts for all four cases are then transformed into probability forecasts f of the event occurrence. Using the probability forecasts f and the binary observations x , the joint distribution $p_{FX}(f, x)$ of the forecast-observations pairs can be examined.

6.2.1 Visual Comparison

Figure 6 illustrates the probability forecasts for the discrete events for the four examples. The plots show the continuous observations z on the y -axis. Whenever an observation exceeds the upper tercile (the horizontal line in the plots), the exceedance event occurs. The forecasts shown on the x -axis are the probability f that an exceedance event occurs. A climatology forecast corresponds to a forecast probability of $1/3$ (the vertical line in the plots). Whenever the forecast probability is greater than $1/3$, the event is predicted to be more likely to occur. If an exceedance event does occur, the forecast-observation pair plots in the upper right shaded area, indicating that the forecast is better than a climatology forecast. Likewise, whenever the forecast is less than $1/3$, and the exceedance event does not occur, the forecast-observation pair plots in the lower left shaded area, again indicating that the forecast is better than a climatology forecast. Individual forecasts that are worse than the climatology forecast plot in the two unshaded areas.

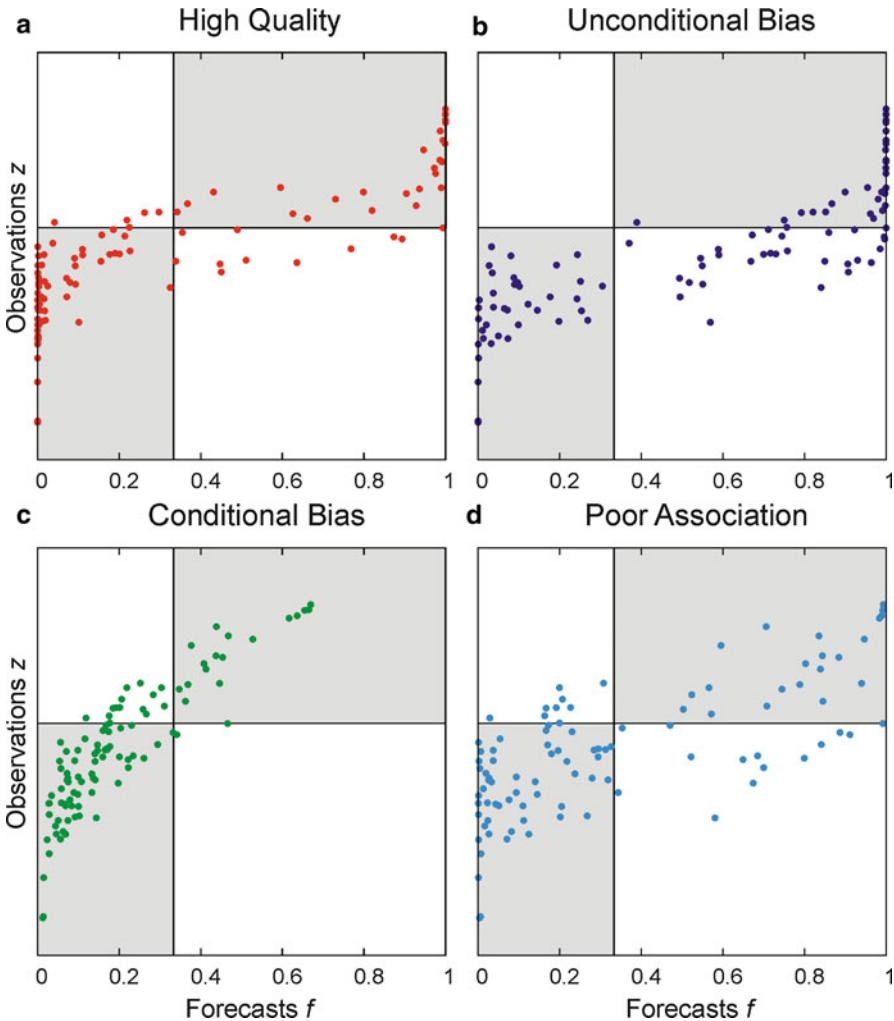


Fig. 6 Comparison of probability forecasts and continuous observations for an upper tercile event for the four hypothetical cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. The panels show 100 forecasts-observation pairs; note that the observations are the same for all four cases and are the same as for the single-valued forecasts (see Fig. 5). The horizontal line shows the upper tercile of the continuous observations. If the observation z is *above the line*, an exceedance event occurs; if z is *below the line*, the event does not occur. The forecast f is the forecast probability that the event occurs. The vertical line shows the climatological probability of an event occurrence (1/3). Forecast-observation pairs that fall in the shaded areas are better than a climatology forecast (less error); the event occurs (does not occur) and the forecast probability is higher (lower) than the climatological probability. Forecast-observation pairs that fall *outside the shaded areas* are worse than a climatology forecast (greater error); the event occurs (does not occur) and the forecast probability is lower (higher) than the climatological probability

Figure 6a illustrates probability forecasts with high forecast quality. Several aspects of forecast quality are apparent. First, the forecasts are sharp; often a probability forecast of 0 or 1 is issued. Furthermore, the forecasts are reliable (the expected outcome is close to the forecast outcome). When the probability forecast is near zero, the event rarely occurs; when the forecast is near one, the event usually occurs. The reliability is also seen for other forecast probabilities. For example, for a probability forecast near 1/3, the event occurs (the observation exceeds the threshold) about one third of the time as expected for reliable forecasts. The probability forecasts have good discrimination; when the event occurs, the probability forecast tends to be higher than the climatological probability, and vice versa. All these aspects contribute to the accuracy of the forecasts (the forecasts and event observations correspond quite well).

Figure 6b illustrates probability forecasts with unconditional bias. This is the case where the corresponding single-valued forecast (the ensemble mean) is systematically higher than the observations (see Fig. 5b). As with the single-valued forecasts, the probability forecasts are no longer reliable. For example, for a probability forecast near 0.2, one would still expect the event to occur about 20% of the time. However, no events occurred when the probability forecast issued was less than the climatological probability of 1/3. Likewise, even when the probability forecast is greater than the climatological probability, the event does not occur as often as its predicted probability. Overall, the accuracy of the forecasts is degraded by unconditional bias; the forecasts systematically overpredict the probability of the event occurrence.

Figure 6c illustrates probability forecasts with conditional bias. This is the case where the corresponding single-valued forecasts have much less variability than in the high quality case (see Fig. 5c). As with the single-valued forecasts, the probability forecasts are much less variable; the forecasts are much less sharp than the high quality case, with forecast probabilities ranging only from about 0 to 0.7. Even though the event occurs more often as the probability forecast increases, the forecasts are not reliable and have less discrimination than the high quality forecasts (different observations are not associated with as strongly different forecast probabilities). In this case, the forecasts systematically underpredict the probability of the event occurrence.

Figure 6d illustrates probability forecasts with poor association. This is the case where the corresponding single-valued forecasts have lower correlation than in the high quality case (see Fig. 5d). As with the single-valued forecasts, the probability forecasts are more scattered than in the high quality forecast case. The forecasts appear to be reliable but are less sharp (e.g., fewer forecast probabilities near zero and one are issued) and have less discrimination.

A more common way to make a visual comparison of probability forecasts is by plotting elements of the joint distribution. For example, a reliability diagram plots elements of the calibration-refinement factorization. Figure 7 shows the reliability diagrams for the four hypothetical probability forecasts for an upper tercile event. The main plot shows the observed relative frequency of the event occurrence $\mu_{X|F}$ as a function of the forecast f issued. For perfectly reliable forecasts, the observed

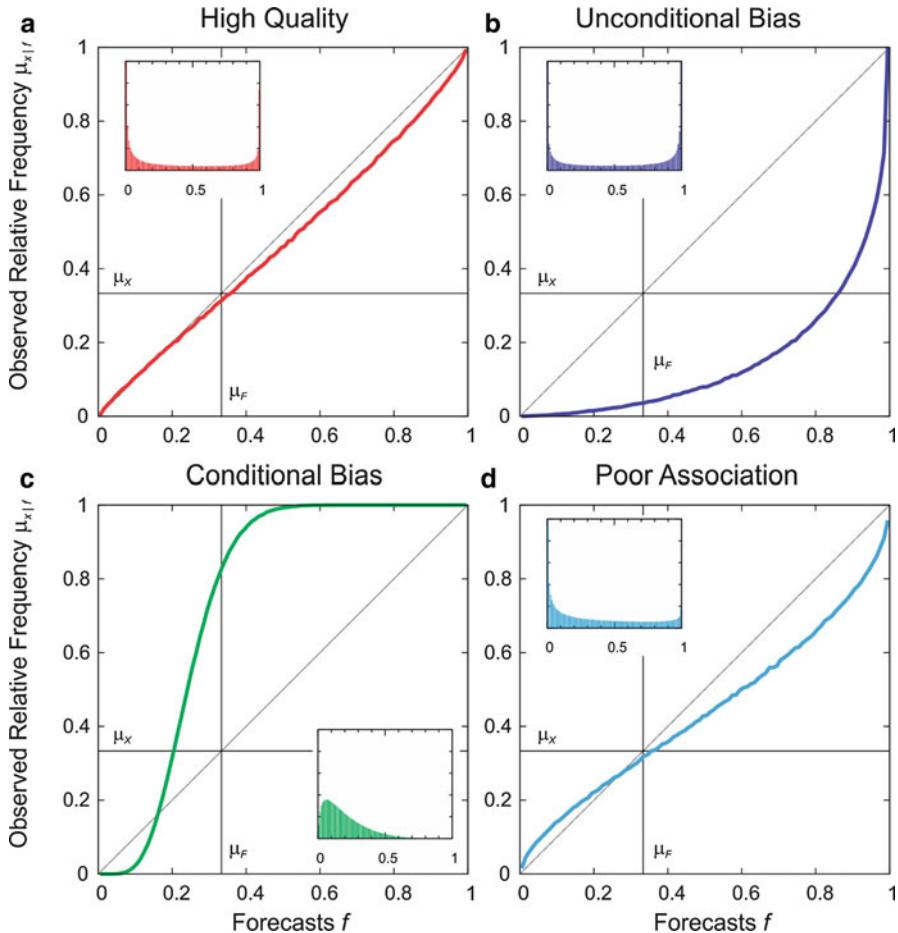


Fig. 7 Reliability diagram for probability forecasts for an upper tercile event for the four cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. The vertical line shows the mean forecast μ_F . The horizontal line shows the mean observation μ_X . A forecast is perfectly reliable if $\mu_{X|F}$ plots on the one-to-one line. A forecast has no resolution if $\mu_{X|F}$ plots on the horizontal μ_X line. The insert plot is the sharpness diagram showing the marginal distribution $p_F(f)$

relative frequency of the event $\mu_{X|F}$ is the same as its forecast probability (i.e., $\mu_{X|F}$ plots on the one-to-one line). As is seen in Fig. 7, all four forecasts are reliable when a forecast probability f of 0 or 1 is issued; the observed relative frequency is 0 or 1 in all four cases. However, only the high quality forecasts (see Fig. 7a) and the poor association forecasts (see Fig. 7d) are fairly reliable when other forecast probabilities are issued. For the unconditional bias case (see Fig. 7b), the forecasts systematically overpredict the probability of the event occurrence (the observed event occurrence frequency is less than forecast). In contrast, for the conditional bias case (see

Fig. 7c), the forecasts systematically overpredict event occurrence for forecast probabilities less than about 0.17 and underpredict the event occurrence for forecast probabilities greater than 0.17.

The insert plot on each reliability diagram shows the marginal distribution of the forecasts $p_F(f)$, or how likely each forecast probability f is issued. Note that most of the forecasts are sharp; forecast probabilities of 0 and 1 are issued with very high frequency. The one exception is the conditional bias case (see Fig. 7c). In this case, high forecast probabilities for the event occurrence are never issued. Furthermore, forecast probabilities of 0 are issued much less frequently than those slightly greater than 0; the most commonly issued forecast probability (the mode of the distribution) is 0.06. As noted above, the ensemble forecasts for the conditional bias case have a much narrower ensemble spread, resulting in probability forecasts that are much less variable (or less sharp) (see again Fig. 6c).

Figure 8 shows a discrimination diagram, which plots elements of the likelihood-base rate factorization. The plots show the conditional distribution of the forecast $p_{F|X}(f|x)$ for the two outcomes – when the event occurs ($x = 1$) because the observation is in the upper tercile, and when it does not occur ($x = 0$). For most cases, the forecasts are strikingly different for the two outcomes. When the event occurs a forecast probability of 1 is often issued; when the event does not occur a forecast probability of 0 is often issued. Hence, the forecasts have significant discrimination. The one exception is for the conditional bias case (see Fig. 8c); when the event occurs the conditional mean forecast is 0.12, and when the event does not occur the conditional mean forecast is 0.35. In this case, the discrimination is low.

6.2.2 Comparison with Verification Measures

Table 5 shows some basic measures of forecast quality. The probability forecasts of an upper tercile event are the most accurate for the high forecast quality case; it has the lowest MSE . In contrast, the other three cases have lower accuracy; their MSE is about twice as high. For the unconditional bias and the conditional bias cases, the reason for lower accuracy is the bias; the mean error (ME) is significant for these two cases (indicating biased forecasts). For the poor association case, the bias is low but the association is much less than for the high quality case.

Table 6 shows elements of the calibration-refinement MSE decomposition from Eq. 7. The high quality case is the most accurate because its forecasts are reliable and have good resolution. For the unconditional bias and conditional bias cases, the forecasts still have good resolution, but are no longer reliable. Note that both these cases have similar results for the calibration-refinement MSE decomposition, indicating that the two types of biases are affecting the reliability measure in a similar way. For the poor association case, the forecasts are less accurate primarily because they have lower resolution.

Table 7 shows elements of the likelihood-base rate MSE decomposition from Eq. 11. The high quality case is accurate because its forecasts have low type-2 conditional bias and high discrimination. Although unconditional bias does not significantly degrade the discrimination, it does contribute to a larger type-2 conditional bias. On the other hand, conditional bias makes the forecasts much less

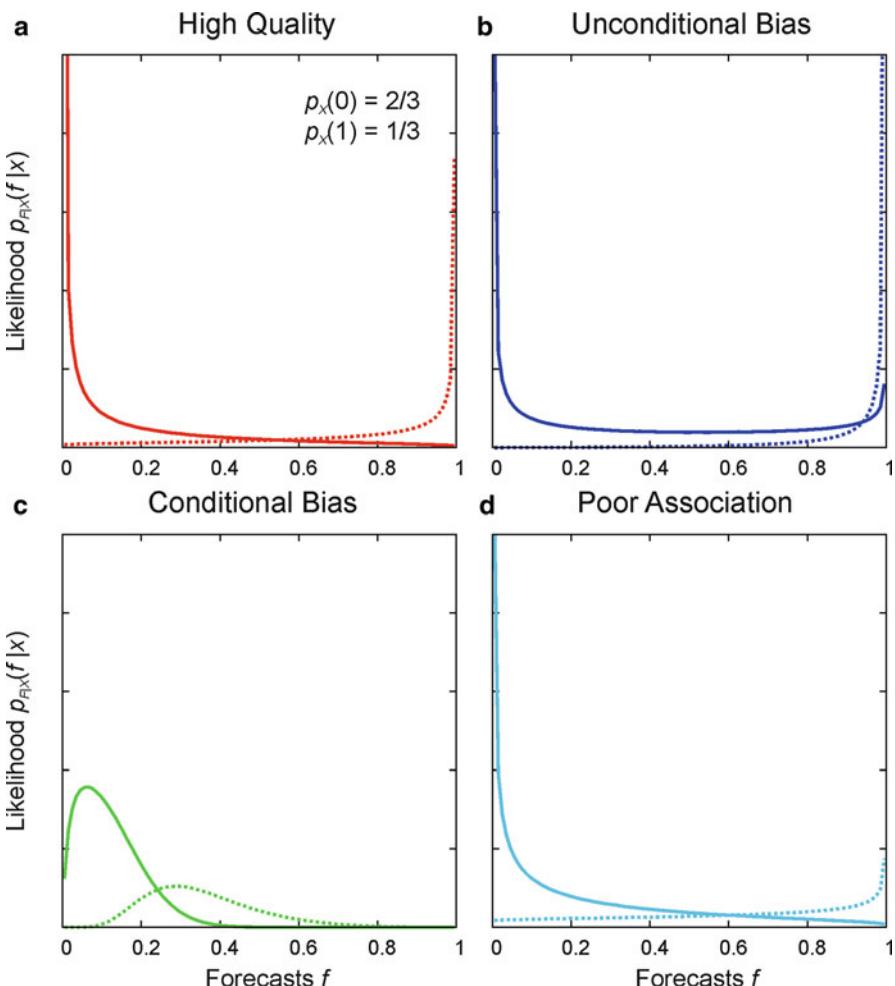


Fig. 8 Discrimination diagram for probability forecasts for an upper tercile event for the four cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. The conditional distributions $p_{F|X}(f|x)$ are shown for the two possible outcomes. The solid line shows the conditional forecast distribution when the event does not occur ($x = 0$). The dashed line shows the conditional forecast distribution when the event does occur ($x = 1$). A forecast has discrimination when the distributions are different for different outcomes. Note that the marginal distribution of the observation $p_X(x)$ is the same for all four cases

variable; its sharpness is 0.022, compared to values over 0.1 for the other three cases. Less variable forecasts mean they have very low discrimination and high type-2 conditional bias. For the poor association case, the forecasts are less accurate than the high quality case because their sharpness and discrimination are lower, and their type-2 conditional bias is higher.

Table 5 Measures of accuracy, bias, and association of probability forecasts for the four cases

Case	Accuracy MSE	Bias ME	Association ρ_{FX}
High quality	0.091	0.013	0.769
Unconditional bias	0.188	0.240	0.665
Conditional bias	0.160	-0.135	0.730
Poor association	0.157	0.030	0.559

Table 6 Elements of the calibration-refinement MSE decomposition showing measures of accuracy, uncertainty, reliability, and resolution for the four probability forecast cases

Case	Accuracy MSE	Uncertainty σ_X^2	Reliability $E_F \left[(\mu_{X F} - F)^2 \right]$	Resolution $E_F \left[(\mu_{X F} - \mu_F)^2 \right]$
High quality	0.091	0.222	0.001	0.132
Unconditional bias	0.188	0.222	0.096	0.130
Conditional bias	0.160	0.222	0.069	0.132
Poor association	0.157	0.222	0.005	0.070

Table 7 Elements of the likelihood-base rate MSE decomposition showing measures of accuracy, sharpness, type-2 conditional bias, and discrimination for the four probability forecast cases

Case	Accuracy MSE	Sharpness σ_F^2	Type-2 bias $E_X \left[(\mu_{F X} - X)^2 \right]$	Discrimination $E_X \left[(\mu_{F X} - \mu_X)^2 \right]$
High quality	0.091	0.141	0.034	0.084
Unconditional bias	0.188	0.154	0.102	0.068
Conditional bias	0.160	0.022	0.150	0.012
Poor association	0.157	0.105	0.085	0.033

Table 8 Elements of the MSE skill score decomposition showing measures of association (potential skill PS), reliability (slope reliability $SREL$), and unconditional bias (standardized mean error SME) for the four probability forecast cases. Climatology (μ_X) is the reference forecast

Case	Skill SS	Association PS	Reliability $SREL$	Bias SME
High quality	0.589	0.591	0.001	0.001
Unconditional bias	0.156	0.442	0.028	0.259
Conditional bias	0.280	0.533	0.171	0.082
Poor association	0.292	0.312	0.016	0.004

Table 8 shows elements of the MSE skill score decomposition from Eq. 19, using climatology (μ_X) as the reference forecast. The high quality case has the highest skill; the skill for the conditional bias and poor association cases are about half that of the high quality case, and for the unconditional bias case it is about one fourth. The decomposition shows that bias (SME) degrades the forecasts for the unconditional bias case, poor reliability ($SREL$) degrades the forecasts for the conditional bias case, and low potential skill (PS) degrades the forecasts for the poor association case.

6.3 Illustration of Forecast Quality for an Ensemble Probability Distribution Forecast

In the previous section we examined the quality of the ensemble forecast as a probability forecast for a discrete event. In this section, we evaluate the forecast quality of the complete ensemble, treating it as a probability distribution forecast of the continuous outcome.

6.3.1 Visual Comparison

Figure 9 illustrates the ensemble forecasts for the four hypothetical cases. The plots are similar to a scatter plot, but with the ensemble forecasts plotted as box plots (y-axis) versus their corresponding observations (on the x-axis).

The ensemble forecasts for the high forecast quality case (see Fig. 9a) appear to be reliable (the box plots are scattered about the one-to-one line), have good discrimination (a high observation tends to have a high ensemble forecast and vice versa), and have strong association. The ensemble forecasts for the unconditional bias case (see Fig. 9b) are identical to those for the high quality case, except that they are shifted upward from the one-to-one line. With forecasts systematically higher than the observations, the accuracy is degraded by unconditional bias. The ensemble forecasts for the conditional bias case (see Fig. 9c) are not reliable; although the ensemble spread is narrower, the slope of box plots no longer follows the one-to-one line. The ensemble forecasts for the poor association case (see Fig. 9d) appear to be reliable (follow the one-to-one line) but have a larger ensemble spread than the high quality case. As a result, the discrimination of the forecasts is also lower.

If one compares the plots of single-valued forecasts (Fig. 5) and probability forecasts for an event (Fig. 6), with those for the ensemble forecasts (Fig. 9), it is evident that ensemble forecasts are of a different nature. In the other two plots, the forecast itself is a number – the ensemble mean for the single-valued forecast, or a probability of an event for the probability forecast. In the case of the complete ensemble, the forecast is instead a probability distribution, as defined empirically by the ensemble members. Unfortunately, one cannot directly apply the distributions-oriented verification approach (as outlined in Sect. 3) to ensemble forecasts, since the forecast is in the form of a function and not a random variable. However, building from the example of the probability forecast of an event, one can construct and evaluate the quality of probability forecasts for any arbitrary binary event, as defined by an event threshold. Conceptually then, one can think of the forecast quality of the

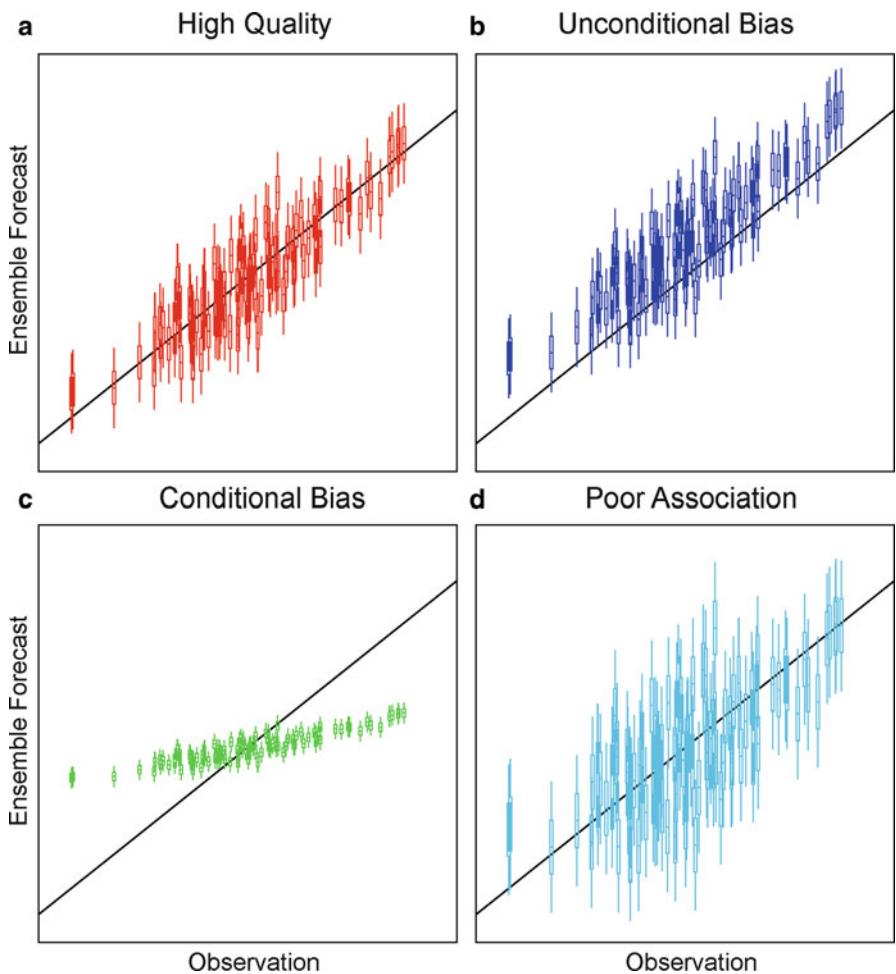


Fig. 9 Comparison of ensemble forecasts and observations for the four cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. Each ensemble forecast is represented by a *box plot*. The *box* contains 25–75% of the ensemble members; the whiskers extend to the 5% and 95% levels. The median is indicated by the *bar* within the *box*. The panels show 100 forecasts-observation pairs; note that the observations are the same for all four cases and are the same as for the single-valued and probability forecasts (see Figs. 5 and 6). The *solid line* shows the 1:1 relationship between the ensemble forecasts and the observations

ensemble forecast as a continuous function of the threshold (see Bradley et al. 2004; Bradley and Schwartz 2011).

Figure 10 shows an example of a forecast quality function for accuracy. The *MSE* of a probability forecast is plotted as a continuous function of the threshold. Rather than using the numerical value of the threshold itself, *MSE* is plotted against the climatological nonexceedance probability of the threshold value. A climatological

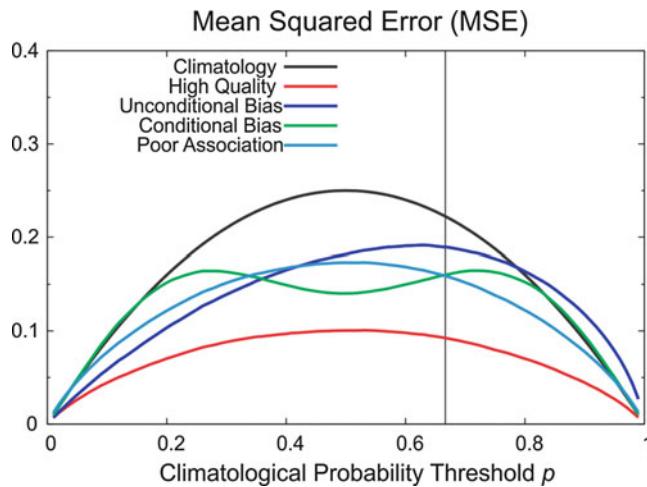


Fig. 10 MSE forecast quality functions for the four ensemble forecast cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association

probability close to zero defines a low outcome event (e.g., the nonexceedance probability of a low flow threshold), whereas a probability close to one defines a high outcome event (e.g., the nonexceedance probability of a high flow threshold). Note that an upper tercile event, as indicated by the vertical line, was previously evaluated in Sect. 6.2.

As can be seen in Fig. 10, the high quality ensemble forecasts are more accurate than the three others; their MSE is lower for all event thresholds. The unconditional bias forecasts, because they are systematically higher, yield more accurate probability forecasts for lower event thresholds than the other two cases but are the least accurate for thresholds greater than 0.42. The conditional bias forecasts have a narrow ensemble spread and are more accurate than the poor association forecasts near the middle tercile (between 0.34 and 0.66 thresholds). However, the poor association forecasts make better probability forecasts for events than the conditional bias forecasts in the lower and upper terciles.

For comparison, the MSE of climatology forecasts are shown for all levels. By definition, the climatology forecast probability is the same as the threshold probability p . Therefore, its MSE is simply the variance $p(1-p)$, which ranges from 0 at the two extremes to a maximum of 0.25 for the median ($p = 0.5$). Note that for some thresholds, the probability forecasts are not skillful, since they are less accurate than climatology forecasts. For example, the unconditional bias forecasts are not skillful for high event thresholds, whereas the conditional bias and poor association forecast are not skillful for low and high event extremes.

This comparison illustrates the drawback of plotting an absolute measure like MSE as a continuous function. MSE is a dimensional measure, with units of probability squared (for probability forecasts). It is valid to compare MSE as a

measure of accuracy for a fixed threshold; however, it is not valid to compare MSE at different thresholds. In other words, the fact that the MSE for low and high thresholds is quite small does not mean the forecasts for these thresholds are more accurate; indeed, the MSE for climatology forecasts are also quite small for these thresholds. A better way to plot forecast quality as a continuous function is using a relative measure. Figure 11 illustrates this using the MSE skill score and its decomposition. Recall, the MSE skill score is the accuracy of a forecast relative to a reference forecast. In Fig. 11, a climatology forecast is used as the reference. The skill score and its decomposition are dimensionless measures and can be compared across threshold values.

It is clear that the high quality ensemble forecasts (see Fig. 11a) make very good probability forecasts. They have high skill (in terms of MSE), high association (potential skill PS), and virtually no conditional bias (slope reliability $SREL$) or unconditional bias (standardized mean error SME), except perhaps at extreme low and high event thresholds.

For the unconditional bias ensemble forecasts (see Fig. 11b), the probability forecasts for nearly all thresholds have high association (or high potential skill). The biases are low at low event thresholds, making them quite skillful forecasts. However, the unconditional bias (SME) increases steadily with threshold, which degrades the skill. For thresholds in the upper tercile, the conditional bias ($SREL$) becomes quite large as well. As a result, the ensemble forecasts for the unconditional bias case are not skillful for high event thresholds (0.79 and greater).

For the conditional bias ensemble forecasts (see Fig. 11c), the probability forecasts also have high association (high potential skill). However, their narrow ensemble spread results in significant conditional bias ($SREL$) at all thresholds, and unconditional biases (SME) that peak at 0.25 and 0.75. As a result, the ensemble forecasts for the conditional bias case are not skillful for low and high event thresholds (below 0.16 or above 0.84).

For the poor association ensemble forecasts (see Fig. 11d), the symmetrical shape of the forecast quality functions resembles those of the high quality forecast. However, the forecasts have lower overall skill. They have lower association and thus lower potential skill (PS). Furthermore, the conditional bias ($SREL$) is much larger at very low and high event thresholds, as are the unconditional bias (SME). As a result of low association and significant biases, the ensemble forecasts for the poor association case are not skillful for low and high event thresholds (below 0.04 or above 0.96).

Other MSE skill score decompositions may be used to plot a relative forecast quality measure as a continuous function. Figure 12 shows the likelihood-base rate (LBR) decomposition of the MSE skill score, with terms based on Eq. 18, using climatology as the reference forecast. This decomposition uses measures that describe forecast quality conditioned on the observations. The high quality forecasts (see Fig. 12a) are sharp, have good discrimination, and have low type-2 conditional bias. The poor association forecasts (see Fig. 12d) resemble the high quality forecasts, but with generally lower sharpness, lower discrimination, and higher type-2 conditional bias, resulting in lower overall skill. The conditional bias forecasts (see

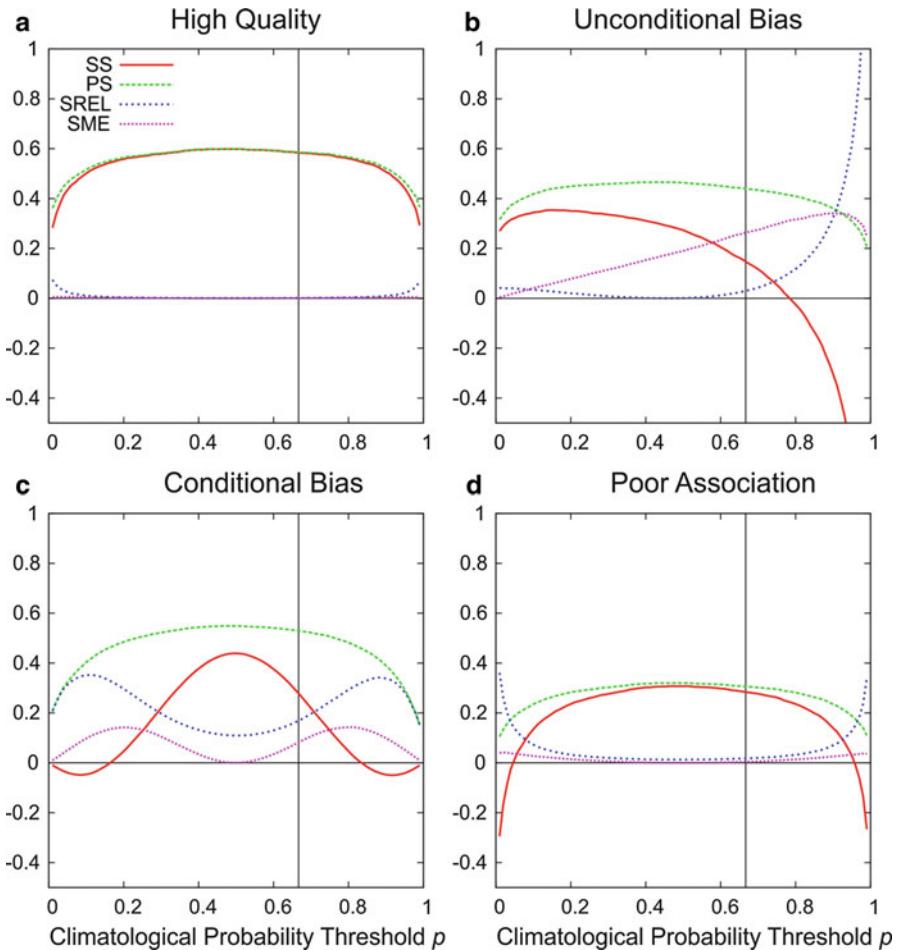


Fig. 11 *MSE* skill score and its decomposition forecast quality functions for the four ensemble forecast cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. Climatology is the reference forecast. Each plot show measures of skill (skill score *SS*), association (potential skill *PS*), reliability (slope reliability *SREL*), and unconditional bias (standardized mean error *SME*). The measures are based on probability forecasts for binary events, defined by the climatological nonexceedance probability threshold p . The vertical line defines the upper tercile event ($p = 0.667$); forecast quality for probability forecasts of this upper tercile event was examined in detail (see Figs. 6, 7, and 8)

Fig. 12c), with their narrow ensemble spread, have very low sharpness (near zero for low and high event thresholds). Since forecasts cannot have discrimination without sharpness, the discrimination is also quite low. Unlike the other cases, the skill function for the unconditional bias forecasts (see Fig. 12b) is not symmetrical; the forecasts are skillful for low and medium event thresholds, but not for high

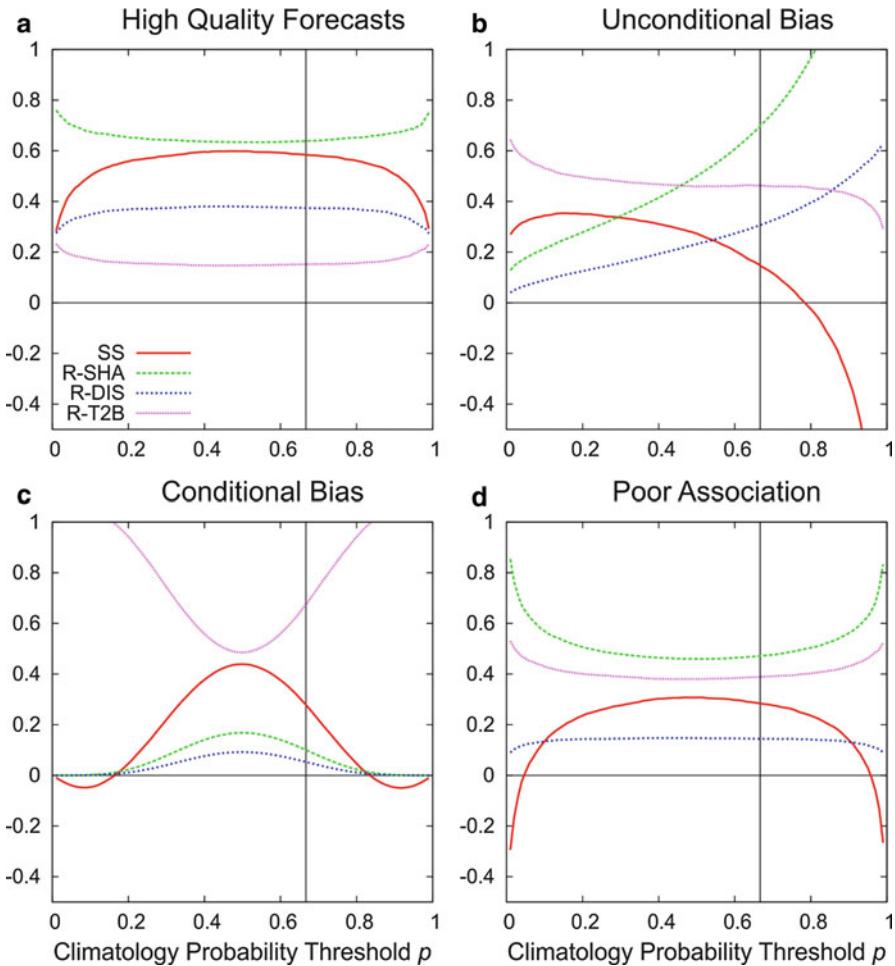


Fig. 12 *MSE* skill score and its likelihood-base rate (*LBR*) decomposition forecast quality functions for the four ensemble forecast cases: (a) high quality forecasts, (b) forecasts degraded by unconditional bias, (c) forecasts degraded by conditional bases, and (d) forecasts degraded by poor association. Climatology is the reference forecast. Each plot shows measures of skill (skill score SS), relative sharpness (R_{SHA}), relative discrimination (R_{DIS}), and relative type-2 conditional bias (R_{T2B}), as defined in Eq. 17. The measures are based on probability forecasts for binary events, defined by the climatological nonexceedance probability threshold p . The vertical line defines the upper tercile event ($p = 0.667$); forecast quality for probability forecasts of this upper tercile event was examined in detail (see Figs. 6, 7, and 8)

thresholds. Interestingly, the sharpness and discrimination are actually lower at low event thresholds and increase dramatically moving to higher thresholds. However, the discrimination increases more slowly than the sharpness, resulting in lower skill at higher thresholds.

Table 9 The continuous ranked probability score (\overline{CRPS}) for the four sets of ensemble forecasts and climatology forecasts

Case	Accuracy \overline{CRPS}
High quality	0.252
Unconditional bias	0.489
Conditional bias	0.463
Poor association	0.438
Climatology	0.564

6.3.2 Comparison with Verification Measures

Unlike the previous cases of single-valued forecasts and probability forecasts for a discrete event, verification measures derived from the joint distribution of forecasts and observations are not applicable to ensemble forecasts in the form of a probability distribution. Instead, we plotted verification measures for probability forecasts – defined by the ensemble forecasts for different threshold levels – as a continuous function of the thresholds, to illustrate the forecast quality of ensemble forecasts. However, measures that summarize aspects of these forecast quality functions are often used in ensemble forecast verification. The most common example is the continuous ranked probability score (Matheson and Winkler 1976; Unger 1985).

The continuous ranked probability score (\overline{CRPS}) is a measure of the overall accuracy of the ensemble forecast. It is related to the ranked probability score and the Brier score. The ranked probability score (RPS) extends the concept of the *MSE* for a binary event (also known as the Brier score) to multicategory events. The \overline{CRPS} extends the RPS to the limit of an infinite number of categories (Hersbach 2000). Mathematically, the \overline{CRPS} is related to the *MSE* as:

$$\overline{CRPS} = \int_{-\infty}^{\infty} MSE(z) dz \quad (21)$$

where $MSE(z)$ is the *MSE* for probability forecasts for an event defined by a threshold outcome z . Conceptually, the \overline{CRPS} summarizes the *MSE* function illustrated in Fig. 10. Table 9 shows the \overline{CRPS} for the four sets of ensemble forecasts. Results for climatology ensemble forecasts are also shown for comparison.

Like the *MSE*, the \overline{CRPS} is a measure of accuracy for the ensemble forecasts. As expected, the high quality ensemble forecasts are the most accurate (lowest \overline{CRPS}), while the climatology ensemble forecasts are the least accurate (highest \overline{CRPS}). The remaining three ensemble forecasts have more similar accuracy. It is clear that all four sets of ensemble forecasts are skillful compared to the climatology reference.

Similar to the \overline{CRPS} , which summarizes the *MSE* function (Fig. 10) over the entire range of thresholds, relative measures like the skill scores and its decomposition can be evaluated. For example, Bradley and Schwartz (2011) showed that the continuous ranked probability skill score ($CRPSS$) is mathematically equivalent to:

$$CRPSS = \int_{-\infty}^{\infty} w(z) SS(z) dz \quad (22)$$

where $SS(z)$ is the Brier skill score for probability forecasts for an event defined by a threshold outcome z , and $w(z)$ is a weighting function that depends on the inherent uncertainty for threshold z (see Eq. 10). Therefore, $CRPSS$ is the weighted-average skill of the function $SS(z)$ over all possible outcomes z . If instead one indexed the skill score function by the climatological probability p (as was done in Figs. 10 and 11), an analogous summary measure of the skill of the ensemble forecast is:

$$\overline{SS} = \int_0^1 w(p) SS(p) dp \quad (23)$$

where $w(p)$ is a weighting function proportional to the inherent uncertainty $p(1-p)$ for threshold p :

$$w(p) = \frac{p(1-p)}{\int_0^1 p(1-p) dp} = 6p(1-p) \quad (24)$$

Here, \overline{SS} represents the weighted-average skill of the skill function shown in Fig. 11. Using a similar approach for all elements of the decomposition yields the summary measures shown in Table 10.

As seen in Fig. 11, the high quality ensemble forecasts have high potential skill (\overline{PS}) and low conditional (\overline{SREL}) and unconditional biases (\overline{SME}). Both the unconditional bias and conditional bias cases have reasonably high potential skill. However, a large \overline{SME} degrades the skill for the unconditional bias ensemble forecasts, and a large \overline{SREL} degrades the skill for the conditional bias forecasts. For the poor association ensemble forecasts, the conditional (\overline{SREL}) and unconditional biases (\overline{SME}) are slightly higher than for the high quality forecasts; however, their skill is much lower because their potential skill is much lower than all the other cases.

Table 10 Summary verification measures for ensemble forecast skill and its decomposition. The summary measures indicate the weighted-average MSE skill \overline{SS} , potential skill \overline{PS} (association), slope reliability \overline{SREL} (conditional bias), and standardized mean error \overline{SME} (unconditional bias)

Case	Skill \overline{SS}	Association \overline{PS}	Reliability \overline{SREL}	Bias \overline{SME}
High quality	0.572	0.576	0.003	0.001
Unconditional bias	0.200	0.441	0.048	0.193
Conditional bias	0.234	0.507	0.199	0.074
Poor association	0.260	0.296	0.028	0.008

Table 11 Summary verification measures for ensemble forecast *MSE* skill and its likelihood-base rate (*LBR*) decomposition. The summary measures indicate the weighted-average skill \overline{SS} , the relative sharpness $\overline{R_{SHA}}$, relative discrimination $\overline{R_{DIS}}$, and relative type-2 conditional bias $\overline{R_{T2B}}$

Case	Skill \overline{SS}	Sharpness $\overline{R_{SHA}}$	Discrimination $\overline{R_{DIS}}$	Type-2 bias $\overline{R_{T2B}}$
High quality	0.572	0.645	0.371	0.154
Unconditional bias	0.200	0.578	0.247	0.469
Conditional bias	0.234	0.090	0.048	0.725
Poor association	0.260	0.489	0.143	0.394

Table 11 shows summary measures based on the *MSE* skill likelihood-base rate decomposition from Eq. 17, using climatology as the reference forecast. As seen in Fig. 12, the high quality forecasts have high sharpness, low type-2 conditional bias, and high discrimination. The unconditional bias forecasts are also sharp but have lower discrimination and larger type-2 conditional bias. The conditional bias forecasts are not sharp; poor discrimination and high type-2 conditional bias degrade their skill. The poor association forecasts still have good sharpness but low discrimination and significant type-2 conditional bias.

7 Practical Considerations

7.1 Diagnostic Verification and Ensemble Forecasts

For those new to forecast verification, it is common to ask what one metric is best to use to assess forecasts or forecast systems. One important lesson of the hypothetical ensemble forecast examples shown in this chapter is that a single measure of forecast quality is insufficient (Murphy 1993). By design, the hypothetical examples had one high quality forecast case and three lower quality forecasts cases; the lower quality forecasts were created by degrading the high quality forecasts in a specific manner. Although the three lower quality ensemble forecasts were dramatically different, their forecast accuracies (or skills) were roughly the same (whether treated as single-valued forecasts, as probability forecasts, or as probability distribution forecasts). It should be obvious then that forecasts with similar accuracy are not necessarily of comparable quality. Examining elements of the joint distribution of forecasts and observations – both visually for a qualitative assessment, and numerically with forecast verification measures for a quantitative assessment – was needed to distinguish the nature of the different forecasts. The distributions-oriented verification framework provides guidance on what aspects and measures of forecast quality are relevant to such a diagnostic assessment.

Different aspects of forecast quality have significant diagnostic and practical implications in ensemble forecasting. For example, when a forecast has poor association, then its potential skill and accuracy are limited. In contrast, when forecast accuracy is degraded by bias, accuracy can be improved by reducing or eliminating diagnosed biases through ensemble postprocessing. Unconditional bias in ensemble

forecasts can easily be removed with simple bias-correction methods (Seo et al. 2006; Hashino et al. 2007). Conditional bias can be improved by calibration methods (Atger 2003; Gneiting et al. 2005; Brown and Seo 2010; among others), but the task is much more complicated. Therefore, it is important to diagnose the reasons for low accuracy and to distinguish between unconditional and conditional biases.

7.2 Absolute Versus Relative Measures

Common absolute measures of forecast quality – like MSE and its decompositions – are often used in forecast verification. Such measures take on the dimensions of the forecast variable. For instance, for a single-valued forecast of flow, MSE has dimensions of flow-squared. For a probability forecast of an event, MSE has dimensions of probability squared. The dimensionality of these measures can limit their applications in hydrologic ensemble forecast comparisons.

For example, consider a situation where one wants to compare the quality of forecasts issued at different locations in a watershed to find where they are most accurate. Using the MSE as a measure of accuracy, one would naturally find that smaller rivers (with lower flows) tend to have lower MSE than those for bigger rivers (with higher flows). To make the comparison possible, one needs a nondimensional measure of accuracy. A common approach is to take the root mean square error (RMSE) and divide by the observation mean, so that differences in the relative accuracy can be assessed.

Likewise, consider a situation where one wants to determine the quality of probability forecasts at a single location for an ensemble prediction system. This is the same situation encountered in the verification example for the ensemble probability distribution forecasts. Here, we first plotted MSE (see Fig. 10) as a function of the defining event threshold (its climatological probability). But forecast probabilities for commonly occurring events (e.g., flow above an upper tercile) tend to be much higher than for rarely occurring events (e.g., an extreme high flow that occurs only 5% of the time). As a result, MSE tends to be higher for commonly occurring events than for rarely occurring events. A smaller MSE is not synonymous with a more accurate forecast; instead, one needs to compare results with the MSE of a climatology forecast (a reference forecast) to make any inferences on their accuracy.

For any comparisons where the different forecast variables have very different climatologies, the use of absolute measures for comparisons is problematic. Instead, relative measures of forecast quality – like MSE skill score and its decompositions shown in Eqs. 15, 16, and 18 – are better suited for forecast comparisons. As the verification example showed, plotting relative measures of forecast quality (see Fig. 11) as a function of threshold allowed an assessment of when probability forecasts for an event were of higher quality, and when they were of lower quality. The results showed that even with the same ensemble forecast, the probability statements issued by ensemble probability distribution function forecasts are not of the same quality; probability statements for certain thresholds were of higher quality than for others. Clearly, the use of relative forecast measures requires an appropriate

reference forecast. By evaluating forecast quality relative to a reference, relative forecast measures allow an “apples-to-apples” comparison of results required for many forecast verification applications.

7.3 Choosing a Verification Data Sample

Although the distribution-oriented forecast verification measures presented in Sect. 5 are defined by the moments of the joint distribution of forecasts and observations, in practice the moments must be *estimated* with a verification data sample. That is, a sample made up of forecasts and their corresponding observation is chosen, and then sample estimates of moments are computed. Because of the assumptions made, the verification data sample must be chosen with some care.

For example, one assumption of the joint-distribution model is that forecast-observation pairs are identically distributed. Hence, one must choose a verification data sample that is a random sample drawn from this joint distribution. Consider the case of forecasts of the flow volume for the upcoming season for a basin in the Rocky Mountains of the western United States. Because of climatological variations in river flows, the distribution of observed seasonal flow volumes for April forecasts (which are a result of snow accumulation over the winter months and their melting rate during the spring season) can be very different than for August forecasts (which are dominated by short-term meteorological events). Yet in practice, in an effort to create a larger sample size, verification data samples are often formed by pooling forecasts issued at different times of the year. In this case, pooling of April and August forecasts would violate the fundamental assumption of stationarity; the joint distribution for April forecasts is not the same as that for August forecasts. Hamill and Juras (2006) show examples where violating the stationarity assumption when choosing a verification data sample can lead to false inferences (unreliable or misleading verification measures). Therefore, selection of a verification data sample must be done with care. Pooling of forecasts issued at different times, forecasts relative to various lead times, or forecasts issued for different locations, should be avoided, unless additional analysis shows that stationarity assumption appears reasonable.

The above example illustrates one case where stratifying the entire forecast-observation data set should be employed. However, data stratification may also be used to better assess how the forecast quality depends on the forecast situation or condition. For instance, it may be useful to assess forecast quality by region (stratify data spatially) and/or by season (stratify data temporally). In other cases one may want to assess forecast quality for particular atmospheric or hydrologic outcomes, such as the occurrence of precipitation or flooding (stratify data by condition). Note that data stratification should involve categories with reasonable sample sizes to obtain reliable verification statistics for each category. This can be a significant obstacle with hydrometeorological ensemble forecast verification, where data samples tend to be small. Consider again the case of seasonal flow volume forecasts. If hindcasting is used to create forecasts for the past 30 years, that means the data sample for April (or August) forecasts is 30 for a particular location. This sample size

is already quite small. However, if an adequate sample size can be maintained, such as when forecasts are issued more frequently, the hindcast period is long, or it is reasonable to pool forecasts issued at different times, then additional data stratification could provide useful insights.

The second assumption of the joint-distribution model is that forecast-observation pairs are independent. The independence assumption can be violated if forecasts or observations are significantly correlated in time or space. For instance, a short-range forecast that is updated frequently would have forecasts and observations that are correlated. Such violations would not necessarily lead to unreliable or misleading verification measures (as long as stationarity assumption is still valid). However, violation of the independence assumption reduces the effective size of the verification data sample. As a result, the sampling uncertainties associated with the estimated verification measures would be higher than for an independent sample.

It is important also to recognize the limitations of observational data sets used in forecast verification. Some observations are not direct measurements of the forecast variable. Consider a quantitative precipitation forecast for a specific duration and areal domain; its corresponding observation might then be estimated by areal interpolation from point precipitation gauge measurements and/or weather radar. Note also that even direct measurements of a forecast variable (e.g., observations of streamflow at a stream-gage) have measurement errors. Changes in data collection and quality control over time, data processing, and data interpolation methods add uncertainty regarding the structure of observational errors and their effect on the verification process. However these observational errors are generally neglected during the verification process, assuming that they are insignificant compared to the forecast error. Methods to account for the observational error are under investigation (e.g., Gorgas and Dorninger 2012). Note that sometimes another reference observation is used to verify forecasts, such as simulated flow values (produced by the same forecasting system from observed atmospheric inputs), to help isolate and analyze a specific source of error (e.g., Jaun and Ahrens 2009; Demargne et al. 2010; Brown et al. 2014).

7.4 Sampling Uncertainties for Small Data Samples

As with any estimate based on a data sample, forecast verification measures estimated with a verification data sample have sampling uncertainties. Obviously, estimates should be based on the largest verification data sample possible. However, as noted above, pooling of forecast-observation pairs to create a larger sample is counterproductive if pooling violates assumptions of the distributions-oriented approach. Therefore, one of key requirements in forecast verification is to archive forecasts from a fixed forecast system for multiple years, or to include a hindcasting capability to retrospectively produce forecasts from a fixed version of the forecast system.

Ideally, evaluating the sampling uncertainties of verification measures should be a routine part of forecast verification to account for the uncertainty associated with a

verification metric, as well as help analyze if one forecast system is significantly different from another for a given metric and given the sampling uncertainty. Unfortunately, assessing sampling uncertainties is not yet common place. Yet for certain verification measures, simple analytical expressions for sampling uncertainty have been derived (Kane and Brown 2000; Stephenson 2000; Thorne and Stephenson 2001; Mason and Graham 2002; Ferro 2007; Jolliffe 2007). For example, Bradley et al. (2008) derived expressions for the sampling uncertainty of the *MSE* and skill scores for probability forecasts; Wilks (2010) has extended this approach to evaluate sampling uncertainty for correlated forecast-observation pairs. In the absence of analytical uncertainty expressions, others have used a more computationally intensive resampling method to assess sampling uncertainties (Efron 1981; Wilks 2011). This method involves randomly resampling forecast-observations pairs to compute verification measures and define their empirical distribution. Examples of the resampling method in forecast verification literature include Mason and Mimmack (1992), Politis and Romano (1994), Wilks (1996), Hamill (1999), Zhang and Casey (2000), Doblas-Reyes et al. (2003), Accadia et al. (2003, 2005), Ebert et al. (2004), Jolliffe (2007), and Ferro (2007), among others.

8 Conclusions

Forecast verification is the process of assessing the quality of forecasts. Using a distributions-oriented approach to forecast verification, aspects of forecast quality are defined based on the joint distribution of forecasts and observations. Hydrometeorological ensemble forecasts can be used to represent many types of forecasts, such as single-valued deterministic forecasts of a continuous outcome, probability forecasts for event occurrences (e.g., a flood event), or ensemble probability distribution forecasts of a continuous outcome. Regardless of the form of the forecast (deterministic or probabilistic) or the observations (discrete or continuous), the joint distribution of forecasts and observations completely describes their relationship.

The many aspects of forecast quality can be examined using the joint distribution or one of its factorizations into a conditional and marginal distribution. Basic aspects like bias, accuracy, skill, and association are defined by the joint distribution. By examining the distribution of observations for specific forecasts, aspects like reliability (type-1 conditional bias) and resolution are defined. By examining the distribution of forecasts for specific observations, aspects like type-2 conditional bias and discrimination are defined. Other aspects such as sharpness depend only on the distribution of forecasts, while uncertainty depends only on the distribution of observations. Common measures of these aspects of forecast quality are defined based on the moments of the joint distribution or its conditional and marginal distributions.

Hypothetical ensemble forecasts were used to illustrate aspects of forecast quality as single-valued forecasts (ensemble mean), probability forecasts for an event, and as ensemble probability distribution forecasts. One of the ensemble forecasts was of high quality, while three others were lower quality – degraded by either unconditional bias,

conditional bias, or poor association. Although all the lower quality forecasts had similar accuracy and skill (the accuracy relative to a climatology reference forecast), the nature of the forecasts themselves were very different. By examining other aspects of forecast quality – either visually for a qualitative assessment or numerically with forecast verification measures for a quantitative assessment – one could easily diagnose the attributes that were degraded.

Some practical considerations of forecast quality for ensemble forecast verification were also discussed. For example, diagnosing the factors that contribute to forecast quality or degrade forecast quality is important and can help determine what ensemble postprocessing methods could be employed to improve the forecasts. Making inferences about forecast quality often requires the selection of appropriate measures (either absolute or relative measures). In practice, forecast verification is carried out with a verification data sample; given the assumptions of the distributions-oriented approach, care must be taken in selecting an appropriate data sample to avoid misleading results. Finally, given the typically small hydrometeorological ensemble forecast verification data samples, the sampling uncertainty associated with forecast verification measures should be assessed as part of the verification process.

References

- C. Accadia, S. Mariani, M. Casaioli, A. Lavagnini, A. Speranza, Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather Forecast.* **18**, 918–932 (2003)
- C. Accadia, S. Mariani, M. Casaioli, A. Lavagnini, A. Speranza, Verification of precipitation forecasts from two limited-area models over Italy and comparison with ECMWF forecasts using a resampling technique. *Weather Forecast.* **20**, 276–300 (2005)
- F. Atger, Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Monthly Weather Review.* **131**(8), 1509–1523 (2003)
- A.A. Bradley, S.S. Schwartz, Summary verification measures and their interpretation for ensemble forecasts. *Mon. Weather Rev.* **139**, 3075–3089 (2011)
- A.A. Bradley, T. Hashino, S.S. Schwartz, Distributions-oriented verification of probability forecasts for small data samples. *Weather Forecast.* **18**, 903–917 (2003)
- A.A. Bradley, S.S. Schwartz, T. Hashino, Distributions-oriented verification of ensemble streamflow predictions. *J. Hydrometeorol.* **5**, 532–545 (2004)
- A.A. Bradley, S.S. Schwartz, T. Hashino, Sampling uncertainty and confidence intervals for the Brier score and Brier skill score. *Weather Forecast.* **23**, 992–1006 (2008)
- G.W. Brier, Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950)
- G.W. Brier, R.A. Allen, Verification of weather forecasts, in *Compendium of Meteorology*, ed. by T.F. Malone (American Meteorological Society, Boston, 1951), pp. 841–848
- J.D. Brown, D.J. Seo, A nonparametric postprocessor for bias correction of hydrometeorological and hydrologic ensemble forecasts. *J. Hydrometeorol.* **11**, 642–665 (2010)
- J.D. Brown, L. Wu, M. He, S. Regonda, H. Lee, D.-J. Seo, Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble

- Forecast Service (HEFS): 1. Experimental design and forcing verification. *J. Hydrol.* **519**, 2869–2889 (2014)
- J. Demargne, J. Brown, Y.Q. Liu, D.J. Seo, L.M. Wu, Z. Toth, Y.J. Zhu, Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmos. Sci. Lett.* **11**, 114–122 (2010)
- F.J. Doblas-Reyes, V. Pavan, D.B. Stephenson, The skill of multi-model seasonal forecasts of the wintertime North Atlantic Oscillation. *Climate Dynam.* **21**, 501–514 (2003)
- E.E. Ebert, L.J. Wilson, B.G. Brown, P. Nurmi, H.E. Brooks, J. Bally, M. Jaeneke, Verification of nowcasts from the WWRP Sydney 2000 forecast demonstration project. *Weather Forecast.* **19**, 73–96 (2004)
- B. Efron, Nonparametric estimates of standard error: the jackknife, the bootstrap and other methods. *Biometrika* **68**, 589–599 (1981)
- C.A. Ferro, Comparing probabilistic forecasting systems with the Brier score. *Weather Forecast.* **22**, 1076–1088 (2007)
- K.J. Franz, H.C. Hartmann, S. Sorooshian, R. Bales, Verification of national weather service ensemble streamflow predictions for water supply forecasting in the Colorado River basin. *J. Hydrometeorol.* **4**, 1105–1118 (2003)
- T. Gneiting, A.E. Raftery, A.H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**, 1098–1118 (2005)
- T. Gneiting, F. Balabdaoui, A.E. Raftery, Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat Methodol.* **69**, 243–268 (2007)
- T. Gorgas and M. Dorninger, Quantifying verification uncertainty by reference data variation. *Meteorologische Zeitschrift.* **21**(3), 259–277 (2012)
- T.M. Hamill, Hypothesis tests for evaluating numerical precipitation forecasts. *Weather Forecast.* **14**, 155–167 (1999)
- T.M. Hamill, J. Juras, Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. Roy. Meteorol. Soc.* **132**, 2905–2923 (2006)
- T. Hashino, A.A. Bradley, S.S. Schwartz, Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.* **11**, 939–950 (2007)
- H. Hersbach, Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570 (2000)
- S. Jaun, B. Ahrens, Evaluation of a probabilistic hydrometeorological forecast system. *Hydrol. Earth Syst. Sci.* **13**, 1031–1043 (2009)
- I. Jolliffe, Uncertainty and inference for verification measures. *Weather Forecast.* **22**, 637–650 (2007)
- I.T. Jolliffe, D.B. Stephenson, Introduction, in *Forecast Verification*, ed. by I.T. Jolliffe, D.B. Stephenson (Wiley, Chichester, 2011), pp. 1–9
- T.L. Kane, B.G. Brown, Confidence intervals for some verification measures – a survey of several methods, in *Preprints, 15th Conference on Probability and Statistics in the Atmospheric Sciences* (American Meteorologic Society, Asheville, 2000)
- R. Krzysztofowicz, The case for probabilistic forecasting in hydrology. *J. Hydrol.* **249**, 2–9 (2001)
- S.J. Mason, N.E. Graham, Areas beneath the relative operating characteristics (ROC) and relative operating levels (ROL) curves: statistical significance and interpretation. *Q. J. Roy. Meteorol. Soc.* **128**, 2145–2166 (2002)
- S.J. Mason, G.M. Mimmack, The use of bootstrap confidence intervals for the correlation coefficient in climatology. *Theor. Appl. Climatol.* **45**, 229–233 (1992)
- J.E. Matheson, R.E. Winkler, Scoring rules for continuous probability distributions. *Manag. Sci.* **22**, 1087–1095 (1976)
- A.H. Murphy, Skill scores based on the mean-square error and their relationships to the correlation coefficient. *Mon. Weather Rev.* **116**, 2417–2425 (1988)
- A.H. Murphy, What is a good forecast – an essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**, 281–293 (1993)

- A.H. Murphy, Forecast verification, in *Economic Value of Weather and Climate Forecasts*, ed. by R.W. Katz, A.H. Murphy (Cambridge University Press, Cambridge, 1997), pp. 19–74
- A.H. Murphy, R.L. Winkler, A general framework for forecast verification. *Mon. Weather Rev.* **115**, 1330–1338 (1987)
- A.H. Murphy, R.L. Winkler, Diagnostic verification of probability forecasts. *Int. J. Forecast.* **7**, 435–455 (1992)
- D.N. Politis, J.P. Romano, The stationary bootstrap. *J. Am. Stat. Assoc.* **89**, 1303–1313 (1994)
- J.M. Potts, Basic concepts, in *Forecast Verification*, ed. by I.T. Jolliffe, D.B. Stephenson (Wiley, Chichester, 2011), pp. 11–29
- D.J. Seo, H.D. Herr, J.C. Schaake, A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci. Discuss.* **3**, 1987–2035 (2006)
- D.B. Stephenson, Use of the “odds ratio” for diagnosing forecast skill. *Weather Forecast.* **15**, 221–232 (2000)
- J.E. Thorne, D.B. Stephenson, How to judge the quality and value of weather forecast products. *Meteorol. Appl.* **8**, 307–314 (2001)
- D.A. Unger, A method to estimate the continuous ranked probability score, in *Nineth Conference on Probability and Statistics in Atmospheric Sciences* (American Meteorological Society, Virginia Beach, 1985), pp. 206–213
- E. Welles, S. Sorooshian, G. Carter, B. Olsen, Hydrologic verification – a call for action and collaboration. *Bull. Am. Meteorol. Soc.* **88**, 503 (2007)
- D.S. Wilks, Statistical significance of long-range “optimal climate normal” temperature and precipitation forecasts. *J. Climate* **9**, 827–839 (1996)
- D.S. Wilks, Sampling distributions of the Brier score and Brier skill score under serial dependence. *Q. J. Roy. Meteorol. Soc.* **136**, 2109–2118 (2010)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd edn. (Academic, Amsterdam, 2011), p. 704
- H. Zhang, T. Casey, Verification of categorical probability forecasts. *Weather Forecast.* **15**, 80–89 (2000)



Verification Metrics for Hydrological Ensemble Forecasts

François Anctil and Maria-Helena Ramos

Contents

1	Introduction	894
2	Metrics and Forecast Attributes	896
3	Common Metrics Applied to the Ensemble Mean	897
3.1	Relative Bias	897
3.2	MAE and RMSE	899
3.3	Pearson Correlation Coefficient	900
3.4	Contingency Tables and Associated Metrics for a Categorical Forecast	900
4	Common Metrics Applied to the Full Ensemble Forecast	902
4.1	Brier Score, for Event Outcomes	902
4.2	Continuous Ranked Probability Score, for Continuous Outcomes	902
4.3	Ignorance Score	904
5	Common Graphical Tools and Metrics Providing Visual Assessments	905
5.1	Rank Histogram and δ Ratio	905
5.2	Reliability Diagram	907
5.3	Probability Integral Transform Diagram	907
5.4	Spread-Skill Plot	909
6	Skill Scores	910
7	Common Metrics Evaluating Forecast Value	912
7.1	Relative Operating Characteristic	912
7.2	Cost-Loss Decision Model	913
8	Conclusion	914
	References	918

F. Anctil (✉)

Département de génie civil et de génie des eaux, Université Laval, Québec, QC, Canada
e-mail: Francois.Anctil@gci.ulaval.ca

M.-H. Ramos

IRSTEA, National Research Institute of Science and Technology for Environment and Agriculture,
UR HBAN, Antony, France
e-mail: maria-helena.ramos@irstea.fr

Abstract

This chapter reviews the most commonly used verification metrics for measuring the performance of hydrological ensemble forecasts. It links metrics to the different attributes of forecast quality and discusses the links between verification variables, metrics, and applications in a broad perspective. It provides an overview of the use of these metrics in forecast evaluation studies and general insights into what forecasters, practitioners, and end-users should consider when applying verification measures in the practice of hydrological ensemble forecasting.

Keywords

Forecast evaluation · Verification metrics · Skill · Scores · Hydrological ensemble forecasts

1 Introduction

The life of many hydrologists would be much simpler if they were able to design a perfect rainfall-runoff model to simulate and predict river discharges – and it is not by failure to attempt. Countless models have been proposed over the last years as progressive improvements over previous ones. Yet, none fully succeeds replicating the variety of streamflow conditions, as we can encounter in river catchments over the world, resulting from an assortment of local climatological, meteorological, and geological conditions, as well as land use and other anthropogenic impacts. Hydrological models remain abstractions of watershed systems, to the point that it is generally difficult to anticipate which model offers the greatest accuracy and predictive capability for specific catchments and hydrologic conditions. As a consequence, the verification of model outputs (often also referred to as “model evaluation”) is a continuing and necessary effort for modelers and operational forecasters to provide guidance in model development, implementation, and operational use.

Even if a perfect model was available to hydrologists, the imperfect knowledge of its input data would still be an issue when drawing conclusions about its predictive capability. Meteorological forcing data (observations for the past or predictions for the future) as well as hydrologic data (discharges or river levels), or any other type of physical measurements needed to set up the parameters of a model or to run real-time data assimilation schemes in forecasting systems, are just pieces of information, available at specific space and time scales, trying to represent the unknown (past or future) reality of complex hydrometeorological systems. They entail errors that propagate through the modeling chain and translate our lack of perfect knowledge into uncertainty that affects the system’s outputs (Ramos et al. 2010; Thiboult et al. 2017).

Living with uncertainties is probably rule number one in forecasting, and handling it is a key of success for many operational systems. For the benefit of forecasters and decision makers, efforts have been put into formalizing and communicating forecast uncertainty. They are a first and necessary step toward evaluating how confident one can be on the outputs of a system, given the errors

and uncertainties present in the whole forecasting chain, from its input data, through its models and techniques, up to the observations that will be used as reference to evaluate the predictions. Since uncertainties exist all along the prediction chain, forecast verification is required at several levels: from the evaluation of the meteorological input data to the assessment of hydrologic outputs at different conditions (e.g., before and after post-processing, with or without a hydrologic data assimilation scheme) and at different configurations (e.g., streamflow volumes or discharges over a critical threshold, single-valued or probabilistic forecasts), for a large span of scales in space and time. Forecast verification is therefore omnipresent in a hydrological ensemble prediction system. It is essential to every person making use of a forecast product to evaluate its value.

Forecast verification can help forecasters to gain confidence in their system and outputs, and help users in decision-making. It serves multiple actors (data providers, model developers, operational forecasters, decision makers, and other forecast users in general) and, consequently, can be performed in different ways depending on the focus of the evaluation and the decision problem stakeholders have to face at the end of the chain (Werner et al. 2016). Despite the different aims a forecaster or a user may have when performing forecast verification, they will often be looking at a number of traditional forecast attributes (e.g., bias, accuracy, reliability, sharpness, resolution; see Bradley et al. 2016) and computing metrics that can numerically convey the quality of the forecasts.

A large part of the performance of a forecast system is measured by the quality of its outputs, or by evaluating how well a forecast compares against a corresponding observation or best estimate of the true outcome. Forecast verification is thus, in short, the process of assessing the quality of a forecast. In hydrology, forecast verification is also called “forecast evaluation,” while in meteorology, the “value of a prediction” specifically refers to how a forecast helps the user make better decisions. Traditionally, forecast verification has been based on quantitative criteria of model performance to provide a systematic and objective evaluation of forecast quality. Plenty of metrics that are used in the verification of hydrologic forecasts originate from the meteorological community (Jolliffe and Stephenson 2012; Wilks 2011; Casati et al. 2008). Others have been developed or adapted to answer to specific needs in hydrology or to overcome typical constraints of hydrological systems (Pappenberger et al. 2008; Brown et al. 2010): for instance, to evaluate how accurate predictions of flood peaks are in magnitude and timing (e.g., Zappa et al. 2013; Liu et al. 2011) or to evaluate highly temporally correlated seasonal low flows (e.g., Wood et al. 2005; Trambauer et al. 2015; Yuan et al. 2015).

Forecast verification has also evolved in hydrologic ensemble prediction from an initial focus mainly put on evaluating ensemble predictions against deterministic forecasts (e.g., Bartholmes et al. 2009) to verification studies focusing on evaluating progresses made from the introduction of new methodological developments (post-processing or data assimilation schemes, for instance) in existing ensemble prediction chains (e.g., Zalachori et al. 2012; Verkade et al. 2013; Bourgin et al. 2014; Brown et al. 2014; Roulin and Vannitsem 2015). As hydrologic ensemble forecasts are more and more deployed for water and risk management, forecast verification in

hydrology has also expanded to encompass evaluating the added-value of ensembles in a decision-making context, through derived management variables or user-problem target verification (e.g., Kim et al. 2007; Roulin 2007; Boucher et al. 2012; Anghileri et al. 2016), fostering the links between the quest for good forecast quality and forecast value.

This chapter presents an overview of the most commonly used metrics for the evaluation of hydrological ensemble forecasts. It does not intend to give an exhaustive list of metrics, but rather it focuses on providing a broad perspective on practical uses of the metrics to applications in hydrological forecasting. We focus on the verification of ensemble forecasts, either when metrics are applied to their average value (e.g., mean or median of all ensemble members) or when they consider the full ensemble with its members. In all cases, we consider that verification metrics are applied to a given forecast lead time (or from similar lead times pooled together) and catchment (or river outlet) (indices are omitted for sake of simplicity). We also refer to “observation” as the variable against which the forecasts are verified. It can be a discharge measurement, a discharge value estimated from a rating curve, or a simulated discharge used as reference for verification purposes.

In the following, Sect. 2 provides a general introduction to metrics and forecast attributes. Section 3 details common metrics applied to the ensemble mean; Sect. 4, metrics applied to the full ensemble forecast; Sect. 5, graphical tools and metrics providing visual assessments; Sect. 6, skill scores; and Sect. 7, metrics evaluating forecast value. Concluding remarks and further examples are provided in Sect. 8.

2 Metrics and Forecast Attributes

The quality of a forecast can be assessed through a wide range of metrics (verification scores) (see, for instance, <http://www.cawcr.gov.au/projects/verification/>) (Wilks 2011; Jolliffe and Stephenson 2012). Many forecasters and practitioners recommend the use of several metrics to better assess the variety of attributes of a forecast (such as reliability, resolution, discrimination, and sharpness; see Bradley et al. 2016): “*any set of forecasts can then be ranked as best, second best, . . . , worst, according to a chosen score, though the ranking need not be the same for different choices of score*” (from: Jolliffe and Stephenson 2012, Chap. 1 Introduction, p. 5).

What metrics to choose depends on the user’s objectives as well as on the characteristics of the forecasts being evaluated. In several situations, it may be interesting to evaluate the relative quality of a forecast with regard to a reference system or a baseline (e.g., to decide which is better between system A and system B). Here, again, there exist several baselines that can be considered, varying from simple approaches such as climatology (always forecasting an average value) or persistence (the last observation is forecast to persist into the future) to more sophisticated benchmarks that take into account analogue-based features (see, for instance, the 23 benchmarks that were designed and used for the assessment of hydrological forecasts in the study proposed by Pappenberger et al. 2015).

Understanding the links between metrics and attributes is necessary to select metrics that are adapted to the system being evaluated and to the aims of the evaluation procedure. This can avoid misevaluation (drawing wrong conclusions), over-evaluation (using redundant metrics and applying excessive efforts to evaluate the same qualities), or under-evaluation (letting aside the evaluation of an attribute that could show off important strengths or weaknesses of the system).

A selection of common verification metrics applied to operational hydrometeorological forecasting and the quality attributes they measure are presented in Table 1 (based on Brown and Demargne 2013 [<https://hepex.irstea.fr/hepex-science-and-challenges-verification-of-ensemble-forecasts-24/>] and Bradley et al. 2016). Whether metrics are applied to deterministic or probabilistic forecasts of discrete (e.g., defined as the variable exceeding a threshold) or continuous events, robust verification analyses require at least long time series of homogeneous datasets over the verification period, and forecast-observation pairs that are representative of the predictive/decision problem at hand, which can be pooled together for the computation of statistic-based metrics. When using ensemble forecasts, it is also often assumed that the members of the ensemble forecast and the verifying observation are sampled from the same distribution.

3 Common Metrics Applied to the Ensemble Mean

We first examine metrics used to verify a single-valued forecast of a continuous forecast variable (e.g., streamflow), such as the relative bias, the MAE and RMSE, and the Pearson correlation coefficient. Secondly, we introduce the metrics derived from the contingency table, which is applied to categorical forecasts (e.g., streamflows exceeding a threshold). As usually done in ensemble forecasting, the ensemble mean will be used to represent a single-valued forecast of the outcome (other measures of central tendency could be used, such as the median or the mode). In this section, $Q_{avg}(k)$ is the k^{th} ensemble average (mean over all ensemble members at time step k) and $Q_{obs}(k)$ is the k^{th} observation of N forecast-observation pairs.

3.1 Relative Bias

To measure the overall (unconditional) bias, the relative bias (BIAS) is used. It measures the ratio between the mean of the ensemble average and the mean of observations, computed over all forecast-observation pairs:

$$\text{BIAS} = \frac{\sum_{k=1}^N Q_{avg}(k)}{\sum_{k=1}^N Q_{obs}(k)} \quad (1)$$

Values of BIAS higher (lower) than 1 indicate an overall overestimation (underestimation) of the observed values.

Table 1 Common verification metrics used in operational hydrometeorological forecasting

Quality attribute	Metric name	Type of forecast	Discrete events?
Bias (difference between average forecast and average observation)	Relative mean error (or relative bias) Frequency bias	Single-valued Both	No Yes
Accuracy (average difference between individual forecasts and observations)	Mean absolute error	Single-valued	No
	Mean square error	Single-valued	No
	Root mean square error	Single-valued	No
	Mean continuous rank probability score (CRPS)	Probabilistic	No
	Brier score	Probabilistic	Yes
	Critical success index (or threat score)	Both	Yes
Correlation (linear relationship between forecasts and observations)	Pearson correlation coefficient	Single-valued	No
	Spearman rank correlation	Single-valued	No
Skill (accuracy of forecast relative to a reference forecast)	Mean absolute error skill score	Single-valued	No
	Mean square error skill score	Single-valued	No
	Mean continuous rank probability skill score	Probabilistic	No
	Brier skill score	Probabilistic	Yes
	Equitable threat score (or Gilbert skill score)	Both	Yes
Reliability (agreement of forecast with observation conditioned on the forecast issued)	Mean square error reliability	Single-valued	No
	Mean CRPS reliability	Probabilistic	No
	Brier score reliability	Probabilistic	Yes
	Reliability diagram	Probabilistic	Yes
	Rank histogram	Probabilistic	Yes
Resolution (differences in outcomes for different forecasts issued)	Success ratio	Both	Yes
	Mean square error resolution	Single-valued	No
	Mean CRPS resolution	Probabilistic	No
	Brier score resolution	Probabilistic	Yes
Type-2 conditional bias (agreement of observation with forecast conditioned on outcome)	Mean square error type-2 conditional bias Brier score type-2 conditional bias	Single-valued Probabilistic	No Yes

(continued)

Table 1 (continued)

Quality attribute	Metric name	Type of forecast	Discrete events?
Discrimination (differences in forecasts for different outcomes)	Mean square error discrimination	Single-valued	No
	Brier score discrimination	Probabilistic	Yes
	Relative operating characteristic score	Both	Yes
	Relative operating characteristic diagram	Both	Yes
Probability of detection (or hit rate)	Probability of detection (or hit rate)	Both	Yes
	Probability of false detection (or false alarm rate)	Both	Yes
	Forecast frequency histogram	Probabilistic	Yes
Sharpness (degree of variability of the forecasts or concentration of the predictive distributions)	Average width of the prediction intervals	Probabilistic	No

3.2 MAE and RMSE

The MAE (mean absolute error) and the RMSE (root mean square error) are often used as measures of accuracy. They are given by:

$$\text{MAE} = \frac{1}{N} \sum_{k=1}^N |Q_{avg}(k) - Q_{obs}(k)| \quad (2)$$

$$\text{RMSE} = \left[\frac{1}{N} \sum_{k=1}^N (Q_{avg}(k) - Q_{obs}(k))^2 \right]^{1/2} \quad (3)$$

Both metrics retain the units of the forecast variable. The MAE applies a linear scoring rule (errors are equally weighted) and describes the average magnitude of the errors without considering their sign. The RMSE applies a quadratic scoring rule (errors are squared before they are averaged) and attributes more weight to large forecast errors. For both metrics, lower values indicate better forecasts. By definition, the RMSE is larger or equal to the MAE. A large difference between the MAE and the RMSE reveals large fluctuations within the error time series. When comparing values of the MAE or RMSE over several catchments, it is useful to make use of a relative measure. Most often, normalized metrics are obtained by dividing the metric of each catchment by the catchment mean observed streamflow, estimated over a historic period or over the verification period.

3.3 Pearson Correlation Coefficient

The Pearson correlation coefficient is a measure of the degree of linear dependence (association) between forecast and observed data. For a population, it is given by:

$$\rho = \frac{\text{cov}(Q_{avg}, Q_{obs})}{\sigma_{Q_{avg}} \sigma_{Q_{obs}}} \quad (4)$$

where cov is the covariance and σ , the standard deviation. Based on a sample of forecast (Q_{avg}) and observation (Q_{obs}) pairs, it is estimated by the sample correlation coefficient:

$$r = \frac{\sum_{k=1}^N (Q_{avg}(k) - \bar{Q}_{avg})(Q_{obs}(k) - \bar{Q}_{obs})}{\left[\sum_{k=1}^N (Q_{avg}(k) - \bar{Q}_{avg})^2 \sum_{k=1}^N (Q_{obs}(k) - \bar{Q}_{obs})^2 \right]^{1/2}} \quad (5)$$

for which $r = 1$ or -1 indicates a perfect positive or negative linear relationship, while the absence of such relationship leads to $r = 0$.

There exists no absolute guideline for interpreting the values of the Pearson coefficient of correlation, i.e., indicating for which threshold value one can say there is a negligible, weak, moderate, or strong correlation in the data analyzed. A satisfactory level of correlation depends on the process under scrutiny. Moreover, even if one rarely expects a strong nonlinear relationship between simulated and observed streamflow time series, it is always advisable to visually inspect their dependence by plotting the data used in order to ensure that the coefficient adequately represents the relationship. This is especially recommended since the Pearson coefficient of correlation is very sensitive to high extreme values (outliers) and insensitive to additive or proportional differences between observations and model simulations (Moriasi et al. 2007). Some people prefer to use the coefficient of determination, which is the squared-value of the coefficient of correlation and is more stringent.

3.4 Contingency Tables and Associated Metrics for a Categorical Forecast

This section describes common metrics used to verify forecasts for a discrete (event) outcome, such as flow above a given flood threshold. To verify such categorical forecasts, a contingency table, often presented in the form of a matrix, gives the counts (or absolute frequencies) for the estimation of the (sample) joint distribution of forecasts and observations (Wilks 2011). The simplest possible situation is a 2×2 contingency table of a categorical yes/no forecast (Table 2). In this case, the possible

Table 2 A 2×2 contingency table for forecast verification

		Event observed	
		Yes	No
Event forecast	Yes	Hits	False alarms
	No	Misses	Correct negatives

entries of the table are related to the following outcomes: “the event was/was not forecast” and “the event was/was not observed.” A perfect forecast system would only produce *hits* and *correct negatives*, and no *misses* or *false alarms*. The cells of a contingency table contain frequency counts, obtained over the sample given by the verification period. When expressed in terms of relative frequency (i.e., dividing each entry of the table by the sample size), the cell counts inform about the joint distribution of forecasts and observations and their marginal distributions. The contingency table thus describes the association between forecasts and observations (events).

To separate “yes” and “no” events, thresholds are either defined by critical events (e.g., operational thresholds above which flooding occurs) or based on streamflow quantiles obtained from long time series of observations (e.g., one may be interested in evaluating discharges above the 80% percentile of observed discharges), which are often used when comparing or pooling metrics for different catchments.

When evaluating if an event was forecast or not, in the case of ensemble forecasts, the ensemble mean is often used. However, it is also possible to consider that an event is forecast if at least $p\%$ of the ensemble members forecast the event – see, for instance, Atger (2001) or Randrianasolo et al. (2010) for examples in precipitation and streamflow verification.

For each configuration of a contingency table, several descriptive statistics can be computed, for instance:

- The probability of detection (or “hit rate”) gives the proportion of the observed “yes” events that were correctly forecast: $\text{POD} = \text{hits}/(\text{hits} + \text{misses})$. It ranges from 0 to 1 (perfect score).
- The false alarm ratio gives the proportion of the forecast “yes” events that actually did not occur: $\text{FAR} = \text{false alarms}/(\text{hits} + \text{false alarms})$. FAR, conditioned on forecasts, is a reliability statistic of the contingency table, ranging from 0 (perfect score) to 1.
- The probability of false detection (also called “false alarm rate”) is conditioned on observations rather than on forecasts: $\text{POFD} = \text{false alarms}/(\text{correct negatives} + \text{false alarms})$. It ranges from 0 to 1 (perfect score) and is a measure of forecast discrimination.
- The bias score (also known as “frequency bias”) measures the ratio of the frequency of forecast events to the frequency of observed events. It indicates whether the forecast system has a tendency to under-forecast (score < 1) or over-forecast (score > 1). It is given by the ratio between $(\text{hits} + \text{false alarms})$ and $(\text{hits} + \text{misses})$.

- Finally, a frequently used metric, particularly when the nonoccurrence of the event is more frequent than its occurrence, is the threat score (TS) or the critical success index (CSI). It is given by the number of hits divided by the total number of hits, misses, and false alarms. The worst possible score is 0 and the best is 1.

4 Common Metrics Applied to the Full Ensemble Forecast

4.1 Brier Score, for Event Outcomes

To verify a probabilistic forecast for a discrete outcome (e.g., flood event), the Brier score (BS) averages the squared differences between pairs of forecast probabilities and the corresponding binary occurrences of the observations. For each realization k of a forecast, p_k is the forecast probability of the event (in ensemble forecasting, this probability is often estimated by the ratio of ensemble members forecasting the event to the size of the ensemble), and o_k is an indicator of the occurrence of the event: it equals 1 if the event occurs and 0 if the event does not occur. As done for the contingency table, critical thresholds are usually used to define an event for the verification of an operational system. Quantile values from the observed streamflow record can also be considered as event thresholds, as it may also facilitate the comparison of forecast quality for different catchments.

For a given lead time, the BS is given by:

$$\text{BS} = \frac{1}{N} \sum_{k=1}^N (p_k - o_k)^2 \quad (6)$$

The Brier score is negatively oriented (smaller score, better performance) and has a minimum value of 0 for a perfect (deterministic) system. As a sum of squared errors, it weights larger errors more than smaller ones. The BS is an overall measure of accuracy, but can be decomposed into three terms of the probability error: reliability, resolution, and uncertainty (Murphy 1973). The decomposition of a score gives a more detailed insight into the performance of the forecast system for the particular event being evaluated.

4.2 Continuous Ranked Probability Score, for Continuous Outcomes

The continuous ranked probability score (CRPS) is probably the most used verification metric in hydrometeorological forecasting for continuous observations to evaluate the overall accuracy of the forecasts. It measures the quadratic distance between the cumulative distribution of the forecasts and the cumulative distribution of the observations (Hersbach 2000). The CRPS (or mean CRPS) is the average, across all forecast-observed pairs, of this integral square difference. Like the BS, the perfect score is 0, and the lower the CRPS, the better the overall

performance of the forecasts. The CRPS is equivalent to the Brier score integrated over all possible threshold values (Jolliffe and Stephenson 2012; Gneiting et al. 2005; Hersbach 2000).

Since the CRPS measures the difference between observations and forecasts expressed as cumulative distributions functions, it allows to consider uncertain observations in its calculation. For a given k^{th} pair of observation-ensemble forecasts, the CRPS is given by:

$$\text{CRPS}(k) = \int_{-\infty}^{\infty} [F'_k(x) - F_k^o]^2 dx \quad (7)$$

where $F'_k(x)$ is the predictive cumulative distribution function of the k^{th} realization, x is the predicted variable, and F_k^o is the corresponding observed cumulative distribution function. If the observation is a unique value, the Heaviside function $H(x \geq x_k)$, where x_k is the observed value, is used. It is a single step-function with the step from 0 to 1 at the observed value of the variable. In practice, and considering that the number of members is sufficient to estimate the predictive distribution, the CRPS is often computed discretely from the empirical predictive distribution of the streamflow, given by the ensemble forecasts.

It is interesting to notice that the error depicted by the CRPS is often not constant; the correspondence between higher observed flows and higher mean CRPS is typical. Figure 1 (left) illustrates this by comparing streamflow time series (inversed upper hydrograph in the right y-axis) and CRPS time series (left y-axis). CRPS values come from the evaluation of 9-day-ahead forecasts issued by an 800-member ensemble prediction system configured using 16 lumped hydrological models driven by the 50 weather ensemble forecasts from the European Centre for Medium-Range Weather Forecasts (from Brochero et al. 2011, Fig. 3).

Since the CRPS is the probabilistic equivalent to the absolute error – the mathematical proof of this equivalence may be found in Baringhaus and Franz (2004) and Székely and Rizzo (2005) – it has become a common procedure to compare the mean

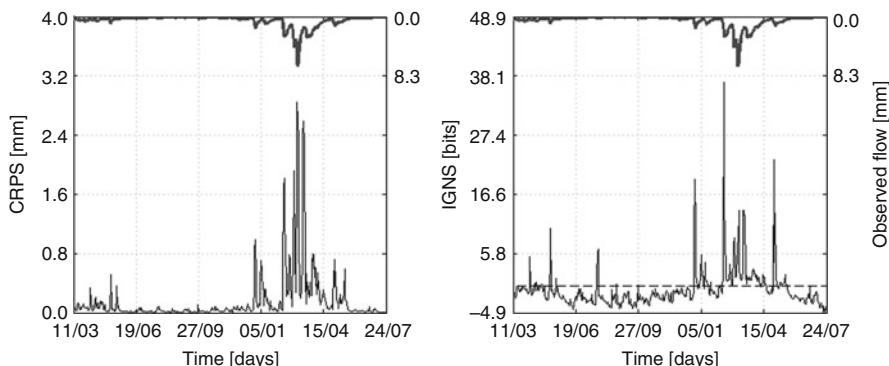


Fig. 1 Comparison of observed streamflow and prediction error time series as depicted by the CRPS (left) and the IGNS (right)

CRPS (average of CRPS values across all pairs of observation-forecast) and the mean absolute error (MAE) of a deterministic forecast. For an ensemble prediction system, the CRPS can also be decomposed into three components (reliability, resolution, and uncertainty), in a similar way to the decomposition of the BS (Hersbach 2000).

4.3 Ignorance Score

Proposed by Good (1952) as the logarithmic score, the ignorance score (IGNS) is an alternative to the Brier score, applicable to probability forecasts for nonbinary discrete events and to full continuous probability distribution forecasts (Wilks 2011). It is defined, for one realization, as the logarithm of the probability $f(x_k)$ assigned to the point x_k corresponding to the observation/event that actually occurred:

$$\text{IGNS} = -\log_2[f(x_k)] \quad (8)$$

The final metric is the average of the values obtained across all realizations of forecast-observed pairs.

The IGNS measures the ignorance or information deficit of a forecaster, compared to a person knowing the true outcome of the event, and ranges from zero (perfect forecast) to infinity. The score, described in detail by Roulston and Smith (2003), is used to evaluate the similarity of the ensemble forecast distribution to the distribution of the observed outcomes; it only depends on the probability assigned to the outcome that occurs, and not to any of the probabilities assigned to the other outcomes. It severely penalizes the bias, since positioning the observation in forecast regions of low probability lead to values that tend to infinity, as well as a larger ensemble spread. Smaller values indicate better performance, with the best value (zero) reached if a probability of 100% is assigned to the actual outcome.

The logarithmic score is highly sensitive to extreme cases (Gneiting and Raftery 2007). In order to rule out the possibility that the results solely reflect the effect of a few outliers, one may resort to trimmed means of the IGNS series excluding the highest and lowest 2% data values, following Weigend and Shi (2000). Infinite values may also be replaced by the next worst noninfinite value (e.g., Boucher et al. 2010). In ensemble forecasting, in order to avoid assigning zero probability to observed outcomes due to the finite sample of ensemble members, it is recommended to estimate the probabilities using plotting position formulas that cannot produce probabilities of zero or one (Wilks 2011). The ignorance score has connections with information theory (Roulston and Smith 2002) and the divergence score (Weijs et al. 2010). Figure 1 (right) compares streamflow and IGNS time series, as in Fig. 1 (left) for CRPS. Note that, in the IGNS case, there is no full correspondence between the higher IGNS values and the higher observed flow values.

5 Common Graphical Tools and Metrics Providing Visual Assessments

5.1 Rank Histogram and δ Ratio

The rank histogram is a graphical approach that was proposed independently by Anderson (1996), Hamill and Colucci (1997), and Talagrand et al. (1997). It evaluates the forecast attributes of reliability, consistency, and bias. In practice, it identifies the rank of the observation within ranked ensemble forecasts and evaluates whether the ensembles apparently include the observations being predicted as equiprobable members (Wilks 2011) or, in other terms, if the observation is equally likely to occur in each of the $d + 1$ bins of an ensemble with d members. The operation is repeated for all N forecasts and corresponding observations in the verification period. The rank histogram is obtained by constructing the histogram of the resulting N rank values.

The interpretation of the rank histogram is based on the assumption that all the members of the ensemble forecasts, along with the observation, have been drawn from the same distribution. Under this hypothesis, if the predictive distribution is well calibrated, then the rank histogram should be close to flatness. An asymmetrical histogram is usually an indication of a bias in the mean of the forecasts. If the rank histogram is symmetric and U-shaped, it may indicate that the predictive distribution is under-dispersed. If it has an arch form, the predictive distribution may be over-dispersed. The rank histogram may thus reveal deficiencies in the reliability attribute of the ensemble forecasts.

For a reliable system, over all $d + 1$ members, the number of elements in each bin of the rank histogram (S_c) has an expected value $N/(d + 1)$, while the deviation (Δ) of the histogram from flatness is measured by the following equation (Talagrand et al. 1997):

$$\Delta = \sum_{c=1}^{d+1} (S_c - h_{ref})^2 \text{ where } h_{ref} = \frac{N}{d + 1} \quad (9)$$

A reliable system has an expectation of $\Delta_0 = dN/d + 1$. The δ ratio ($\delta = \Delta/\Delta_0$), proposed by Talagrand et al. (1997), can be used as a measure of the reliability of an ensemble prediction system for a scalar variable. A ratio value that is considerably larger than 1 is a proof of lack of reliability. Figure 2 shows six examples of rank histograms for six different catchments, with results ranging from well-calibrated ensembles (flat histogram and small ratio value) to under-dispersed ensembles (U-shaped histograms and large ratio values) (from Velázquez et al. 2010, Fig. 5). The ratio metric facilitates comparing the performance of ensemble systems over catchments or assessing the evolution of the performance as a function of forecast lead time (e.g., Abaza et al. 2013).

Finally, some authors have pointed out a few flaws of the rank histogram. For instance, a U-shaped histogram, usually taken as a sign of under-dispersion, can sometimes be caused by conditional bias (Hamill 2001). Also, it must be noted that the rank histogram does not include a representation of the sharpness of the ensemble forecasts (Wilks 2011).

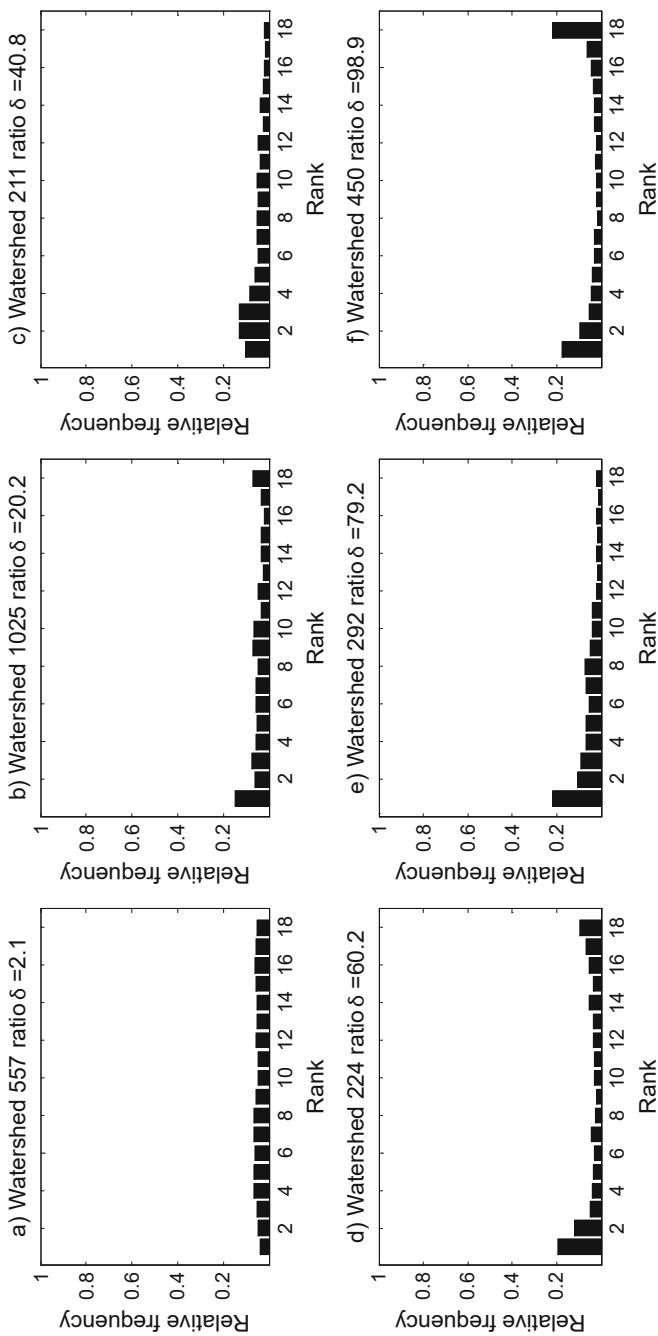


Fig. 2 Six examples of rank histograms with their respective ratio δ values

5.2 Reliability Diagram

The reliability diagram (sometimes also called “attribute diagram”) is a graphical method for assessing reliability, resolution, and sharpness of a probability forecast. It requires a fairly large dataset, since it is based on a partition in bins of the verification dataset into subsamples conditional on forecast probability. In practice, given that m denotes the different M thresholds of probability to assess, the reliability of the system can be directly measured from the comparison of these M thresholds with the conditional probability of observation as a function of the forecast. Since observation of the event is dichotomous ($r_k = 1$ if the event occurred and $r_k = 0$ otherwise) such conditional probability or relative frequency observed \bar{o}_m is given by:

$$\bar{o}_m = \frac{1}{N} \sum_{k=1}^N r_k \quad \text{where } r_k = \begin{cases} 1 & \text{if } x_k \in I_m \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where N is the number of forecast-observation pairs used in the verification. The goal is to have well-calibrated forecast systems, where the relative frequency is essentially equal to the probability of the forecast, i.e., $\bar{o}_m = I_m$ (Wilks 2011). The reliability diagram is the plot of the conditional probability versus the probability of the forecast.

In some cases, verification results can also be sensitive to sampling variability (Bradley et al. 2003; Clark and Slater 2006), and it may be worth assigning confidence limits to a reliability diagram using a bootstrap technique. Figure 3 shows the companion reliability plots to the rank histograms of Fig. 2, but for discharges larger than the 50% quantile of the observation time series (from Velázquez et al. 2010, Fig. 7).

5.3 Probability Integral Transform Diagram

Reliability is also often evaluated with the probability integral transform (PIT) diagram (Gneiting et al. 2007; Laio and Tamea 2007). The PIT diagram is the cumulative distribution of the PIT values, which are defined by the values of the predictive distribution function at the observations, computed at each time step. In the case of a reliable forecast, the observations uniformly fall within the predictive distribution and the PIT diagram coincides with the 1:1 diagonal (Fig. 4). If the PIT diagram is systematically above (below) the diagonal, the observed values are too frequently located in the lower (upper) parts of the forecast distribution, suggesting a systematic bias of the forecasts toward overprediction (underprediction). If the PIT diagram tends to resemble a horizontal line, observations fall too frequently in the tails of the forecast distribution, indicating that forecasts are too narrow. On the contrary, if the PIT diagram is closer to a vertical line, too many observations fall in the midrange of the forecast distribution, indicating that the forecasts make an ensemble that is too wide. In order to numerically compare results among catchments, the area between the curve of the PIT diagram and the 1:1 diagonal is often used (Renard et al. 2010): the smaller this area, the

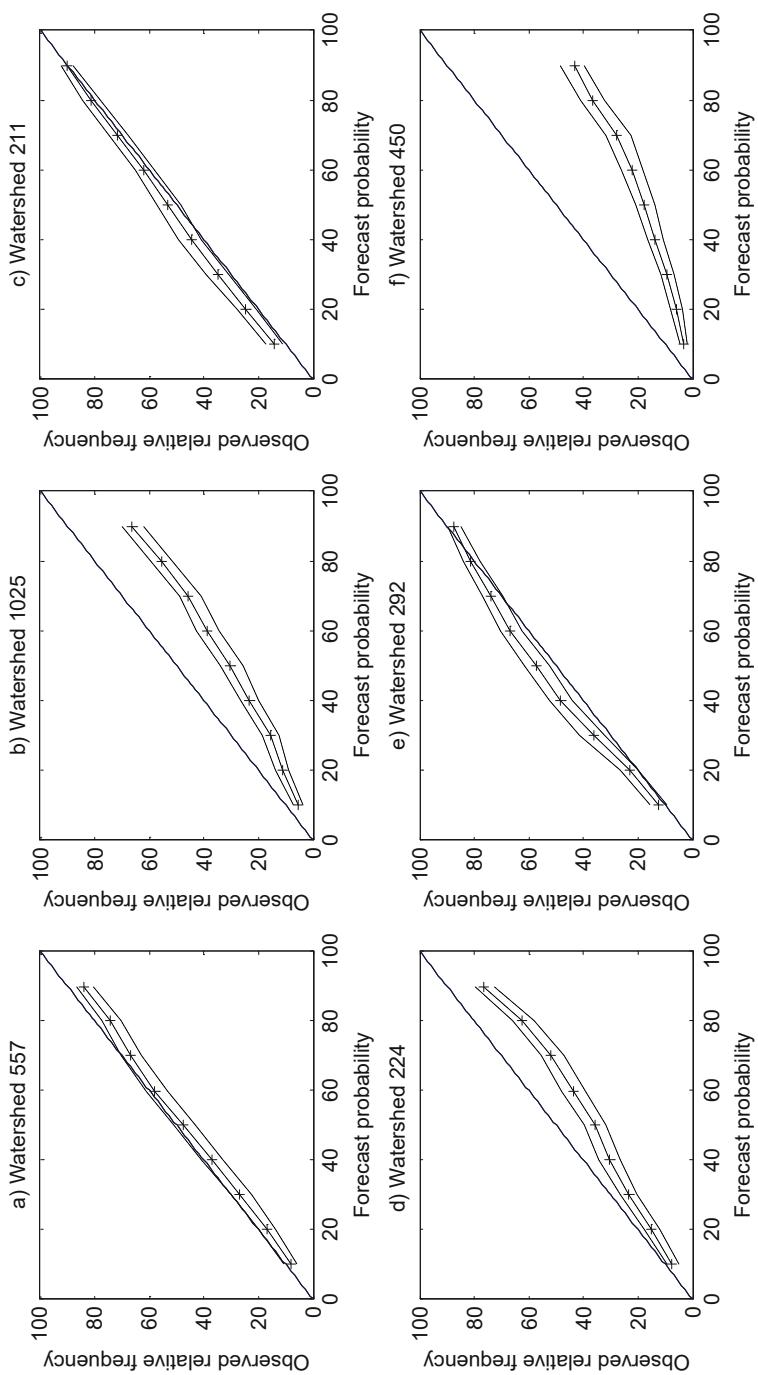


Fig. 3 Reliability diagrams for the same catchments as in Fig. 2 but for discharges larger than the 50% quantile of the observation time series. Dashed lines depict the 95% confidence interval

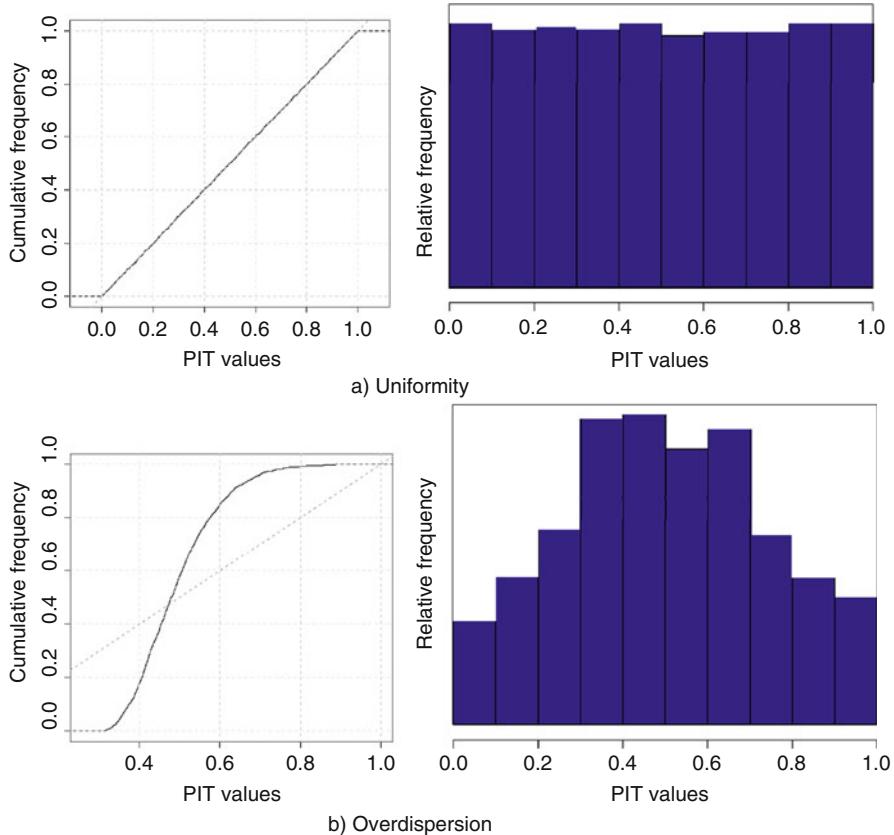


Fig. 4 Examples of PIT diagram (left) and PIT histogram (right) illustrating the visual assessment for (a) reliable ensemble forecasts and (b) overdispersive ensemble forecasts

more reliable the ensemble forecasts. The PIT diagram is an appealing tool since it does not require subjective binning of data or the use of a threshold to transform the forecasts into binary forecasts (see examples in Shrestha et al. 2015; Crochemore et al. 2016). Also, as noted in Wilks (2011), in the limit of infinite sample size, or if the ensemble distribution is represented as a smooth, continuous probability distribution function, the rank histogram is identical to the PIT histogram (i.e., the histogram of the PIT values). The same interpretive diagnostics for rank histograms apply to PIT histograms.

5.4 Spread-Skill Plot

The spread-skill plot evaluates the ability of the ensemble spread (variance) to depict the forecast error of the data, expressed as the RMSE of the ensemble means (Holt et al. 2009). One can conclude that the ensemble successfully predicts the

forecast error when the RMSE and spread are equal, but it is underdispersive whenever the RMSE is superior to the spread (Palmer et al. 2005). Assuming exchangeability between all ensemble members and a large ensemble size, the relationship between the RMSE and the ensemble spread may be approximated by Fortin et al. (2014):

$$\text{RMSE} \approx \left(\frac{1}{N} \sum_{k=1}^N s_k^2 \right)^{1/2} = \left(\bar{s}_k^2 \right)^{1/2} \quad (11)$$

where s_k^2 is the variance of the k^{th} predictive distribution. For smaller sizes (d) of ensembles, the following approximation should be used instead:

$$\text{RMSE} \approx \left[\left(\frac{d+1}{d} \right) \frac{1}{N} \sum_{k=1}^N s_k^2 \right]^{1/2} = \left(\frac{d+1}{d} \bar{s}_k^2 \right)^{1/2} \quad (12)$$

Figure 5 illustrates a typical application of the spread-skill plot, comparing the influence of meteorological ensemble forecasts of different resolutions (from Abaza et al. 2013, Fig. 6 of the Corrigendum). For the first catchment (Châteauguay, Québec, Canada), forecasts issued from a regional meteorological ensemble prediction system lead to lower RMSE and a better agreement between the RMSE and the spread, for forecast horizons greater than about 24 h – below this threshold, a postcalibration scheme is needed to compensate for the fact of not accounting for all principal sources of uncertainty (Thibault et al. 2016). Gains of using a higher resolution meteorological ensemble prediction system are less striking on the second catchment (Écorces, Québec, Canada), but it nonetheless improves the RMSE-spread agreement for forecast horizons greater than about 40 h.

6 Skill Scores

Skill scores are calculated to compare the accuracy of the forecasts over a reference forecast. They can measure the improvement of a given forecast relative to a reference forecast, for instance, or how much better one forecasting system is in comparison to another forecasting system given the same variables and location. They are thus a relative measure (or scaled representation) of forecast quality and therefore allow for meaningful forecast comparisons, for different locations or conditions for example.

The general skill score formulation for a given verification score and reference forecast is:

$$\text{Skill Score} = \frac{(\text{Score}_{\text{forecast}} - \text{Score}_{\text{reference}})}{(\text{Score}_{\text{perfect}} - \text{Score}_{\text{reference}})} \quad (13)$$

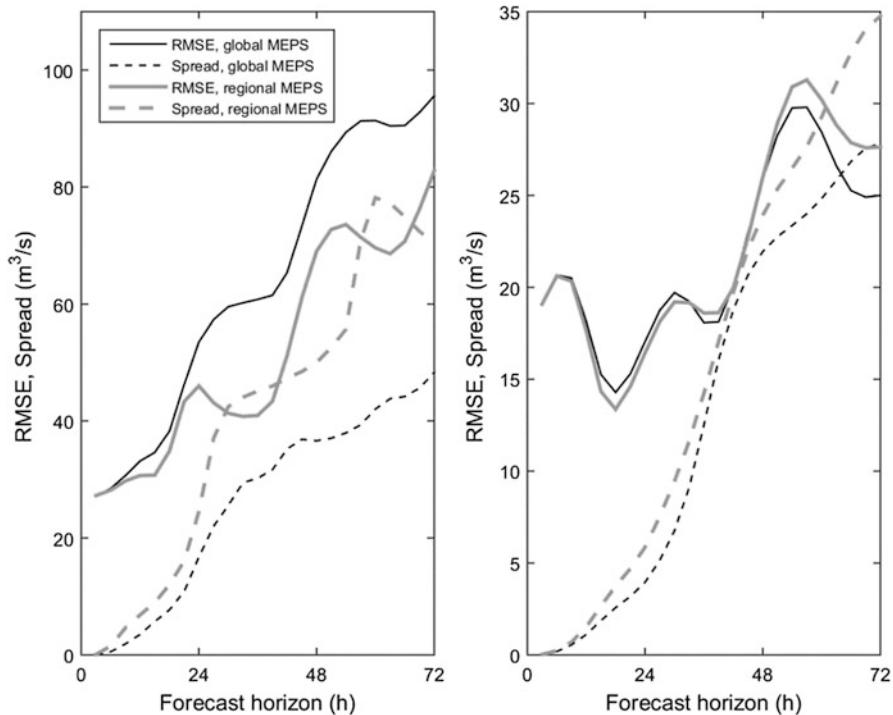


Fig. 5 Evolution of the RMSE and spread of a hydrological ensemble prediction system based either on a global (100 km) or a regional (33 km) meteorological ensemble prediction system, as a function of forecast horizon

where $\text{Score}_{\text{perfect}}$ is the value of the accuracy measure that would be achieved by perfect forecasts. For a metric for which the perfect score is zero, such as BS or CRPS, the skill score can be written as:

$$\text{Skill Score} = 1 - \frac{\text{Score}_{\text{forecast}}}{\text{Score}_{\text{reference}}} \quad (14)$$

Skill scores range from negative infinity to positive one. A perfect forecast yields skill values of 1. Scores equal to 0 mean that the forecast system has similar skill than the reference forecast. Negative (positive) scores indicate a poorer (better) forecasting system than the reference. A skill score is often interpreted as a percentage improvement over the reference forecast (Eq. 13 \times 100%; Wilks 2011).

Skill scores can be computed for many of the metrics used in forecast verification and are typically used when calculating the BS and CRPS scores. It is also worth noting that, in hydrology, the Nash and Sutcliffe Efficiency Index (EI) is a typical skill score long used to evaluate model simulations and determine whether the hydrological model is better than a one-parameter “no-knowledge” model that

gives the mean of the observations as simulation for all time steps. Another common skill score is the persistence index (PI) (Kitanidis and Bras 1980; Corradini et al. 1986; Anctil et al. 2004). It offers an alternative to the EI by exploiting a scaling term based on the information often used for data assimilation prior to issuing the forecast: the last available streamflow observation. The PI indicates whether the forecasts are on average better than a one-parameter “naïve” model that gives a prediction of the previous observation for all forecast horizons. PI statistics are particularly well designed for assessing short-range deterministic forecasts.

7 Common Metrics Evaluating Forecast Value

The measure of forecast value is associated to how a forecast system helps the user to make a better decision. This means that the context of evaluating forecast value can be very user-specific and have different focuses when one is interested in quantifying the added value of forecasts of good quality for specific applications: for instance, measuring the gain in lead time from anticipating drought conditions, the gain in insurance losses when forecasting floods and taking action upon, the gain in energy production from the use of improved forecasts in models of hydropower reservoir optimization, or evaluating transport costs by using calibrated probabilistic forecasts in a simulation model for river navigation.

Murphy (1993) wrote that “*...forecasts possess no intrinsic value. They acquire value through their ability to influence the decisions made by users of the forecasts.*” Investigating the forecast quality/value relationship can throw lights on the expected benefits of research and operational studies seeking to improve the quality of hydrometeorological forecasts. Some metrics and approaches for the evaluation of forecast quality can include variables relevant to decision-making in their metrics.

7.1 Relative Operating Characteristic

The relative operating characteristic (ROC) curve (Peterson et al. 1954; Mason 1982) is widely used in the evaluation of probabilistic or ensemble-based forecasts and is considered as a measure of their potential usefulness. It is used to determine how well forecast discriminates between events and nonevents, thus measuring forecast resolution. It is not sensitive to forecast bias and, therefore, it does not inform about reliability, but, instead, it allows noncalibrated forecasts (i.e., forecasts that have not been bias corrected) to be compared.

The ROC curve is a graphical representation where the probability of detection (POD) is plotted against the probability of false detection (POFD), both derived from contingency tables built by using different predictive probability thresholds (e.g., at least 5%, 25%, 50%, etc., of ensemble members exceeding a given critical threshold). The event is defined by the critical threshold, which is often

taken from the values of quantiles of the observation time series. A perfect forecast would imply a zero POFD, while the POD would be equal to 1. Thus, the closer the curve to the upper left corner of the ROC graph, the better the forecast. The area under the ROC (AUC) can be used to estimate this ability of the forecasts to correctly anticipate the occurrence or nonoccurrence of the events, and can be a useful metric to inter-compare forecasting systems. A perfect forecast system is represented by an AUC equal to 1, while an AUC below 0.5 indicates no skill (i.e., a forecast system where a hit is just as likely as a false alarm; a “random forecast”).

7.2 Cost-Loss Decision Model

The usefulness of a forecasting system may also be assessed from the cost-loss decision model defined by Richardson (2000). Consider a decision maker who must choose either to take action or to do nothing: the choice depending exclusively on the belief that a given weather event X will occur or not. If the event occurs and no action is taken, then a cost L has to be incurred. Taking action has a cost C whatever the outcome, but if the event occurs, a part of the loss L_1 is prevented. The decision maker is interested in following a strategy that minimizes the cost. When only historical information from long time series of past observations is available, there are two choices: always protect or never protect. The first instance results in an average expense of $C + \mu(L - L_1)$ where μ is the frequency of occurrence of the event in the historical record. If protective action is never taken, the average expense is μL . Therefore, the optimal strategy consists in always taking protective action if $C + \mu(L - L_1) < \mu L$ (that is if $C < \mu L_1$) and never taking protective action otherwise. The average expense E is then:

$$E_{\text{climate}} = \min\{C + \mu(L - L_1), \mu L\} \quad (15)$$

In the case of a perfect forecast, it is possible to take action only when the event is going to occur. The average expense would then be:

$$E_{\text{perfect}} = \mu(C + L - L_1) \quad (16)$$

We can define the relative value V of a forecast system as the reduction in expense in terms of a proportion of what would be obtained by a perfect forecast:

$$V = \frac{E_{\text{climate}} - E_{\text{forecast}}}{E_{\text{climate}} - E_{\text{perfect}}} \quad (17)$$

The decision maker will benefit from the forecast when $V > 0$.

Consider the case of a deterministic forecast, a hit rate H (or, as seen previously, POD) and a false-alarm rate F (or, POFD), V is a function of H , F , μ , and α (Richardson 2000):

$$V = \frac{\min(\alpha, \mu) - F\alpha(1 - \mu) + H\mu(1 - \alpha) - \mu}{\min(\alpha, \mu) - \mu\alpha} \quad (18)$$

where $\alpha = C/L_1$, the cost of taking an action expressed as a fraction of that part of the potential loss that is protected by that action, which is known as the cost-loss ratio.

The relative value depends only on the parameter α , which describes the decision-making situation, the parameter μ , which characterizes the hydrologic (climatic) context, and the rates H and F , which quantify the performance of the forecast system. When working with a probabilistic forecast, a decision maker has to choose the probability threshold at which action should be taken. For each value of the probability threshold, the relative value of the probabilistic forecast can be calculated. The user can then choose the value that results in the largest value of V .

Examples of applications of the cost-loss decision model in hydrology can be seen in Roulin (2007), Van den Bergh and Roulin (2010), Verkade and Werner (2011), Abaza et al. (2013), and Thibault et al. (2017). In the traditional application of the cost-loss model, the decision maker minimizes the expected expense and is considered risk neutral. Alternative approaches have recently emerged, where risk aversion when making decisions is explicitly considered through the use of a utility function – see an application to early warning flood systems in Matte et al. (2017), and the references therein. A typical application is illustrated in Fig. 6 (from Abaza et al. 2014, Fig. 16). As expected, the value of the forecasts decreases for longer forecast horizons.

8 Conclusion

Many verification metrics can be found in the literature to evaluate the quality of hydrometeorological forecasts according to their different attributes. In weather prediction, the Joint Working Group on Forecast Verification Research from the World Weather Research Programme (WWRP) and the Working Group on Numerical Experimentation (WGNE) maintains a reference website (<http://www.cawcr.gov.au/projects/verification/>) describing standard and newly developed verification metrics. In hydrologic prediction, the HEPEX website (www.hepex.org) has also proposed discussions and examples of applications, along with indications of freely available verification tools and packages, notably the Ensemble Verification System (EVS, Brown et al. 2010; Demargne et al. 2010), which has been broadly applied to the verification of hydrologic variables.

As many metrics exist, many applications of these metrics can also be found. Table 3 presents some (nonexhaustive) examples of applications in hydrological forecasting, illustrating the variety of purposes, contexts, and approaches in the literature. In particular, note that combinations of metrics, customized to the application and value to be demonstrated, are often considered. While the first applications of forecast verification in hydrology have mainly focused on using metrics traditionally applied in the evaluation of weather ensemble predictions, a growing tendency has been observed to adapt the use of verification metrics to hydrologic variables, at relevant spatial and temporal scales of dominant

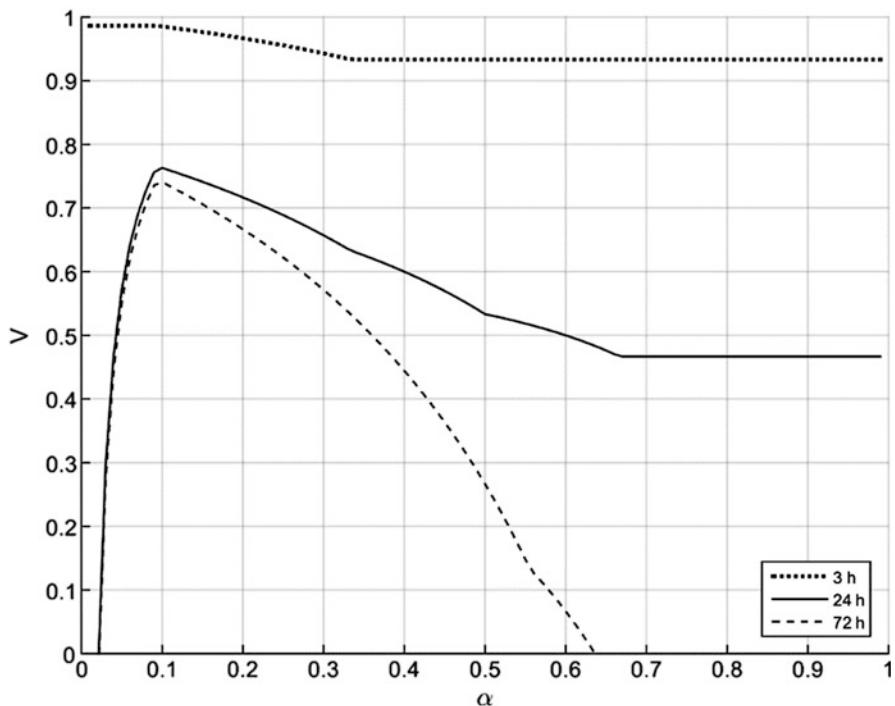


Fig. 6 Economic value V plotted against cost/loss ratio α for the 3 h, 24 h, and 72 h forecast horizons

hydrological process in a river basin and targeting water-related decisional contexts. Additionally, attention has been paid to analyzing hydrologic time series of forecast-observed pairs in respect to their typical characteristics (e.g., periodicity and persistence) and to the homogeneity of the hydrological processes behind their main features (e.g., high and low flows, rising and falling limbs of hydrographs). The aim is to better identify the strengths and weakness of a forecasting system when focusing on particular concerns of water resources and hydrometeorological risk managers.

A common conclusion arising from these diverse studies, and from the verification literature in general, is that there is no unique verification metrics that can reflect all the attributes one seeks to evaluate in a forecast. The use of several metrics is largely recommended. For “several,” one may still ask, “But, how many?” Certainly some metrics may be redundant in forecast assessment. The best guidance is to first choose at least one metric for each forecast attribute of interest. The choice of the most appropriate metrics for a given system and its particular use(s) is a step-by-step approach, with tries and errors, that need to be built in each case and with time, as experience is gained (see Cloke and Pappenberger 2008 for an example of a six-step approach to aid choosing measures and evaluate their suitability for a task at hand).

Ultimately, the goal of applying standard verification metrics, or developing new ones, is to use them efficiently to guide improvements over a current situation. As

Table 3 Some examples showing the diversity of applications of verification metrics in hydrology

Geographic context, spatial domain, variable	Metrics used	Verification period	Verification against
To diagnose how rainfall input and parametric uncertainty influence flow simulation uncertainty in a distributed hydrologic model (Carpenter and Georgakakos 2004):			
5 watersheds and subcatchments in the southern Central Plains of the United States. Hourly flow simulations, lead time 0 (simulation)	Ensemble dispersion through a normalized interquartile range (90th and 10th percentile)	25 to 30 flow events for each watershed between June 1993 and May 1999	Observed streamflows
To evaluate bias correction methods for ensemble streamflow volume forecasts (Hashino et al. 2007):			
Des Moines River basin, in Iowa, north-central USA. Monthly flow volumes, issued sequentially for each month, lead times of 1–12 months	Mean square error (MSE) skill score using climatology as a reference	1949 to 1996 historic data (forecasts generated on the first of each month)	Observed streamflows
To test a global approach to producing hydrological ensemble forecasts in river basins where in situ data are sparse (Voisin et al. 2011):			
4 outlets at the Ohio River basin, USA. Daily gridded runoff forecasts, lead times up to 15 days	Bias, RMSE, Pearson correlation, Rank histograms, CRPSS	2002–2007	A reference discharge simulation using observed meteorological data
To investigate the potential of radar-based ensemble flash-flood forecasts, including evaluation against deterministic discharge forecasts (Liechti et al. 2013):			
3 catchments in the southern Swiss Alps. Hourly runoff, lead time up to 8 h	Brier skill score, FAR and POD, ROC area	1389 hourly time steps (June 2007–December 2010)	Observed discharges
To establish a baseline for future enhancements, and to guide the operational use of the ensemble forecasting system studied (Brown et al. 2014):			
8 catchments in USA. Daily averages of streamflows, lead times 1–14 days, and time aggregated discharges	Relative mean error (RME) of the ensemble mean, correlation coefficient, CRPS, BSS, and decompositions, reliability diagram, ROC	Hindcasts for a 20-year period between 1979 and 1999	Observed streamflows and flow simulations using observed meteorological data

(continued)

Table 3 (continued)

Geographic context, spatial domain, variable	Metrics used	Verification period	Verification against
To evaluate a pan-European (EFAS, European Flood Awareness System) operational suite (Alfieri et al. 2014):			
38,452 grid points of the EFAS European river network. Daily streamflows, lead times up to 10 days	Nash–Sutcliffe efficiency, forecast bias, coefficient of variation of the RMSE, CRPSS	Operational forecasts and hindcasts from 2009 to 2012	A reference discharge simulation using observed meteorological data
To evaluate the benefits of using ensemble predictions for reservoir inflow to hydropower plants, in comparison to the deterministic values given by the control member of the ensemble and by the ensemble mean (Fan et al. 2014):			
São Francisco River basin and subcatchments, in Minas Gerais, Southeast Brazil. Hourly streamflow, lead time up to 16 days	MAE, CRPS, rank histogram, ROC curve, Brier skill score, visual inspection, threshold exceedance diagrams for flood events	Three wet seasons in 2010–2013 and three selected major flood events	Observed hydrographs and inflows estimated by water balance
To assess the impact of data assimilation for ESP seasonal water supply (Franz et al. 2014):			
North Fork of the American River Basin (NFARB) in northern California, USA. Water supply values (total discharge, m ³)	RMSE, percent bias, correlation coefficient, CRPSS, containing ratio, discrimination diagram, reliability diagram	26 to 58 years of historic data for January to April, depending on data availability	Discharges and SWE observations
To evaluate a postprocessing technique applied to construct probabilistic inflow forecasts for several catchments and lead times simultaneously (Engeland and Steinsland 2014):			
5 catchments in southwestern Norway. Daily discharges, lead time 1–10 days	Predictive QQ-plots, average width of the 95% forecast intervals, CRPS, energy score, Nash–Sutcliffe efficiency	1 September 2005–30 August 2009	Observed discharges
To investigate how data assimilation and postprocessing contribute to the skill of hydrological ensemble forecasts (Bourgin et al. 2014):			
202 catchments in France. Hourly streamflows, lead times up to 48 h	Bias, RMSE, PIT, normalized mean interquartile range (NMIQR), CRPSS	2005–2009	Observed discharges
To evaluate the assimilation of satellite soil moisture retrievals into a rainfall-runoff model for flood prediction in a large, sparsely monitored catchment (Alvarez-Garreton et al. 2015):			
1 semi-arid river basin in Queensland, Australia. Daily streamflow, lead time 0 (simulation)	Nash–Sutcliffe efficiency, RMSE, peak volume error, rank histogram, POD, FAR, CRPS	1 June 2003–2 March 2014	Streamflow records

(continued)

Table 3 (continued)

Geographic context, spatial domain, variable	Metrics used	Verification period	Verification against
To investigate the impact of errors in the forcing, in the model structure and parameters, and in the initial conditions on hydrological forecasts; to test the postprocessing of hydrological ensembles (Roulin and Vannitsem 2015):			
The Ourthe Orientale at Mabompré in the Ardennes region in Belgium. Daily discharges, lead time up to 10 days	Bias or mean error (ME), RMSE, spread, CRPS, and decomposition	March 2008–December 2012	Observed discharges and reference discharge simulation using observed meteorological data

many other innovations in forecasting technologies, verification metrics are useless unless they are effectively communicated, understood, and acted upon (Ramos et al. 2007; Demeritt et al. 2013; Werner et al. 2016). Results from forecast verification may help on carrying out improvements to be made to an existing system. Moreover, when communicated to forecast users and understood by them, they enable decision makers to make risk-based decisions that account for the past performance of the forecasting system for the situation at hand, which potentially leads to improved, more efficient decisions in many circumstances. They also facilitate the understanding of how both forecast errors and quality attributes cascade into forecast users' systems (for instance, reservoir operations or flood warning) and, therefore, support research developments to improve operational forecasts.

The efficient use and communication of verification information in real time forecasting is still a challenge for many forecasting systems (Werner et al. 2016). Future studies could investigate, for instance, how knowledge of past forecast performance in conditions similar to the real-time situation can guide forecasters when interpreting model outputs. Additionally, new developments can be expected on the evaluation of extreme (rare) events (e.g., major floods and droughts) and the reporting of the quality of a forecasting system in post-event analyses, when only a limited number of forecasts is available for the evaluation of highly impacting events. For these conditions, the need to evaluate the sampling uncertainty of the verification metrics is an additional challenge.

References

- M. Abaza, F. Anctil, V. Fortin, R. Turcotte, A comparison of the Canadian global and regional meteorological ensemble prediction systems for short-term hydrological forecasting. *Mon. Weather Rev.* **141**, 3462–3472 (2013). Corrigendum. *Mon. Weather Rev.* **142**, 2561–2562
- M. Abaza, F. Anctil, V. Fortin, R. Turcotte, Sequential streamflow assimilation for short-term hydrological ensemble forecasting. *J. Hydrol.* **519**, 2692–2706 (2014)
- L. Alfieri, F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, P. Salamon, Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.* **517**, 913–922 (2014)

- C. Alvarez-Garreton, D. Ryu, A.W. Western, C.-H. Su, W.T. Crow, D.E. Robertson, C. Leahy, Improving operational flood ensemble prediction by the assimilation of satellite soil moisture: comparison between lumped and semi-distributed schemes. *Hydrol. Earth Syst. Sci.* **19**, 1659–1676 (2015)
- F. Anctil, C. Michel, C. Perrin, V. Andréassian, A soil moisture index as an auxiliary ANN input for stream flow forecasting. *J. Hydrol.* **286**, 155–167 (2004)
- J.L. Anderson, A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Clim.* **9**, 1518–1530 (1996)
- D. Anghileri, N. Voisin, A.F. Castelletti, F. Pianosi, B. Nijssen, D.P. Lettenmaier, Value of long-term streamflow forecast to reservoir operations for water supply in snow-dominated catchments. *Water Resour. Res.* **52**(6), 4209–4225 (2016). <https://doi.org/10.1002/2015WR017864>
- F. Atger, Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Process. Geophys.* **8**, 401–417 (2001)
- L. Baringhaus, C. Franz, On a new multivariate two-sample test. *J. Multivar. Anal.* **88**(1), 190–206 (2004)
- J.C. Bartholmes, J. Thielen, M.H. Ramos, S. Gentilini, The European Flood Alert System EFAS – part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.* **13**(2), 141–153 (2009)
- M.A. Boucher, J.P. Laliberté, F. Anctil, An experiment on the evolution of an ensemble of neural networks for streamflow forecasting. *Hydrol. Earth Syst. Sci.* **14**, 603–612 (2010)
- M.-A. Boucher, D. Tremblay, L. Delorme, L. Perreault, F. Anctil, Hydro-economic assessment of hydrological forecasting systems. *J. Hydrol.* **416**, 133–144 (2012). <https://doi.org/10.1016/j.jhydrol.2011.11.042>
- F. Bourgin, M.-H. Ramos, G. Thirel, V. Andréassian, Investigating the interactions between data assimilation and post-processing in hydrological ensemble forecasting. *J. Hydrol.* **519**, 2775–2784 (2014)
- A.A. Bradley, T. Hashino, S.S. Schwartz, Distributions-oriented verification of probability forecasts for small data samples. *Weather Forecast.* **18**, 903–917 (2003)
- A.A. Bradley, J. Demargne, J.J. Franz, Attributes of forecast quality, in *Handbook of Hydrometeorological Ensemble Forecasting*, ed. by Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. Cloke, J. Schaake (Springer, Berlin/Heidelberg, 2016), 46p. https://doi.org/10.1007/978-3-642-40457-3_2-1
- D. Brochero, F. Anctil, C. Gagné, Simplifying a hydrological ensemble prediction system with a backward greedy selection of members, part I: optimization criteria. *Hydrol. Earth Syst. Sci.* **15**, 3307–3325 (2011)
- J.D. Brown, J. Demargne, D.J. Seo, Y. Liu, The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Softw.* **25**(7), 854–872 (2010). <https://doi.org/10.1016/j.envsoft.2010.01.009>
- J.D. Brown, M. He, S. Regonda, L. Wu, H. Lee, D.-J. Seo, Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS Hydrologic Ensemble Forecast Service (HEFS): 2. Streamflow verification. *J. Hydrol.* **519**, 2847–2868 (2014)
- T.M. Carpenter, K.P. Georgakakos, Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. *J. Hydrol.* **298**, 202–221 (2004)
- B. Casati, L.J. Wilson, D.B. Stephenson, Forecast verification: current status and future directions. *Meteorol. Appl.* **15**(1), 3–18 (2008)
- M.P. Clark, A.G. Slater, Probabilistic quantitative precipitation estimation in complex terrain. *J. Hydrometeorol.* **7**, 3–22 (2006)
- H. Cloke, F. Pappenberger, Evaluating forecasts of extreme events for hydrological applications: an approach for screening unfamiliar performance measures. *Meteorol. Appl.* **15**, 181–197 (2008)
- C. Corradini, F. Melone, L. Ubertini, A semi-distributed adaptive model for real-time flood forecasting. *Water Resour. Bull.* **22**, 1031–1038 (1986)

- L. Crochemore, M.-H. Ramos, F. Pappenberger, Bias correcting precipitation forecasts to improve the skill of seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.* **20**, 3601–3618 (2016). <https://doi.org/10.5194/hess-20-3601-2016>
- J. Demargne, J.D. Brown, Y. Liu, D.-J. Seo, L. Wu, Z. Toth, Y. Zhu, Diagnostic verification of hydrometeorological and hydrologic ensembles. *Atmos. Sci. Lett.* **11**(2), 114–122 (2010)
- D. Demeritt, S. Nobert, H.L. Cloke, F. Pappenberger, The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol. Process.* **27**, 147–157 (2013). <https://doi.org/10.1002/hyp.9419>
- K. Engeland, I. Steinsland, Probabilistic postprocessing models for flow forecasts for a system of catchments and several lead times. *Water Resour. Res.* **50**, 182–197 (2014). <https://doi.org/10.1002/2012WR012757>
- F.M. Fan, W. Collischonn, A. Meller, L.C.M. Botelho, Ensemble streamflow forecasting experiments in a tropical basin: the São Francisco River case study. *J. Hydrol.* **519**, 2906–2919 (2014)
- V. Fortin, M. Abaza, A. Anctil, R. Turcotte, Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeorol.* **15**, 1708–1713 (2014)
- K.J. Franz, T.S. Hogue, M. Barik, Assessment of SWE data assimilation for ensemble streamflow predictions. *J. Hydrol.* **519**(Part D), 2737–2746 (2014)
- T. Gneiting, A.E. Raftery, A.H. Westveld, T. Goldman, Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Weather Rev.* **133**, 1098–1118 (2005). <https://doi.org/10.1175/MWR2904.1>
- T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* **102**, 359–378 (2007)
- T. Gneiting, F. Balabdaoui, A.E. Raftery, Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **69**, 243–268 (2007)
- I.J. Good, Rational decisions. *J. R. Stat. Soc. C* **14**, 107–114 (1952)
- T. Hamill, Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Weather Rev.* **129**, 550–560 (2001)
- T.M. Hamill, S.J. Colucci, Verification of Eta RSM short-range ensemble forecasts. *Mon. Weather Rev.* **125**, 1312–1327 (1997)
- T. Hashino, A.A. Bradley, S.S. Schwartz, Evaluation of bias-correction methods for ensemble streamflow volume forecasts. *Hydrol. Earth Syst. Sci.* **11**, 939–950 (2007)
- H. Hersbach, Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570 (2000)
- H. Holt, J. Pullen, C. Bishop, Urban and ocean ensembles for improved meteorological and dispersion modeling of the coastal zone. *Tellus* **61A**, 232–249 (2009)
- I.T. Jolliffe, D.B. Stephenson, *Forecast Verification: A practitioner's Guide in Atmospheric Science*, 2nd edn. (Wiley, New York, 2012). <https://doi.org/10.1002/9781119960003>
- Y.-O. Kim, H. Eum, E.G. Lee, I.H. Ko, Optimizing operational policies of a Korean multireservoir system using sampling stochastic dynamic programming with ensemble streamflow prediction. *J. Water Resour. Plan. Manag.* **133**, 4–14 (2007). [https://doi.org/10.1061/\(ASCE\)0733-9496\(2007\)133:1\(4\)](https://doi.org/10.1061/(ASCE)0733-9496(2007)133:1(4))
- P.K. Kitanidis, R.L. Bras, Real-time forecasting with a conceptual hydrologic model. 2. Applications and results. *Water Resour. Res.* **16**(6), 1034–1044 (1980)
- F. Laio, S. Tamea, Verification tools for probabilistic forecasts of continuous hydrological variables. *Hydrol. Earth Syst. Sci.* **11**(4), 1267–1277 (2007)
- K. Liechti, M. Zappa, F. Fundel, U. Germann, The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrol. Earth Syst. Sci.* **17**, 3853–3869 (2013)
- Y. Liu, J.D. Brown, J. Demargne, D.-J. Seo, A wavelet-based approach to assessing timing errors in hydrologic predictions. *J. Hydrol.* **397**(3–4), 210–224 (2011)
- S.J. Mason, A model for assessment of weather forecast. *Aust. Meteorol. Mag.* **30**, 291–303 (1982)
- S. Matte, M.-A. Boucher, V. Boucher, T.-C. Fortier Filion, Moving beyond the cost–loss ratio: economic assessment of streamflow forecasts for a risk-averse decision maker. *Hydrol. Earth Syst. Sci.* **21**, 2967–2986 (2017)

- D.N. Moriasi, J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, T.L. Veith, Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Trans. ASABE* **50**, 885–900 (2007)
- A.H. Murphy, A new vector partition of the probability score. *J. Appl. Meteorol.* **12**(4), 595–600 (1973)
- A. Murphy, What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**(2), 281–293 (1993)
- T. Palmer, R. Buizza, R. Hagedorn, A. Lawrence, M. Leutbecher, L. Smith, Ensemble prediction: a pedagogical perspective. *ECMWF Newslet.* **106**, 10–17 (2005). ECMWF, Reading
- F. Pappenberger, K. Scipal, R. Buizza, Hydrological aspects of meteorological verification. *Atmos. Sci. Lett.* **9**, 43–52 (2008)
- F. Pappenberger, M.H. Ramos, H.L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, P. Salamon, How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.* **522**, 697–713 (2015)
- W.W. Peterson, T.G. Birdsall, W.C. Fox, The theory of signal detectability. *Trans. IRE Prof. Group Inf. Theory* **2–4**, 171–212 (1954)
- M.H. Ramos, J. Bartholmes, J. Thielen-del Pozo, Development of decision support products based on ensemble forecasts in the European Flood Alert System. *Atmos. Sci. Lett.* **8**, 113–119 (2007). <https://doi.org/10.1002/asl.161>
- M.H. Ramos, T. Mathevet, J. Thielen, F. Pappenberger, Communicating uncertainty in hydro-meteorological forecasts: mission impossible? *Meteorol. Appl.* **17**, 223–235 (2010)
- A. Randrianasolo, M.H. Ramos, G. Thirel, V. Andreassian, E. Martin, Comparing the scores of hydrological ensemble forecasts issued by two different hydrological models. *Atmos. Sci. Lett.* **11**(2), 100–107 (2010)
- B. Renard, D. Kavetski, G. Kuczera, M. Thyer, S.W. Franks, Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. *Water Resour. Res.* **46**, W05521 (2010)
- D.S. Richardson, Skill and relative economic value of ECMWF ensemble prediction system. *Q. J. R. Meteorol. Soc.* **126**, 649–667 (2000)
- E. Roulin, Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrol. Earth Syst. Sci.* **11**, 725–737 (2007)
- E. Roulin, S. Vannitsem, Post-processing of medium-range probabilistic hydrological forecasting: impact of forcing, initial conditions and model errors. *Hydrol. Process.* **29**(6), 1434–1449 (2015)
- M.S. Roulston, L.A. Smith, Evaluating probabilistic forecasts using information theory. *Mon. Weather Rev.* **130**, 1653–1660 (2002)
- M.S. Roulston, L.A. Smith, Combining dynamical and statistical ensembles. *Tellus A* **55**, 16–30 (2003). <https://doi.org/10.1034/j.1600-0870.2003.201378.x>
- D.L. Shrestha, D.E. Robertson, J.C. Bennett, Q.J. Wang, Improving precipitation forecasts by generating ensembles through postprocessing. *Mon. Weather Rev.* **143**, 3642–3663 (2015)
- G. Székely, M. Rizzo, A new test for multivariate normality. *J. Multivar. Anal.* **1**(93), 58–80 (2005)
- O. Talagrand, R. Vautard, B. Strauss, Evaluation of probabilistic prediction systems, in *Workshop on Predictability*, ed. by for Medium-Range Weather Forecasts, E. C., Shinfield Park, Reading (1997), pp. 1–25
- A. Thiboult, F. Anctil, M.-A. Boucher, Accounting for three sources of uncertainty in ensemble hydrological forecasting. *Hydrol. Earth Syst. Sci.* **20**, 1809–1825 (2016)
- A. Thiboult, F. Anctil, M.H. Ramos, How does the quantification of uncertainties affect the quality and value of flood early warning systems? *J. Hydrol.* **551**, 365–373 (2017). <https://doi.org/10.1016/j.jhydrol.2017.05.014>
- P. Trambauer, M. Werner, H.C. Winsemius, S. Maskey, E. Dutra, S. Uhlenbrook, Hydrological drought forecasting and skill assessment for the Limpopo River basin, southern Africa. *Hydrol. Earth Syst. Sci.* **19**, 1695–1711 (2015). <https://doi.org/10.5194/hess-19-1695-2015>
- J. Van den Bergh, E. Roulin, Hydrological ensemble prediction and verification for the Meuse and Scheldt basins. *Atmos. Sci. Lett.* **11**, 64–71 (2010). <https://doi.org/10.1002/asl.250>

- J.-A. Velázquez, F. Anctil, C. Perrin, Performance and reliability of multimodel hydrological ensemble simulations based on seventeen lumped models and a thousand catchments. *Hydrol. Earth Syst. Sci.* **14**, 2303–2317 (2010)
- J.S. Verkade, M.G.F. Werner, Estimating the benefits of single value and probability forecasting for flood warning. *Hydrol. Earth Syst. Sci.* **15**, 3751–3765 (2011)
- J.S. Verkade, J.D. Brown, P. Reggiani, A.H. Weerts, Post-processing ECMWF precipitation and temperature ensemble reforecasts for operational hydrologic forecasting at various spatial scales. *J. Hydrol.* **501**, 73–91 (2013)
- N. Voisin, F. Pappenberger, D.P. Lettenmaier, R. Buizza, J.C. Schaake, Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River Basin. *Weather Forecast.* **26**, 425–446 (2011)
- A.S. Weigend, S. Shi, Predicting daily probability distributions of S&P500 returns. *J. Forecast.* **19**, 375–392 (2000)
- S.V. Weijns, R. van Nooijen, N. van de Giesen, Kullback–Leibler divergence as a forecast skill score with classic reliability–resolution–uncertainty decomposition. *Mon. Weather Rev.* **138**, 3387–3399 (2010)
- K. Werner, J.S. Verkade, T.C. Pagano, Application of hydrological forecast verification information, in *Handbook of Hydrometeorological Ensemble Forecasting*, ed. by Q. Duan, F. Pappenberger, J. Thielen, A. Wood, H. Cloke, J. Schaake (Springer, Berlin/Heidelberg, 2016), 22p. https://doi.org/10.1007/978-3-642-40457-3_7-1
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences: An Introduction* (Academic, 2011), Amsterdam, 676p
- A.W. Wood, A. Kumar, D.P. Lettenmaier, A retrospective assessment of National Centers for Environmental Prediction climate model-based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res. Atmos.* **110**, D04105 (2005). <https://doi.org/10.1029/2004JD004508>
- X. Yuan, J. Roundy, E. Wood, J. Sheffield, Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins. *Bull. Am. Meteorol. Soc.*, 1895–1912 (2015). <https://doi.org/10.1175/BAMS-D-14-00003.1>
- I. Zalachori, M.H. Ramos, R. Garçon, T. Mathevret, J. Gailhard, Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Adv. Sci. Res.* **8**, 135–141 (2012)
- M. Zappa, F. Fundel, S. Jaun, A ‘Peak-Box’ approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. *Hydrol. Process.* **27**(1), 117–131 (2013). <https://doi.org/10.1002/hyp.9521>



Verification of Meteorological Forecasts for Hydrological Applications

Eric Gilleland, Florian Pappenberger, Barbara Brown,
Elizabeth Ebert, and David Richardson

Contents

1	Introduction	924
2	Meteorological Forecast Verification Aspects of Interest to Hydrology	925
2.1	Variables and Verifying Data	925
2.2	Benchmark and Skill	926
2.3	Threshold-Based Evaluation	926
2.4	Verification Area	927
2.5	Averaging, Smoothing, and Accumulation	927
3	Issues for the Verification of Point Forecasts in Meteorology	928
3.1	Matching Forecasts and Observations	928
3.2	Dealing with Inhomogeneity	930
3.3	Verifying Ensemble Forecasts for Point Locations	931
4	Issues for the Verification of Gridded Forecasts in Meteorology	931
4.1	Sources of Gridded Observations	931
4.2	Choosing and Preprocessing Gridded Observations for Verification Analysis	933
4.3	Double-Penalty and Small-Scale Errors	934
5	Methods for Verifying High-Resolution Gridded Forecasts	935
5.1	Overview	935
5.2	Filter Methods	936
5.3	Displacement Methods	937
6	Application of Spatial Techniques to Meteorological Ensemble Forecasts	938
7	Verification of Extreme Events	941

E. Gilleland (✉) · B. Brown

Research Applications Laboratory, Weather Systems and Assessment Program, National Center for Atmospheric Research NCAR, Boulder, CO, USA

e-mail: ericg@ucar.edu; bgb@ucar.edu

F. Pappenberger · D. Richardson

European Centre for Medium-Range Weather Forecasts, ECMWF, Reading, UK
e-mail: florian.pappenberger@ecmwf.int; david.richardson@ecmwf.int

E. Ebert

Research and Development Branch, Bureau of Meteorology, BoM, Melbourne, Australia
e-mail: beth.ebert@bom.gov.au

7.1	New Scores for Deterministic Forecasts	941
7.2	Verifying Extreme Events for Ensembles	942
7.3	Extreme Forecast Index (EFI)	943
8	Inference in the Presence of Correlation	944
9	Conclusions	946
	References	947

Abstract

This chapter illustrates how verification is conducted with operational meteorological ensemble forecasts. It focuses on the main aspects of importance to hydrological applications, such as verification of point and spatial precipitation forecasts, verification of temperature forecasts, verification of extreme meteorological events, and feature-based verification.

Keywords

Forecast verification · Model evaluation · Spatial statistics · Spatial forecast verification · Extreme-value verification

1 Introduction

Operational meteorological ensembles from numerical weather prediction (NWP) are regularly evaluated and verified to ensure that the forecasts are useful, to understand and monitor their quality and understand deficiencies to guide improvements of the forecasts. For example, Fig. 1 shows a verification of ensemble forecasts from various weather forecasting centers for precipitation over the extra tropics. Such forecasts are then used to drive hydrological models and produce

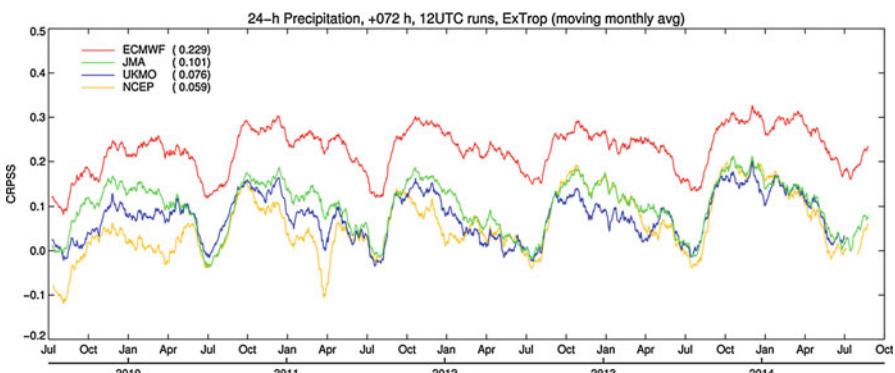


Fig. 1 Continuous rank probability skill score of ECMWF, JMA, UKMO, and NCEP for precipitation accumulated over 24-h and 3-day lead time (extratropics – poleward of 20°). The number next to the centers is the average score over the 5-year period. Skill is measured against climatology

forecasts for flood, drought, water resources, and other hydrological relevant variables.

In this chapter, some standard and more cutting-edge verification methods from meteorology that are relevant to hydrology are described. In particular, it is illustrated how verification is conducted with operational meteorological ensemble forecasts with a focus on the main aspects of importance to hydrological applications, such as verification of point and spatial precipitation forecasts, verification of temperature forecasts, verification of extreme meteorological events, and feature-based verification.

2 Meteorological Forecast Verification Aspects of Interest to Hydrology

2.1 Variables and Verifying Data

Meteorological forecasts produce a large number of variables both at single locations and on spatial fields. Some of them are directly used in hydrological models (e.g., precipitation or 2 m temperature), while others, such as geopotential height at 500 hPa or tropospheric moisture transports, influence hydrological predictability indirectly. For example, the flooding in Central Europe in 2013 was caused by a certain synoptic weather pattern which typically leads to heavy precipitation (Haiden et al. 2014). Atmospheric moisture transport is a prerequisite for heavy precipitation and, although both are correlated, they are not identical. Indeed, predictions for atmospheric moisture transports have a higher skill than precipitation forecasts. Therefore, verification of synoptic patterns is important, but cannot stand on its own if the forecast verification is supposed to be useful for hydrology. The evaluation of surface variables is essential.

The type and property of surface variable that is evaluated may significantly differ according to the hydrological application. A flash flood forecasting system, for instance, is interested in peak precipitation rates, whereas a hydrological forecast used for navigation may be more focused on daily average precipitation. In a tropical forecasting system, evaporation plays a major role and the quality of a forecast of evaporation is of high importance. Forecasts in snow-dominated catchments are more focused on forecasting snow melt, temperatures, and rain or snow events. This means that it is difficult for any meteorological forecasting center or provider to produce evaluation metrics that are relevant to all users; rather, many of them will be of general interest and some designed for specific end users.

Hydrologists also need to bear in mind that meteorological forecasts are sometimes verified against spatial “analyses.” An analysis field is often derived from many sources of observations (e.g., ground based or satellites) and merged in a data assimilation framework with a meteorological model. Hence, error structures will depend on the quality of the individual components. However, it is very often the only way to evaluate forecasts in a data-sparse environment on a larger scale

(e.g., global as it is done for the Global Flood Awareness System; Alfieri et al. 2012). Such an analysis is not equivalent to observations, which can have different error structures and dependencies. Therefore, verification against analyses can lead to different results than verification against observations. This issue is considered further in Sect. 4. In addition, most hydrological models are calibrated using observations. Therefore, deriving forecast performance quality from an analysis field needs to be treated with necessary caution. It may be necessary for hydrologists to conduct their own forecast verification, based on the same type of observations used for the calibration of the hydrological model.

2.2 Benchmark and Skill

To fully understand the performance and value of a forecast, it is necessary to evaluate forecasts against a benchmark. The most widely used hydrological score for benchmarking is the one introduced by Nash and Sutcliffe (1970), in which the hydrological model is compared against a climatological mean flow. A climatological reference is also widely used in meteorological verification. The choice of the benchmark strongly influences whether a model has skill or not (Pappenberger et al. 2015). For example, if one chooses to evaluate hydrological forecasts against a forecast that always predicts zero discharge, then it is very likely that the forecast will be classified as skillful. A more difficult benchmark, e.g., discharge persistence (last observed flow), may lead to lower skill. Hydrologists who use meteorological verification need to be aware that a skillful precipitation forecast does not necessarily mean that the discharge forecast will be skillful (and vice versa!). The hydrological model is a nonlinear transformation of the meteorological forecasts. For example, a climatological precipitation distribution (e.g., mean precipitation) will not lead to a climatological discharge distribution (e.g., mean discharge).

2.3 Threshold-Based Evaluation

Nonlinear transformation of a meteorological forecast through the hydrological model is important to consider when applying and interpreting verification metrics that are based on threshold exceedances. Flooding, for instance, is usually caused by intense precipitation exceeding a certain threshold for a certain time period. The exact characteristics of a meteorological event that may cause a flood depend strongly on particular characteristics of a catchment and its initial conditions (Merz and Bloeschl 2003). Standard, threshold-based meteorological evaluation approaches (e.g., precipitation exceeding 50 mm in 24 h) are therefore only relevant if the threshold considered is also relevant in the context of the hydrological application envisaged.

2.4 Verification Area

Meteorological verification often aggregates scores over latitude-longitude boxes corresponding to the grid of the model or the verifying analysis. Catchments are irregularly shaped, and a heavy precipitation event predicted in a rectangular box may actually fall in a neighboring catchment and hence produce a poor hydrological prediction. In addition, mean skill scores over the entire continent will only partially represent the true skill of even the largest catchments in the continent. Therefore, it is important to ensure that meteorological verification is done on an appropriate scale for the hydrological application in which it is used. Hydrologists may also need to assess the meteorological forecasts at the basin or subbasin scale to better understand the quality of the hydrological forecasts.

2.5 Averaging, Smoothing, and Accumulation

Hydrological models integrate components of a meteorological forecast over space, time, and multiple variables in a nonlinear way. For example, water from different parts of the Rhine catchment takes different amounts of time to reach the outlet. A drop of precipitation takes 7 days from the upper part of the catchment to the outlet, whereas precipitation falling close to the outlet of the catchment needs only 1 day. Therefore, a discharge prediction at the outlet will incorporate forecasts over a range of days as well as observations from multiple days. As an idealized example, assume a catchment has an upper, medium, and lower section. The travel time from the upper section to the outlet is 3 days, the travel time from the medium section to the outlet is 2 days, and the travel time from the lower section to the outlet is 1 day. In this imaginary catchment, all precipitation becomes discharge and no water is lost. A discharge forecast issued today for tomorrow will only contain precipitation from the past observations (observations from 2 days ago for the upper part, yesterday's precipitation for the medium part, and today's precipitation for the lower part). The 2-day forecast will incorporate the precipitation from the 1-day forecast in the lower areas of the catchment, today's observations in the medium part and yesterday's observations in the upper part. The 3-day discharge forecast still carries observed precipitation in the upper part from today and 1 day as well as 2 days of precipitation forecast for the lower and medium part, respectively. Only a 4-day discharge forecast will be purely dependent on meteorological precipitation forecasts of lead time 1, 2, and 3 days. Therefore a hydrological forecast not only spatially integrates but also mixes lead times and observations in a physically coherent manner, which would be difficult to assess in a precipitation-only verification study. In addition, as already mentioned, a discharge forecast is also influenced by forecast variables other than precipitation (e.g., evaporation, temperature), neatly dictating a physically coherent correlation structure between the variables.

The size of the catchment is therefore clearly of importance. It has been shown that forecasts for larger catchments tend to have greater skill simply due to the larger averaging effect. However, as the previous example illustrates, an incorrect spatial

covariance structure can lead to an amplified loss in skill in the discharge predictions of a catchment, similar to the double-penalty effect (see Sect. 4.2). Not all hydrological models have a grid-type structure, some of them are lumped. Calculating verification scores for a spatial field, a spatial average, or a single point in a catchment can lead to significantly different results; hence an evaluation of meteorological forecasts for hydrological applications has to be informed by the choice of the hydrological model.

In addition, not all parts of the catchment will contribute equally to the average error of the discharge forecast. Imagine a catchment in which one part has very high soil moisture, for example, due to specific geology, and another part does not. In the first part of the catchment, high soil moisture will lead to more discharge, whereas in the part with low soil moisture, most precipitation will become part of a groundwater store. This result means that any forecast errors in the high soil moisture area will be amplified, whereas in the other part, they will be damped. A simple spatial averaging of errors may not give appropriate information for a hydrological forecaster.

In summary, most meteorological forecast evaluations will have to be performed in conjunction with hydrological skill evaluations to allow for a full understanding of the forecast performance and allow for efficient diagnostics to improve the coupled hydrometeorological forecasting system.

3 Issues for the Verification of Point Forecasts in Meteorology

3.1 Matching Forecasts and Observations

While NWP forecasts are often verified against spatial analyses, measurements made at point locations remain the “gold standard” for small-scale accuracy. NWP meteorological forecasts of surface weather variables, such as temperature and precipitation, are routinely verified against observations from the global network of surface reporting stations coordinated by the WMO (2014). These synoptic observations (often referred to as SYNOPs) are made using a common reporting procedure and measurement standards. Typically, the gridded NWP data are interpolated to each observation location, often using the nearest model grid-point value for precipitation and a bilinear interpolation from the four nearest model grid points for temperature. Scores, such as bias, root-mean-square error (RMSE), standard deviation of error and mean absolute error (MAE) for deterministic forecasts, and Brier score and continuous ranked probability score (CRPS) for probability forecasts, are then computed and aggregated over geographical regions and over a large set of cases (months or seasons). For precipitation, scores are commonly calculated for a number of defined thresholds (e.g., 10 mm in 24 h), using a variety of contingency table-based measures including the Peirce skill score (PSS, sometimes called “true skill score”), the equitable threat score (ETS), and frequency bias index (FBI). Definitions of these and other commonly used verification scores can be found in Wilks (2011),

and examples of their application to NWP forecast verification are given by McBride and Ebert (2000) and Wolff et al. (2014).

A number of issues with this basic approach are discussed below. Nevertheless, it can provide useful information about the overall performance of an NWP system, highlighting important biases and illustrating changes in performance over time. For example, Fig. 2 shows the trend in mean error and standard deviation over the last 10 years of error for 2-m temperature forecasts over Europe for the ECMWF high-resolution model (HRES). Verification is against synoptic observations available on the Global Telecommunication System (GTS). A correction for the difference between model orography and station height was applied to the temperature forecasts, but no other post-processing was applied to the model output. There is a clear seasonal cycle in the scores as well as a general improvement over time in the standard deviation of the error. Differences between nighttime and daytime errors are most apparent in the biases. A recurring feature of the 2-m temperature forecast in recent years is a negative nighttime temperature bias in winter and early spring. Although Fig. 2 provides a useful summary, it inevitably hides much valuable information. For example, there is a significant geographical variation in bias due to different local climatological conditions, and a substantial part of the bias in some locations is related to particular meteorological situations. Too-rapid afternoon cooling in snow-covered forested areas in Scandinavia is one example that contributes to the overall cold bias.

A number of issues need to be addressed when verifying NWP forecasts against point observations. A fundamental issue is that the NWP model generates output as an area average over a grid box, with a grid length that can be only a few kilometers for a high-resolution limited area model or a few tens of kilometers for a global ensemble forecast system. If an EPS runs on a 30-km grid, the precipitation at each

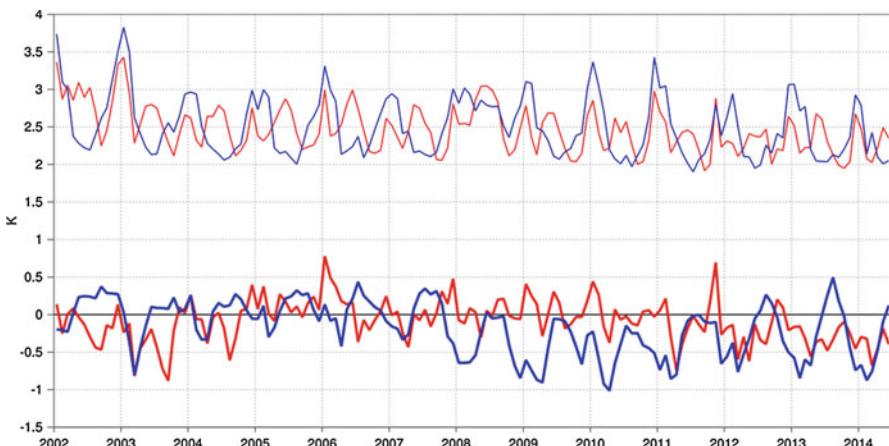


Fig. 2 Verification of 2-m temperature forecasts against European SYNOP data on the GTS for 60-h (nighttime, blue) and 72-h (daytime, red) forecasts. Lower pair of curves shows bias, and upper curves are standard deviation of error

grid point should be seen as representing the average value for a 1,000-km² area. In a situation with a small-scale severe weather event, such as an intense summer storm, there can be very large differences between the rainfall measured at an individual station and the average value over a larger region representative of a model grid box. There also can be significant discrepancies between model and observed temperatures: stations located within the same model grid box may have significantly different climatologies due to local variations, especially in areas of complex orography, coastal areas, or urban locations (urban heat island effects are generally not represented in NWP models). Such basic differences between the modeled variable and the observations are often referred to as representativeness issues. It is important to take account of these in forecast verification, especially in the verification of extreme events (Ebert et al. 2003; Goeber et al. 2008).

3.2 Dealing with Inhomogeneity

Observation sites are distributed very inhomogeneously across the globe. Even in Europe, where the meteorological ground station density is relatively high, less than 10% of land grid boxes (at 16-km resolution) contain one or more SYNOP stations that provide real-time data for general verification use. Thus, there is considerable undersampling, which renders verification results for quantities such as precipitation less robust. This situation becomes a particular issue when the duration of the verification sample is relatively short, such as is generally the case for studies that compare current and future operational NWP model configurations, when conditional verification is performed, or in the context of verification of extremes.

In order to generate robust results, it is often necessary to aggregate scores over a wide geographical area as well as over a significant period of time (typically a 3-month season). Differences in climate between stations can affect some skill scores (Hamill and Juras 2006). One way to address this issue is to scale the quantity to be verified by the local climate of the station. Doing so allows homogeneous samples to be accumulated over larger areas and in mountainous domains. For example, choosing a precipitation threshold as a quantile of the climate distribution may be more appropriate than selecting an absolute threshold such as 5 mm/day, which may occur much more frequently at one location than another (Jenkner et al. 2008).

The Stable Equitable Error in Probability Space (SEEPS) score (Rodwell et al. 2010; Haiden et al. 2012) was developed to mitigate many of these issues as far as possible for precipitation forecasts. SEEPS assesses rainfall forecasts in three categories (no rain, light rain, and heavy rain), where the boundary between “light” and “heavy” rain marks the highest one third (1/3) of the climatological cumulative rainfall distribution. The representativeness is further accounted for by applying the verification to precipitation accumulated over 24-h periods, which provides more relevant results than verification of instantaneous fields, such as precipitation rates. North et al. (2013) showed that SEEPS can also be used successfully for 6-h accumulations.

3.3 Verifying Ensemble Forecasts for Point Locations

The scores described in the previous two subsections can all be applied both to the individual members of an ensemble forecast and to the ensemble mean, verified as a deterministic forecast (e.g., Duda et al. 2014). One needs, however, to be aware of the intrinsic smoothing of the ensemble mean field in the interpretation of the verification results.

The information on distributions contained in an ensemble forecast also needs to be assessed. Different scores are used to assess the probabilistic aspects of ensemble forecasts. Ensemble forecasts for temperature and precipitation are both often verified using the continuous ranked probability score (CRPS), which assesses the probability distribution derived from the ensemble. All the issues of representativeness discussed above apply equally to the ensemble verification. An additional issue for the verification of ensemble forecasts in meteorology is in the interpretation of the spread and reliability of an ensemble prediction system (EPS). In a well-tuned ensemble system, the RMS error of the ensemble mean (EM) forecast should, on average, match the ensemble standard deviation (spread) (e.g., Molteni et al. 1996); note, however, that care must be used in selecting the method to estimate the standard deviation (Fortin et al. 2016). Typically, an EPS is underdispersive for surface weather elements, indicating that the EPS does not fully capture all the uncertainty in the forecasting system. However, it is important to account for errors, and particularly the representativeness, of the observations when comparing the EM error (computed against observations) with the spread (computed purely between model fields on the model grid). The representativeness can account for a substantial proportion of the apparent underdispersion of the ensemble forecast. A number of standard measures that assess the calibration or reliability of EPS rank histograms (PIT diagrams) and reliability diagrams are sensitive to representativeness issues, and care needs to be taken in interpretation of the results if the observation uncertainty is not properly taken into account (Saetra et al. 2004).

4 Issues for the Verification of Gridded Forecasts in Meteorology

4.1 Sources of Gridded Observations

Verification of gridded forecasts from NWP models against gridded observations addresses issues of spatial completeness and avoids the issue of different scales being represented by the model grid and point observations. Gridded observations, especially if available at high resolution, are particularly useful in estimating rainfall over catchments (which are on the mesoscale) in order to simulate and predict streamflows.

Mesoscale surface analyses, in particular, are useful for verifying model predictions of hydrometeorological variables such as temperature, wind, and humidity. Their higher resolution (1–10 km, typically) represents topographical effects

fairly well and offers the possibility to verify forecasts from both high- and low-resolution models. A disadvantage of many, if not most, mesoscale analyses is the use of a model short-range forecast as the first-guess field for the analysis of the surface-level observations. In this case, model biases are present in the final product to a greater or lesser degree, depending on the density of surface observations and how strongly the individual observations are weighted in the analysis scheme (Lazarus et al. 2002). The effect on the verification, particularly in model intercomparisons, is that the model being verified may be incorrectly assessed as more accurate than it actually is if it is the one supplying the first guess. If the observations are sufficiently dense in space, a model-independent mesoscale analysis (e.g., the Vienna Enhanced Resolution Analysis; Steinacker et al. 2000) would provide a better gridded dataset to use in verification. Information on the data density used in the analysis should be reported.

Rain gauge analyses have long been used to verify gridded precipitation forecasts. Most nations collect daily rainfall data at manned and automatic stations, and these can be mapped onto a grid using kriging or other objective analysis techniques. The number of gauges measuring rainfall at sub-daily time scales is much lower, so it is less common to produce gauge-based rainfall analyses for hourly or 3-hourly accumulations. Some nations have volunteer observing networks that report daily rainfall outside of real time. These additional observations can substantially boost the overall number of gauges, warranting a second nonreal-time rainfall analysis that is more accurate than the real-time analysis.

The accuracy of the gauge-based rainfall analysis depends on a number of factors. The most important is the station density, or the number of stations per analysis grid cell. Studies have shown that no improvement in the objective analysis scheme can adequately make up for a lack of sufficient observations (e.g., Bussières and Hogg 1989). A related factor is the spatial and temporal resolution of the analysis – finer resolution is prone to greater errors due, again, to sampling considerations. The nature of the rainfall itself also impacts on the accuracy, with convective rainfall more difficult to estimate accurately than large-scale rainfall because its spatiotemporal coherence is lower. Gauge measurements can be affected by wind-induced undercatch; this is particularly important in storm situations or when the precipitation falls as snow. When using gauge analyses for verification, it is important to take these factors into account.

Radar is another excellent source of gridded precipitation estimates in many parts of the world. The spatial resolution of radar data is typically 1–2 km, and the temporal resolution is 5–15 min. Radar quantitative precipitation estimates (QPE) are based on the reflection of microwave pulses by suspended hydrometeors, with the reflectivity transformed to rain rate through statistical fits to the data. Radar data first undergo a complex quality control process to remove artifacts from ground clutter, beam blockage, attenuation, range effects, biological targets, and anomalous propagation before they can be converted from reflectivity to rain rate. The traditional Z-R relationships that have been used for several decades are making way for more accurate QPE techniques based on data from newer dual-polarization radars that transmit horizontal and vertical pulses to collect more information on droplet

size, shape, and composition (Cifelli and Chandrasekar 2010). Even with the newer technology, there remain large uncertainties in the radar rainfall estimates. Radar QPE is more accurate within about 100 km from the radar, and bias correction using underlying rain gauge data is always recommended (Seo and Breidenbach 2002).

QPE mosaics from multiple radars in a network may provide high-resolution rainfall estimates over large domains that can be used for weather monitoring and nowcasting, rainfall estimation, assimilation into NWP models, and model verification. These mosaics are more spatially complete, although there may still be gaps between the radar rainfall fields. Often radar and gauge data are blended to produce a **multisensor analysis** that takes advantage of the greater accuracy of gauge measurements and the better spatial and temporal coverage of the radar observations, giving spatially complete high-resolution fields of hourly precipitation (e.g., NWS 2014; EUMETNET 2014).

Satellite observations offer an alternative source of gridded precipitation estimates that can be used for model verification, especially where gauge networks are inadequate and radar QPE are unavailable. Microwave rainfall estimates from several low earth-orbiting satellites, sometimes augmented with IR rainfall or cloud motion information from geostationary satellites, are combined into quasi-global rainfall fields at a number of centers. Because they are not limited to land areas, as is the case with gauge and radar, they can observe rainfall from weather systems approaching the land, which is very important in the case of tropical cyclones. Satellite rainfall can also be used to fill gaps in radar mosaics over land.

Kidd and Levizzani (2011) and IPWG (2016) provide a useful review of the status of satellite precipitation estimation. Temporal and spatial resolutions range from 30 min to 3 hourly, 8 km to 0.25° latitude/longitude, respectively. Most operational products are underpinned by radar and radiometer precipitation estimates from the TRMM (Tropical Rain Measuring Mission) satellite or more recently the GPM (Global Precipitation Measurement) core satellite launched in 2014, which has advanced detection capabilities enabling measurement of light rain and snow. Nevertheless, satellite precipitation retrievals based on passive microwave or IR data struggle to detect cool season rainfall from low clouds over land, leading to severe underestimates during mid-latitude winter. Bias correction and/or blending with gauge data can greatly improve the accuracy by reducing the effects of retrieval and sampling errors, as done with the TRMM 3B42 and GPM integrated multisatellite retrievals for GPM (IMERG) product (Huffman et al. 2007, 2013).

4.2 Choosing and Preprocessing Gridded Observations for Verification Analysis

In many cases, for hydrological purposes, the most important aspect of the forecast to get right is the location of the rain (it must be over the catchment of interest), followed by the quantitative amount. Spatial verification, discussed in the next section, is ideally suited for assessing the predicted location of rain, but requires high-quality spatial observations. Gauge analyses, radar, and satellite all provide

gridded rainfall observations with varying degrees of detail, accuracy, and coverage. The choice of data to use for verification will depend on the availability of observations and the spatial and temporal characteristics of the forecast. For verification of daily rainfall from a global NWP model, a rain gauge analysis may be adequate, provided the surface network is sufficiently dense to justify a grid that is similar in scale to the model. Where available, radar data that has been bias corrected and blended with gauge data provides additional spatial and temporal detail and is well suited for verifying shorter period accumulations, for example, from higher-resolution model forecasts. Satellite rainfall is the least accurate of the three data sources discussed above and should be used with caution when verifying model QPFs.

After identifying a gridded observation dataset, if the forecasts and observations are not at the same grid scale, it is necessary to map one or both onto a common grid for verification. The fairest approach is to map the finer scale product onto the grid of the coarser product because this method does not make unrealistic demands on the resolution of the coarser product. However, for some types of evaluation, such as verification of catchment average rain, it may make more sense to remap onto a finer grid in order to “cookie cut” the forecast and observations to fit the irregular area.

Methods used for grid remapping include interpolation, nearest-neighbor sampling, and area-weighted averaging. The choice of remapping method affects the verification results and should be appropriate for the variable being verified. For rainfall, area-weighted averaging is recommended in preference to interpolation because it preserves the total rainfall (important for hydrological applications) and does not smooth the intensity distribution (Accadia et al. 2003). For temperature, verification on a grid, bilinear interpolation is appropriate in x-y space, but it also is necessary to adjust the temperature for any vertical differences in altitude between the forecast and observed grid points. This adjustment can be done using a standard lapse rate or a lapse rate derived from the model forecast or analysis (Minder et al. 2010).

4.3 Double-Penalty and Small-Scale Errors

The inherent predictability of weather is usually lower at fine scales where local convective processes and topographic effects have a strong influence than at synoptic scales where larger-scale dynamical processes control the evolution of the fields. When verification is conducted at fine scales, forecast errors tend to be more pronounced because correctly predicting the fine detail is more difficult than predicting the larger-scale structure. For example, a model may predict precipitation at a certain location and time, but the observed precipitation occurs at a slightly different location and time. This displacement leads to the “double penalty” in verification, where the forecast is penalized for predicting rain when/where it did not occur and again for failing to predict rain when/where it did occur. The forecast may still provide very useful information on structure and intensity to a forecaster (who does not interpret the forecast at face value), but it typically scores worse than a

coarser-resolution forecast using standard grid-point-to-grid-point verification metrics, such as the RMSE (cf. Mass et al. 2002; Gilleland et al. 2009, 2010a).

Moreover, forecast verification at finer resolution puts greater demands on the observations. The difficulty in measuring precipitation accurately was discussed earlier. Spatial and temporal aggregation dampens the influence of random errors in the observations, increasing the accuracy of the analysis, but, at the same time, it removes useful information on location, timing, and intensity of observed events that may be relevant to the hydrological response. Small-scale errors in the observed rainfall can lead to poorer verification scores by introducing an additional error term (e.g., Ciach and Krajewski 1999), leading the user to believe that the forecasts are poorer than they truly are (Mass et al. 2002; Gilleland et al. 2009, 2010a).

5 Methods for Verifying High-Resolution Gridded Forecasts

5.1 Overview

To address the challenges to verification presented by high-resolution forecasts, numerous new verification methods have been proposed in order to inform users about forecast performance in a spatially meaningful way (Gilleland et al. 2009, 2010a; Brown et al. 2011; Gilleland 2013). The location and amplitude errors diagnosed by these schemes give hydrological users an indication of how much the meteorological forecasts can be trusted to predict the right amount of rainfall in the right place. The methods generally fall into two main categories: filter and displacement.

The category of **filter methods** (Ebert 2008) includes *neighborhood* methods, utilizing smoothing filters, and *scale separation* methods, based on band-pass filters. Smoothing filters can be applied to one or both of the forecast and observed raw or thresholded (binary) fields. The aim is to assess the performance at different scales and diagnose at which scale the forecast has sufficient accuracy to be useful. Band-pass filters quantify forecast performance at distinctly separate levels of detail that are considered to represent physically meaningful scales of interest (e.g., synoptic versus convective scale). Filter methods can be used to verify all types of precipitation fields and are particularly well suited for “messy” situations that may have no clearly identifiable rain features.

Displacement, or location, methods are primarily aimed at informing about positional errors and are especially well suited for situations when clearly identifiable rain features are present. In most cases, they include metrics for verifying intensity as well. This category is further divided into *feature-based methods*, which first identify features within the field (e.g., attempting to identify individual storms), and *field deformation* techniques, which are applied to the entire field. However, the latter techniques are often applied to specific features in addition to the entire field (e.g., Gilleland et al. 2008). Field deformation techniques attempt to morph the natural grid of the forecast field into one that better aligns with the observation field before applying the traditional grid-point-to-grid-point verification.

5.2 Filter Methods

The first class of filter methods is often called **neighborhood methods**. Here a smoothing filter replaces each grid-point value with an average or other statistical quantity based on the distribution of values from nearby grid points. The nearby grid points are determined by two main factors: the desired shape of the neighborhood (e.g., square or disk) and the size of the neighborhood.

The simplest neighborhood method is the upscaling method, which simply smooths the forecast so that small-scale and location errors have a much-reduced impact on the results. This method is often used to compare models with different spatial resolutions (WMO 2008).

When a smoothing filter is applied to a binary field (e.g., rain exceeding a given threshold), the value at a given grid point is interpreted as the event frequency within the neighborhood centered on that grid point. The fractions skill score (FSS) is one of the more commonly used smoothing filter methods at operational centers including the UK Met Office and MeteoSwiss (Roberts and Lean 2008; Weusthoff et al. 2010; Duc et al. 2013; Mittermaier et al. 2013). For each neighborhood size, the FSS is calculated as

$$\text{FSS} = 1 - \frac{\sum_{s=1}^n (\hat{p}_s - p_s)^2}{\sum_{s=1}^n \hat{p}_s^2 + \sum_{s=1}^n p_s^2}$$

where p and \hat{p} represent the frequency of observed and forecast events, respectively, in the neighborhood centered on point s , and n is the number of grid points in the field. A perfect forecast at a given scale achieves $\text{FSS} = 1$. One useful application is to identify at which scale (neighborhood size) the forecast has better skill (in terms of FSS) than a randomly distributed forecast with mean observed frequency \bar{p} (Mittermaier and Roberts 2010).

Scale separation methods utilize a spectral representation of a function, whereby the function is written as a sum of coefficients multiplied by basis functions. For example, in Fourier decomposition, the function is represented by sine and cosine basis functions, whereas a wavelet decomposition utilizes locally supported basis functions from a relatively wide class of possibilities. Another smoothing filter approach borderlines on this band-pass category in that it applies a wavelet decomposition to one or both fields, sets the smallest wavelet coefficients to zero, and recomposes the field in question (Briggs and Levine 1997). The result is a smoother field so that the technique is analogous to the upscaling method.

Briggs and Levine (1997) also apply traditional verification measures like anomaly correlation and RMSE to the component fields of the wavelet decomposition, which analyzes forecast performance in a similar way as applying verification measures to a series at each frequency of the spectral decomposition of a time series. This latter technique is a scale separation, or band-pass filter, technique. Casati et al. (2004) and Casati (2010) apply a similar technique, but to the differences in binary fields of the forecast minus the observations. One note when applying wavelet

decomposition techniques, especially for precipitation fields, is that care must be taken in choosing the specific wavelet basis function. For example, the Daubechies wavelet used by Briggs and Levine (1997) will generally result in wavelet artifacts that could lead to spurious results; for precipitation, a Haar wavelet as used by Casati et al. (2004) might be better suited.

5.3 Displacement Methods

The two primary categories for displacement methods include the *field deformation* approaches, which are applied to the entire fields, and the *feature-based* approaches, which are applied to individual features (such as precipitation areas) within the fields. The *field deformation* methods are all applicable on the features in the latter case. The *feature-based* approaches allow for additional analyses, such as a spatially meaningful contingency table and substantially greater diagnostic information about feature attributes that can be aggregated nicely over time to give highly informative results about the character of forecast errors.

Some of the earliest **field deformation** approaches attempted to realign the forecast grid so that the intensity values better match those of the observation field. The magnitude of the shifting, distortion, and amplification required to match the forecast with the observations is a measure of the forecast error.

The simplest field deformation methods simply move the entire grid around rigidly until some objective function is minimized (e.g., Hoffman et al. 1995; Ebert and McBride 2000). Other techniques similarly minimize an objective function, but individual grid points can move independently (usually after performing a rigid or affine transformation) in order to allow for nonlinear deformations of the grid (e.g., Alexander et al. 1998; Keil and Craig 2009; Gilleland et al. 2010c). Once a deformation has been found, numerous possibilities for summarizing information result. Direct inspection of vector fields showing the optimal deformation (possibly aggregated over time or space) allows for a visual display of region-specific location errors. Rigid transformations give readily interpretable information about the large-scale amount of location errors in specific directions.

Simpler than warping a field is to apply a binary image summary measure (e.g., Hausdorff distance, partial Hausdorff distance, Baddeley's delta metric, mean error distance, etc.) to the forecast and observed binary fields (such as the event field determined by where the value exceeds a given threshold; see e.g., Gilleland 2011; Schwedler and Baldwin 2011). Such methods, generally, all give the same information, are quick to compute, and give useful summaries of forecast performance in terms of location/pattern errors.

AghaKouchak et al. (2010) introduced three geometrical binary field summary measures that can be used to compare the geometrical properties of predicted and observed spatial fields. The connectivity index measures the connectivity and organization of the field, the shape index measures its roundness, and the area index measures the dispersiveness of the pattern. This last measure was also

introduced as a property for individual features in the technique introduced by Davis et al. (2006a, b, 2009) described below.

Feature-based methods, sometimes referred to as object based, give information about forecast performance for specific features within a field. For example, the contiguous rain area (CRA) method of Ebert and McBride (2000) is used at the Australian Bureau of Meteorology. It uses a rigid transformation to estimate the displacement error. The size, intensity, and pattern of the matched features in the forecasts and observations can be directly compared, and the fractional contributions from each type of error can be estimated. Various summary measures are included in the feature-based technique referred to as model object-based diagnostic evaluation (MODE, Davis et al. 2006a, b, 2009) along with other feature-specific information such as feature centroid, major axis angle, intersection area, etc. The MODE technique is used at the NWS Weather Prediction Center to operationally verify precipitation forecasts over the United States from a number of NWP models (<http://www.hpc.ncep.noaa.gov/verification/mode/mode.php>).

As an extension of the feature-based methods, the structure-amplitude-location (SAL) method of Wernli et al. (2008) is used at the Finnish Meteorological Institute and other operational centers. This method evaluates the realism of model precipitation features in a spatial region such as a hydrological basin, without attempting to match individual forecast and observed features. The location error is a function of the difference between the forecast and observed centers of mass of the rainfall fields. The differences between the forecast and observed structure (“peakiness”) and area-averaged amplitude reflect errors in the rainfall intensity distribution, which is important for runoff calculations.

Many of the above methods are clearly aimed at handling spatial displacement errors in order to minimize the impact of double penalties. It is possible, and perhaps worthwhile, to know whether or not an error is actually a spatial displacement, or whether it is merely a timing error that subsequently results in an error of spatial displacement. Some work has been done to incorporate space and time into the verification analyses, but as these methods are newer, they are not covered in full here (e.g., Gilleland et al. 2010b; Bullock 2011; Zimmer and Wernli 2011; Ebert et al. 2013).

6 Application of Spatial Techniques to Meteorological Ensemble Forecasts

When considering the application of spatial verification methods to meteorological ensemble predictions on a spatial grid, the standard approaches for ensemble verification can be applied. In particular, ensemble predictions can be treated in three standard ways: (i) as a set of *individual deterministic forecasts*, (ii) as a *probabilistic forecast*, and (iii) as a *distribution of forecasts*.

With regard to (i), the individual ensemble member forecasts can be evaluated using any of the spatial methods described in Sect. 5, leading to a set (or distribution) of verification results; or the ensemble mean (or some other representative forecast)

can be evaluated, which would provide a measure of the “average” spatial forecast performance (although losing some potentially useful information about the forecast uncertainty).

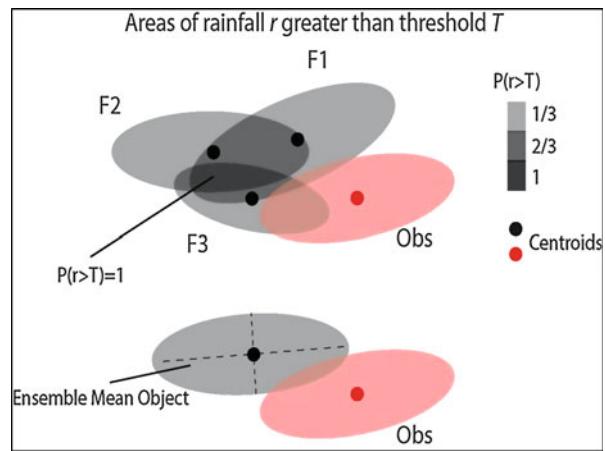
With regard to approach (ii), the ensemble predictions on a grid can be transformed to a grid of probability values for a specific event (e.g., the exceedance of a threshold value of precipitation) either directly by counting the members exceeding the threshold or by application of a more advanced post-processing method. This approach leads to a spatial probabilistic forecast field for a specific event. The probability field itself often has spatially coherent structures that lend themselves well to application of spatial verification methods. These probability fields can then be compared to the observed binary fields associated with the same specified threshold (used to define the probabilities). Spatial verification methods can be applied to these derived probability fields in the same way that they are applied to the raw forecast fields. Although approach (ii) has not been applied frequently in practice, it appears that all of the spatial methods described in Sect. 4 could be useful in such evaluations.

The feature-based approaches, in general, are well suited to alternative (iii), in which the characteristics of the features and their differences can be examined and summarized. For example, it can be highly informative to know if a forecast projects the right number, size, shape, location, and orientation of features within a study domain, regardless of whether or not it matches with the observations at every time step.

Treating an ensemble of predictions as a set of individual deterministic forecasts [approach (i)] is perhaps the most straightforward option for evaluation of ensembles using the various spatial methods. In particular, any spatial method can be applied to the individual members. Then the performance of the members can be evaluated by ranking them according to a particular spatial verification measure (e.g., Keil and Craig 2007; Micheas et al. 2007). Alternatively, the distribution of a particular spatial statistic for one ensemble forecasting system can be summarized and compared to the distributions for competing systems. Overall performance of an ensemble can be summarized using information from the distribution of performance of the individual members. For example, Atger (2001) and Marsigli et al. (2008) applied neighborhood approaches to examine the distributions of characteristics of ensemble member performance, and Zacharov and Rezacova (2009) used an average FSS value (based on the FSS values for individual ensemble members) to summarize the spread-skill relationship for high-resolution ensemble predictions of precipitation.

Application of feature-based verification approaches to ensemble forecast evaluation provides a straightforward approach toward merging spatial and ensemble forecast evaluation methods (Wernli et al. 2008). The diagram in Fig. 3 demonstrates three different approaches that might be taken to examine an ensemble feature prediction. The figure shows an ensemble forecast of a precipitation area (features F1, F2, and F3) compared to a single observed precipitation feature (note that, in this example, the intensity of precipitation is not considered; we are only concerned with the shape and location of the precipitation areas). If the ensemble predictions were converted to a probabilistic forecast, four probability values would be possible:

Fig. 3 Schematic of ensemble features and their properties. Features are defined by, for example, forecast and observed precipitation exceeding a threshold, T (Figure courtesy of Chris Davis (NCAR))



0, 0.33, 0.67, and 1 (depending on the overlap of the four forecast areas). Another approach for examining these forecasts would be to compute the average, as depicted in the bottom diagram in Fig. 3. Each of these two approaches has advantages and disadvantages in terms of the kinds of information that can be gleaned from the comparison. In the case of conversion to a probabilistic forecast, the resulting features have little in common with the shape of the original forecast features. This result also characterizes the mean feature (note that this is a general flaw in the evaluation of ensemble means rather than looking at the whole ensemble). Therefore, neither of these approaches ideally represents the information content of the whole ensemble forecast. An alternative approach is to examine the distributions of attributes (e.g., areas, intensities, locations, displacements) to understand how the overall characteristics of the ensemble features differ from those from the observed field.

Gallus (2009) undertook one of the first applications of the feature-based approach to examine and compare ensembles of attributes of precipitation features. Specifically, he applied the CRA and MODE techniques to examine rain rate, volume, areal coverage, and displacement errors to examine spread-skill relationships and other attributes of ensemble performance. A more recent example based on the application of the MODE technique is shown in Fig. 4. In this example, the feature areas associated with high-resolution ensemble forecasts of precipitation are compared to the matched areas of the observed features. A great deal can be learned about the areas of the ensemble-based features compared to their matching observed feature. For example, many of the mean area values do not correspond well to the observed feature area, but, in many cases, one or more of the forecast areas are similar to the observed area. In addition, many of the mean forecast values are dominated by a single member. Similar analyses could be applied to other forecast attributes (e.g., intensity summaries within the features, orientation, shape, and location of the feature; cf. Micheas et al. 2007; Davis et al. 2006a, b, 2009; Lack et al. 2010). Such analyses are amenable to a variety of summaries that provide

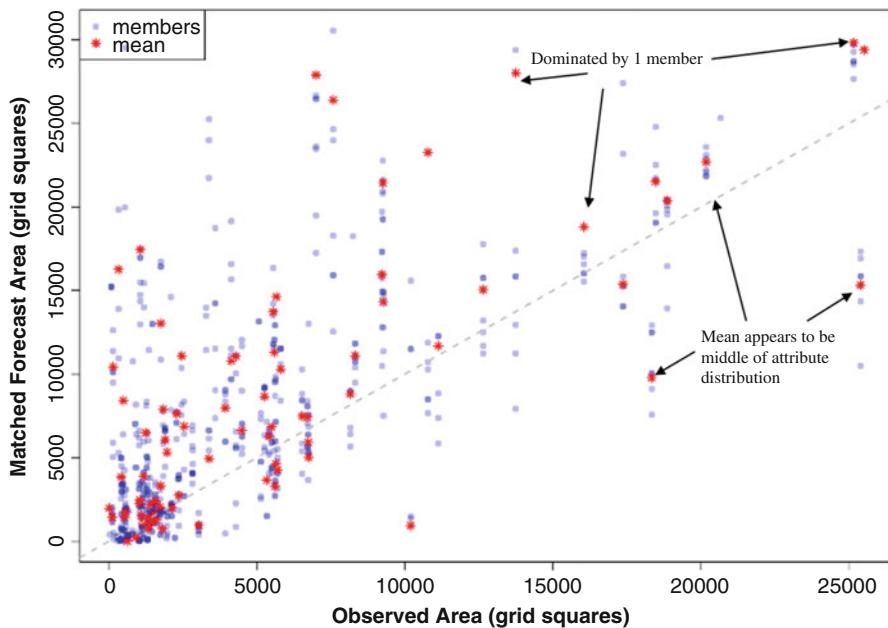


Fig. 4 Example of feature areas defined by application of the MODE technique to a high-resolution (9-km) ensemble forecast of precipitation. Feature areas predicted by individual ensemble members are shown as *blue dots*; ensemble mean areas are depicted by *star symbols* in *red*. *Grid squares* have approximate dimensions of 9×9 km (Figure courtesy of Tara Jensen (NCAR))

physically meaningful information to forecast developers, hydrologists, and other users, including more traditional verification measures, such as rank histograms and Brier scores.

The development of approaches to utilize spatial methods to evaluate ensemble predictions is still in its infancy. While efforts toward implementation of these new methods have been underway for a number of years, they have yet to realize full acceptance and application. Ongoing and future efforts will demonstrate the full benefits of application of spatial methods for evaluation of ensemble predictions.

7 Verification of Extreme Events

7.1 New Scores for Deterministic Forecasts

The ability of a forecast model to capture extreme meteorological events is a priority for hydrological concerns. However, the rarity of such events poses difficulties for their verification, both in terms of gaining meaningful information from summary scores, which are mostly aimed at informing about average performance, as well as

in handling the associated estimation uncertainty that is often very high because of the reduced number of occurrences.

Stephenson et al. (2008) summarize the behavior of several typical categorical verification scores as the base rate (i.e., the rate of occurrence of the observed event) deteriorates to zero. They proposed new summary statistics that better inform about a forecast model's ability to predict rare events, which Ferro and Stephenson (2011) subsequently modified to address issues that were since raised with the originally proposed summaries. These new measures are called the extreme dependency index (EDI) and the symmetric extreme dependency index (SEDI). If F is the false alarm rate (number of false alarms divided by the number of nonobserved events) and H is the hit rate (number of hits divided by the number of observed events), the EDI and SEDI are given by

$$\text{EDI} = \frac{\log F - \log H}{\log F + \log H}$$

$$\text{SEDI} = \frac{\log F - \log H - \log(1 - F) + \log(1 - H)}{\log F + \log H + \log(1 - F) + \log(1 - H)}$$

Both of these scores have values of 1 for perfect forecasts and 0 for forecasts with no skill. The motivation for these new statistics comes from extreme-value theory. Specifically, a relationship exists between these statistics and an extreme (or tail) dependence index (cf. Coles 2001, Sect. 8.4; Reiss and Thomas 2007).

7.2 Verifying Extreme Events for Ensembles

One way to focus verification for probabilistic forecasts for specific subgroups of observations such as extremes is to use the threshold-weighted continuous rank probability score (CRPS^t) (Gneiting and Ranjan 2011; Lerch 2012). This score can be conditioned on different discharge signatures (similar to weather regimes, see Lerch and Thorarinsdottir 2013). It is defined by

$$\text{CRPS}^t(f, y) = \int (F(z) - 1_A(y \leq z))^2 u(z) dz$$

$$\text{with } 1_A(y \leq z) := \begin{cases} 1 & \text{if } y \leq z \in A \\ 0 & \text{if } y \leq z \notin A \end{cases}$$

where F is the predictive cumulative distribution function corresponding to the probability density function f of the forecast and y is the observation. u is a nonnegative weight function of the forecast z , with $(z) = 1_A(z \geq r)$, which is equal to 1 for z values of the observation that are larger than or equal to a threshold

$r \epsilon \mathbb{R}$ (otherwise the function is 0). The score was used by Pappenberger et al. (2015) to evaluate river discharges for high and low flows in comparing different benchmarks and their fit for purpose for forecasting floods and droughts.

7.3 Extreme Forecast Index (EFI)

The extreme forecast index (EFI) was developed at ECMWF as a tool to provide forecasters with guidance on potential extreme weather events (Lalaurette 2003; Zsótér 2006). The EFI is a summary measure of the difference between the ensemble forecast distribution and the model climate distribution for a given weather element. It is specifically designed to highlight occasions when there is a significant shift in the ensemble forecast toward the climate extreme. The methodology of comparing a model forecast to an underlying climatology has been applied in both meteorology and hydrology (Lalaurette 2003; Thielen-del Pozo et al. 2009; Bartholmes et al. 2009; Cloke et al. 2010; Alfieri et al. 2011). This approach can be applied in areas where there are no or very few observations, presenting a major advantage to many other approaches. In addition, it can be seen as a post-processing or calibration method as it is computed entirely in the model space. One of the main assumptions associated with the EFI is that the behavior in terms of occurrence of anomalies in the model space is similar to that in the observation space.

The EFI is defined as:

$$\text{EFI} = \frac{2}{\pi} \int_0^1 \frac{p - F(p)}{\sqrt{p(1-p)}} dp$$

where $F(p)$ is a function denoting the proportion of ensemble members lying below the p th quantile of the climate record. The term $\sqrt{p(1-p)}$, which takes its minimum for $p = 0.5$ and its maximum at both ends of the probability range, is used to give more weight to the tails of the distribution. This weighting can also be interpreted as using the Anderson–Darling (Anderson and Darling 1952) test as a modification of the well-known Kolmogorov–Smirnov test. Given that $0 \leq F(p) \leq 1$, EFI values will lie in the same interval with values of 1.0 obtained when all the ensemble members are above (positive) or below (negative) the climate distribution. Dutra et al. (2013) present a semi-analytical technique to calculate the EFI.

The EFI is naturally extremely sensitive to the uncertainties in the underlying climatic distribution, which has been explored in detail by Zsótér et al. (2014). The climatic distribution needs to be derived from a large sample size in particular if certainty in the tails of the distribution is required. The EFI is routinely applied for variables such as temperature (Ghelli et al. 2010), precipitation (Haiden et al. 2014), snowfall (Tsonevsky and Richardson 2012), and wind (Petroliagis and Pinson 2012). It can be extended to many other areas, for example, to verify the amount of energy available for convection (CAPE, convective available potential energy).

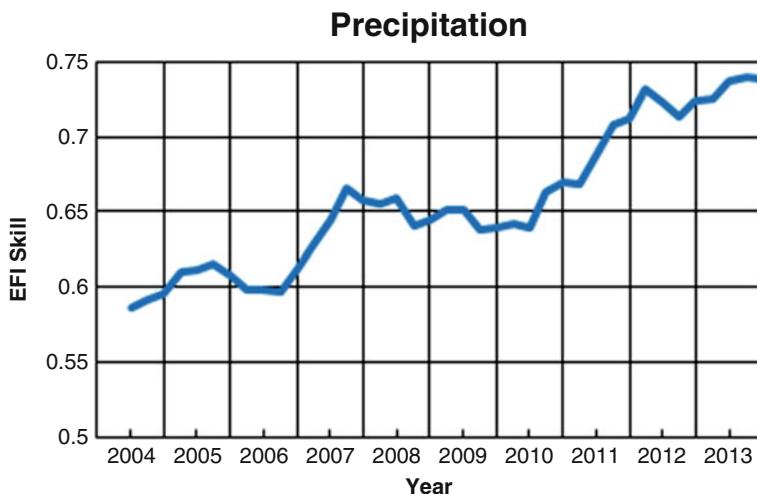


Fig. 5 Performance of the EFI for precipitation at forecast day 4 (24-h accumulation over period 72–96 h ahead); an extreme event is taken as an observation exceeding 95th percentile of station climate, and *curves* show a four-season running mean of relative operating characteristic (*ROC*) area skill scores (final point includes spring March–May 2014)

Meta-verification of the EFI has been performed using synoptic observations over Europe. An extreme event is judged to have occurred if the observation exceeds the 95th percentile of the observed climate for that station (calculated from a 15-year sample, 1993–2007). The ability of the EFI to detect extreme events is assessed using the relative operating characteristic (*ROC*) score. Results show that the EFI has substantial ability to provide early warnings of extreme events, confirming the subjective experience of forecasters, and that the performance has improved over recent years (Fig. 5).

Other indices for extreme weather forecasting can also be useful. For example, the Shift in Probability Space-index (SPS) provides information on how extreme the tail is. By using both these parameters, the level of extremity could be revealed. In order to measure the abnormality of such extreme situations, the Shift of Tails-index (SOT) should also be considered (Zsótér 2006; Tsonevsky and Richardson 2012).

8 Inference in the Presence of Correlation

Comparing one forecasting system or model against another is a frequent application of forecast verification. In the context of ensemble prediction, when verifying ensembles of forecasts composed of multiple models, the question often arises as to which multi-model ensemble member or members are better than others. For example, in aggregating forecasts over the members (e.g., when finding the ensemble mean), one might want to weight better performing models more heavily than

models that do not perform as well. In order to make such determinations, one might consider comparing summary statistics for the two models against the same observation. For example, one might consider testing the hypothesis that the difference in mean square error between two forecast members is zero (i.e., $H_0: E[MSE1] - E[MSE2] = 0$) against the two-sided alternative that the difference is different from zero (Diebold and Mariano 1995; Hering and Genton 2011; Gilleland 2013).

Correlation becomes a concern for conducting such a test because it will generally lead to overconfidence (lower variability) in the estimation of uncertainty concerning the difference. Correlation can come from dependence in time and space, as well as contemporaneous correlation between the two forecasts. Even if the two forecasts are completely independent of each other, contemporaneous correlation may still exist because of the use of the same set of observations in calculating the statistic of interest.

The economic literature is relatively rich in terms of investigating this particular issue. In particular, the seminal work by Diebold and Mariano (1995) introduced the idea of comparing the accuracy of time series forecasts with a test for the null hypothesis of equal forecast accuracy between two competing models while accounting for temporal correlation (henceforth, the DM test). Giacomini and White (2006) unified much of this theory into a generalized framework that includes the evaluation of point, interval, probability, and density forecasts. Hering and Genton (2011) further refined the DM test procedure, extended it to the spatial domain, and applied the test to a wind speed data set. Importantly, they also showed that their modifications result in a test that has proper size and power even in the face of contemporaneous correlation. Moreover, it works for any loss function, as well as correlation, and does not require any distributional assumption for the underlying data, only that the resulting test statistic is distributed as a standard normal.

Gilleland (2013) applied the test of Hering and Genton (2011) to quantitative precipitation forecasts and further incorporated methods to account for phase, or location, errors in addition to spatial correlation. The procedure provides a test for competing forecast models that accounts for spatial correlation and evaluates errors in location, which might be caused by timing errors or small-scale errors. The author found that accounting for such errors simultaneously with intensity errors is important in terms of the power of the tests. The additional deformation component in the testing procedure allows the test to better detect differences in skill when they actually exist.

The DM test for two time series forecasts, F_1 and F_2 , against the same set of observations, O , is carried out by first finding the loss, $g(F_1, O)$ and $g(F_2, O)$, at each time point (e.g., the loss might be MSE, absolute error, correlation, straight differences, etc.). The null hypothesis is that the expected difference between these two loss functions is zero, i.e.,

$$H_0 : E[g(F_1, O)] - E[g(F_2, O)] = 0.$$

The difference in loss functions is called the loss differential, say d . The asymptotic distribution of the sample *mean* loss differential as the sample size goes to infinity is

normal with (population) mean μ and variance $2\pi s_d(0)$, obtained in such a way as to account for temporal correlation. The large-sample standard normal test statistic for forecast accuracy is therefore given by:

$$S = \frac{\bar{d}}{\sqrt{2\pi s_d(0)/n}}$$

where n is the length of the time series and \bar{d} is the sample mean loss differential.

The key to incorporation of correlation information is in estimating the spectral density at frequency zero, which is accomplished through a weighted sum of the sample autocovariance for a k -step forecast, which could be either a single forecast or a time series of forecasts with a given lead time. The result is a direct consequence that the variance of a sum is the same as the sum of the covariances of the lagged (in time, space, or both) terms. That is,

$$2\pi \hat{s}_d(0) = \sum_{\tau=-(k-1)}^{k-1} \frac{1}{n} \sum_{t=|\tau|+1}^n (d_t - \bar{d})(d_{t-|\tau|} - \bar{d}).$$

A major drawback to the originally proposed empirical estimate above is that the sum over all lagged empirical terms is always zero, and clearly the variance is not generally zero. Hering and Genton (2011) suggest modifying the above estimate by replacing the empirical autocovariances with a fitted parametric covariance model, which, by definition, is nonnegative definite, so that the terms will not cancel each other out. For example, they use the exponential model given by:

$$2\pi \hat{s}_d(0) = \sigma^2 + 2\sigma^2 \sum_{\tau=1}^{n-1} e^{-3\tau/\theta}$$

where θ is the distance beyond which the correlation is less than about 0.05 (i.e., the practical range) and σ^2 is the lag-zero variance. The extension of the test to the spatial setting is analogous, replacing the autocovariance function with a parametric spatial covariance function of similar form.

9 Conclusions

Weather forecast verification is a relatively mature field, but one that continues to adapt to new demands brought on by changing technologies and model improvements, as well as one that continues to develop and adopt new methods, sometimes from other fields such as economics. Because hydrologic forecasts use many of the same variables that are used in meteorology, many of these methods can be useful in a hydrological context.

Even in the realm of meteorology, specific verification methods and analysis are highly dependent on specific user requirements. In this chapter, we have described some of the specific requirements for hydrology, such as the often complicated

meteorological conditions that affect discharge (e.g., direct and indirect connections with meteorology), and some of the newly developed methods and scores to better address verification challenges that are common to meteorological and hydrological applications. Challenges addressed in this chapter include handling double-penalty issues and small-scale error accumulation when verifying high-resolution forecasts, properly handling important rare events, and addressing spatial correlation when conducting statistical tests.

Additional challenges that are not addressed herein include multidimensional verification, which is particularly important for hydrologic applications, where it is likely that a combination of more than one variable can result in important-to-predict events, as well as the ability to verify across multiple temporal scales in order to seamlessly account for both short- and long-term forecasting (Ebert et al. 2013).

References

- C. Accadia, S. Mariani, M. Casaioli, A. Lavagnini, Sensitivity of precipitation forecast skill scores to bilinear interpolation and a simple nearest-neighbor average method on high-resolution verification grids. *Weather Forecast.* **18**, 918–932 (2003)
- A. AghaKouchak, N. Nasrollahi, J. Li, B. Imam, S. Sorooshian, Geometrical characterization of precipitation patterns. *J. Hydrometeorol.* **12**(2), 274–285 (2010). <https://doi.org/10.1175/2010JHM1298.1>
- G.D. Alexander, J.A. Weinman, J.L. Schols, The use of digital warping of microwave integrated vapor imagery to improve forecasts of marine extratropical cyclones. *Mon. Weather Rev.* **126**, 1469–1496 (1998)
- L. Alfieri, D. Velasco, J. Thielen, Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. *Adv. Geosci.* **29**, 69–75 (2011). <https://doi.org/10.5194/adgeo-29-69-2011>. www.adv-geosci.net/29/69/2011/
- L. Alfieri, P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, F. Pappenberger, GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.* **13**, 141–153 (2012)
- T. Anderson, D. Darling, Asymptotic theory of certain goodness of fit criteria based on stochastic processes. *Ann. Math. Stat.* **23**, 193–212 (1952)
- F. Atger, Verification of intense precipitation forecasts from single models and ensemble prediction systems. *Nonlinear Processes Geophys.* **8**, 401–417 (2001)
- J.C. Bartholmes, J. Thielen, M.H. Ramos, S. Gentilini, The European Flood Alert System EFAS – part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.* **13**, 141–153 (2009)
- W.M. Briggs, R.A. Levine, Wavelets and field forecast verification. *Mon. Weather Rev.* **125**, 1329–1341 (1997)
- B.G. Brown, E. Gilleland, E.E. Ebert, Forecasts of spatial fields, in *Forecast Verification*, ed. by I. Jolliffe, D.B. Stephenson. The Atrium, Southern Gate, Chichester, UK (2011), pp. 95–117
- R. Bullock, Development and implementation of MODE time domain object-based verification, in *24th Conference on Weather and Forecasting*, Seattle, 24–27 Jan 2011. American Meteorological Society
- N. Bussieres, W. Hogg, The objective analysis of daily rainfall by distance weighting schemes on a mesoscale grid. *Atmosphere–Ocean* **27**, 521–541 (1989)
- B. Casati, New developments of the intensity-scale technique within the spatial verification methods inter-comparison project. *Weather Forecast.* **25**(1), 113–143 (2010). <https://doi.org/10.1175/2009WAF222257.1>
- B. Casati, G. Ross, D.B. Stephenson, A new intensity-scale approach for the verification of spatial precipitation forecasts. *Meteorol. Appl.* **11**, 141–154 (2004)

- G.J. Ciach, W.F. Krajewski, On the estimation of radar rainfall error variance. *Adv. Water Resour.* **22**, 585–595 (1999)
- R. Cifelli, V. Chandrasekar, Dual-polarization radar rainfall estimation, in *Rainfall: State of the Science*, ed. by Y.M. Tistek, M. Gebremichael. American Geophysical Union, Washington, DC (2010), pp. 105–125. doi:10.1029/2010GM000930
- H.L. Cloke, C. Jeffers, F. Wetterhall, T. Byrne, J. Lowe, F. Pappenberger, Climate impacts on river flow: projections for the Medway catchment, UK, with UKCP09 and CATCHMOD. *Hydrol. Process.* (2010). <https://doi.org/10.1002/hyp.776>
- S.G. Coles, *An Introduction to Statistical Modeling of Extreme Values* (Springer, London, 2001)
- C. Davis, B. Brown, R.G. Bullock, Object-based verification of precipitation forecasts. Part I: methodology and application to mesoscale rain areas. *Mon. Weather Rev.* **134**(7), 1772–1784 (2006a). <https://doi.org/10.1175/MWR3145.1>
- C. Davis, B. Brown, R.G. Bullock, Object-based verification of precipitation forecasts. Part II: application to convective rain systems. *Mon. Weather Rev.* **134**(7), 1785–1795 (2006b). <https://doi.org/10.1175/MWR3146.1>
- C. Davis, B. Brown, R.G. Bullock, J. Halley Gotway, The method for object-based diagnostic evaluation (MODE) applied to numerical forecasts from the 2005 NSSL/SPC spring program. *Weather Forecast.* **24**(5), 1252–1267 (2009). <https://doi.org/10.1175/2009WAF2222241.1>
- F.X. Diebold, R.S. Mariano, Comparing predictive accuracy. *J. Bus. Econ. Stat.* **13**, 253–263 (1995)
- L. Duc, K. Saito, H. Seko, Spatial-temporal fractions verification for high-resolution ensemble forecasts. *Tellus A* **65**, 18171 (2013). <https://doi.org/10.3402/tellusa.v65i0.18171>
- J.D. Duda, X. Wang, F. Kong, M. Xue, Using varied microphysics to account for uncertainty in warm-season QPF in a convection-allowing ensemble. *Mon. Weather Rev.* **142**, 2198–2219 (2014)
- E. Dutra, M. Diamantakis, I. Tsonevsky, E. Zsoter, F. Wetterhall, T. Stockdale, D. Richardson, F. Pappenberger, The extreme forecast index at the seasonal scale. *Atmos. Sci. Lett.* **14**(4), 256–262 (2013)
- E.E. Ebert, Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteorol. Appl.* **15**, 51–64 (2008). <https://doi.org/10.1002/met.25>. Available at <http://www.ecmwf.int/newsevents/meetings/workshops/2007/jwgv/METspecialissueemail.pdf>
- E.E. Ebert, J.L. McBride, Verification of precipitation in weather systems: determination of systematic errors. *J. Hydrol.* **239**, 179–202 (2000)
- E.E. Ebert, U. Damrath, W. Wergen, M.E. Baldwin, The WGNE assessment of short-term quantitative precipitation forecasts. *Bull. Am. Meteorol. Soc.* **84**, 481–492 (2003)
- E. Ebert, L. Wilson, A. Weigel, M. Mittermaier, P. Nurmi, P. Gill, M. Göber, S. Joslyn, B. Brown, T. Fowler, A. Watkins, Progress and challenges in forecast verification. *Meteorol. Appl.* **20**, 130–139 (2013). <https://doi.org/10.1002/met.1392>
- EUMETNET, ODYSSEY, the OPERA Data Centre (2014), <http://www.eumetnet.eu/odyssey-opera-data-centre>. Viewed 10 Apr 2014
- C.A.T. Ferro, D.B. Stephenson, Extremal dependence: improved verification measures for deterministic forecasts of rare binary events. *Weather Forecast.* **26**, 699–713 (2011). <https://doi.org/10.1175/WAF-D-10-05030.1>
- V. Fortin, M. Abaza, F. Anctil, R. Turcotte, Why should ensemble spread match the RMSE of the ensemble mean? *J. Hydrometeorol.* **15**, 1708–1713 (2016)
- W.A. Gallus Jr., Application of object-based verification techniques to ensemble precipitation forecasts. *Weather Forecast.* **25**, 144–158 (2009)
- A. Ghelli, A. Garcia-Mendez, F. Prates, M. Dahoui, Extreme weather events in summer 2010: how did the ECMWF forecasting systems perform? *ECMWF Newsl.* **125**, 7–11 (2010)
- R. Giacomini, H. White, Tests of conditional predictive ability. *Econometrica* **74**, 1545–1578 (2006)
- E. Gilleland, Spatial forecast verification: Baddeley's delta metric applied to the ICP test cases. *Weather Forecast.* **26**, 409–415 (2011). <https://doi.org/10.1175/WAF-D-10-05061.1>

- E. Gilleland, Testing competing precipitation forecasts accurately and efficiently: the spatial prediction comparison test. *Mon. Weather Rev.* **141**(1), 340–355 (2013). <https://doi.org/10.1175/MWR-D-12-00155.1>
- E. Gilleland, T.C.M. Lee, J. Halley Gotway, R.G. Bullock, B.G. Brown, Computationally efficient spatial forecast verification using Baddeley's Delta image metric. *Mon. Weather Rev.* **136**(5), 1747–1757 (2008). <https://doi.org/10.1175/2007MWR2274.1>
- E. Gilleland, D. Ahijevych, B.G. Brown, B. Casati, E.E. Ebert, Intercomparison of spatial forecast verification methods. *Weather Forecast.* **24**(5), 1416–1430 (2009). <https://doi.org/10.1175/2009WAF2222269.1>
- E. Gilleland, Coauthors, Spatial forecast verification: Image warping. NCAR Technical Note NCAR/TN-482+STR (2010). doi:10.5065/D62805JJ
- E. Gilleland, D.A. Ahijevych, B.G. Brown, E.E. Ebert, Verifying forecasts spatially. *Bull. Am. Meteorol. Soc.* **91**(10), 1365–1373 (2010a). <https://doi.org/10.1175/2010BAMS2819.1>
- E. Gilleland, L. Chen, M. DePersio, G. Do, K. Eilertson, Y. Jin, E.K. Lang, F. Lindgren, J. Lindström, R.L. Smith, C. Xia, Spatial forecast verification: image warping. *NCAR Tech. Note* (2010b), NCAR/TN-482 + STR doi:10.5065/D62805JJ
- E. Gilleland, J. Lindström, F. Lindgren, Analyzing the image warp forecast verification method on precipitation fields from the ICP. *Weather Forecast.* **25**(4), 1249–1262 (2010c). <https://doi.org/10.1175/2010WAF2222365.1>
- T. Gneiting, R. Ranjan, Comparing density forecasts using threshold- and quantile-weighted scoring rules. *J. Bus. Econ. Stat.* **29**(3), 411–422 (2011)
- M. Goebel, E. Zsoter, D.S. Richardson, Could a perfect model ever satisfy the forecaster? On grid box mean versus point verification. *Meteorol. Appl.* **15**, 359–365 (2008)
- T. Haiden, M. Rodwell, D. Richardson, A. Okagaki, T. Robinson, T. Hewson, Intercomparison of global model precipitation forecast skill in 2010/11 using the SEEPS score. *Mon. Weather Rev.* **140**, 2720–2733 (2012)
- T. Haiden, L. Magnusson, I. Tsonevsky, F. Wetterhall, L. Alfieri, F. Pappenberger, P. de Rosnay, J. Muñoz-Sabater, G. Balsamo, C. Albergel, R. Forbes, T. Hewson, S. Malardel, D. Richardson, ECMWF forecast performance during the June 2013 flood in Central Europe. Technical Memorandum No. 723, European Centre for Medium Range Weather Forecasts, Reading, England, vol 723 (2014)
- T.M. Hamill, J. Juras, Measuring forecast skill: is it real skill or is it the varying climatology? *Q. J. Roy. Meteorol. Soc.* **132**, 2905–2923 (2006)
- A.S. Hering, M.G. Genton, Comparing spatial predictions. *Technometrics* **53**, 414–425 (2011)
- R.N. Hoffman, Z. Liu, J.-F. Louis, C. Grassotti, Distortion representation of forecast errors. *Mon. Weather Rev.* **123**, 1758–2770 (1995)
- G.J. Huffman, R.F. Adler, D.T. Bolvin, G. Gu, E.J. Nelkin, K.P. Bowman, Y. Hong, E.F. Stocker, D.B. Woff, The TRMM multi-satellite precipitation analysis: quasi-global, multi-year, combined-sensor precipitation estimates at fine scale. *J. Hydrometeorol.* **8**, 38–55 (2007)
- G.J. Huffman, D.T. Bolvin, D. Braithwaite, K. Hsu, R. Joyce, P. Xie, S.H. Yoo, NASA global precipitation measurement (GPM) integrated multi-satellite retrievals for GPM (IMERG), in *Algorithm Theoretical Basis Document, Version 4.1*. NASA, Greenbelt, MD (2013)
- IPWG, International Precipitation Working Group: Data and Products (2016), <http://www.isac.cnr.it/~ipwg/data.html>. Viewed 18 June 2016
- J. Jenkner, C. Frei, C. Schwierz, Quantile-based short-range QPF evaluation over Switzerland. *Meteorol. Z.* **17**, 827–848 (2008)
- C. Keil, G.C. Craig, A displacement-based error measure applied in a regional ensemble forecasting system. *Mon. Weather Rev.* **135**, 3248–3259 (2007). <https://doi.org/10.1175/MWR3457.1>
- C. Keil, G.C. Craig, A displacement and amplitude score employing an optical flow technique. *Weather Forecast.* **24**(5), 1297–1308 (2009). doi:10.1175/2009WAF2222247.1
- C. Kidd, V. Levizzani, Status of satellite precipitation retrievals. *Hydrol. Earth Syst. Sci.* **15**, 1109–1116 (2011)

- S.A. Lack, G.L. Limpert, N.I. Fox, An object-oriented multiscale verification scheme. *Weather Forecast.* **25**(1), 79–92 (2010). <https://doi.org/10.1175/2009WAF2222245.1>
- F. Lalaurette, Early detection of abnormal weather conditions using a probabilistic extreme forecast index. *Q. J. Roy. Meteorol. Soc.* **129**(594), 3037–3057 (2003)
- S.M. Lazarus, C.M. Ciliberti, J.D. Horel, K.A. Brewster, Near-real-time applications of a mesoscale analysis system to complex terrain. *Weather Forecast.* **17**, 971–1000 (2002)
- S. Lerch, *Verification of Probabilistic Forecasts for Rare and Extreme Events* (Ruprecht-Karls-Universitaet Heidelberg, Heidelberg, 2012)
- S. Lerch, T.L. Thorarinsson, Comparison of nonhomogeneous regression models for probabilistic wind speed forecasting. *Tellus A* **65**, 21206 (2013)
- C. Marsigli, A. Montani, T. Pacagnella, A spatial verification method applied to the evaluation of high-resolution ensemble forecasts. *Meteorol. Appl.* **15**, 125–143 (2008)
- C.F. Mass, D. Ovens, K. Westrick, B.A. Colle, Does increasing horizontal resolution produce more skillful forecasts? *Bull. Am. Meteorol. Soc.* **83**, 407–430 (2002)
- J.L. McBride, E.E. Ebert, Verification of quantitative precipitation forecasts from operational numerical weather prediction models over Australia. *Weather Forecast.* **15**, 103–121 (2000)
- R. Merz, G. Blöschl, A process typology of regional floods, *Water Resour. Res.* **39**, 1340 (2003). doi:10.1029/2002WR001952, 12
- A. Micheas, N.I. Fox, S.A. Lack, C.K. Wikle, Cell identification and verification of QPF ensembles using shape analysis techniques. *J. Hydrol.* **344**, 105–116 (2007)
- J.R. Minder, P.W. Mote, J.D. Lundquist, Surface temperature lapse rates over complex terrain: lessons from the Cascade Mountains. *J. Geophys. Res. Atmos.* **115**(D14) (2010). <https://doi.org/10.1029/2009JD013493>
- M. Mittermaier, N. Roberts, Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Weather Forecast.* **25**, 343–354 (2010)
- M. Mittermaier, N. Roberts, S.A. Thompson, A long-term assessment of precipitation forecast skill using the Fractions Skill Score. *Meteorol. Appl.* **20**, 176–186 (2013)
- F. Molteni, R. Buizza, T.N. Palmer, T. Petroliagis, The ECMWF ensemble prediction system: methodology and validation. *Q. J. Roy. Meteorol. Soc.* **122**, 73–119 (1996)
- J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models. Part I: a discussion of principles. *J. Hydrol.* **10**, 282–290 (1970)
- R. North, M. Trueman, M. Mittermaier, M.J. Rodwell, An assessment of the SEEPS and SEDI metrics for the verification of 6 h forecast precipitation accumulations. *Meteorol. Appl.* **20**, 164–175 (2013)
- NWS, Advanced Hydrologic Prediction Service (2014), <http://water.weather.gov/precip/about.php>. Viewed 10 Apr 2014
- F. Pappenberger, M.H. Ramos, H.L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, P. Salamon, How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.* **522**, 697–713 (2015)
- T. Petroliagis, P. Pinson, Early indication of extreme winds utilising the Extreme Forecast Index, *ECMWF Newsletter* No. 132 – Summer 2012, 2012
- R.-D. Reiss, M. Thomas, *Statistical Analysis of Extreme Values: With Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd edn. (Birkhäuser, Basel, 2007), 530pp
- N.M. Roberts, H.W. Lean, Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Mon. Wea. Rev.* **136**, 78–96 (2008)
- M.J. Rodwell, D.S. Richardson, T.D. Hewson, T. Haiden, A new equitable score suitable for verifying precipitation in numerical weather prediction. *Q. J. Roy. Meteorol. Soc.* **136**, 1344–1363 (2010)
- Ø. Saetra, H. Hersbach, J.-R. Bidlot, D.S. Richardson, Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Weather Rev.* **132**, 1487–1501 (2004)
- B.R.J. Schwedler, M.E. Baldwin, Diagnosing the sensitivity of binary image measures to bias, location, and event frequency within a forecast verification framework. *Weather Forecast.* **26**, 1032–1044 (2011). <https://doi.org/10.1175/WAF-D-11-00032.1>

- D.J. Seo, J.P. Breidenbach, Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *J. Hydrometeorol.* **3**, 93–111 (2002)
- R. Steinacker, C. Häberli, W. Pöttschacher, A transparent method for the analysis and quality evaluation of irregularly distributed and noisy observational data. *Mon. Weather Rev.* **128**, 2303–2316 (2000)
- D.B. Stephenson, B. Casati, C.A.T. Ferro, C.A. Wilson, The extreme dependency score: a non-vanishing measure for forecasts of rare events. *Meteorol. Appl.* **15**, 41–50 (2008). <https://doi.org/10.1002/met.53>
- J. Thielen-del Pozo, J. Bartholmes, M.-H. Ramos, A. de Roo, The European Flood Alert System. Part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125–140 (2009)
- I. Tsonevsky, D.S. Richardson, D.S., Application of the new EFI products to a case of early snowfall in Central Europe. *ECMWF Newsletter* 133, Autumn 2012, p. 4, 2012
- H. Wernli, M. Paulat, M. Hagen, C. Frei, SAL – a novel quality measure for the verification of quantitative precipitation forecasts. *Mon. Weather Rev.* **136**, 4470–4487 (2008)
- T. Weusthoff, F. Ament, M. Arpagaus, M.W. Rotach, Assessing the benefits of convection-permitting models by neighborhood verification: examples from MAP D-PHASE. *Mon. Weather Rev.* **138**, 3418–3433 (2010)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*, 3rd edn. (Academic, Amsterdam, 2011), 704 pp
- WMO, Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models – Revision 2, WMO/TD-No.1485 WWRP 2009-1 (2008)
- WMO, *Manual on the Global Data-Processing and Forecasting System (GDPFS)*. WMO-no.485, World Meteorological Organization, Geneva, vol 485 (2014)
- J. Wolff, M. Harrold, T.A. Fowler, J. Halley Gotway, L. Nance, B.G. Brown, Beyond the basics: evaluating model-based precipitation forecasts using traditional, spatial, and object-based methods. *Weather Forecast.* **29**, 1451–1472 (2014)
- P. Zacharov, D. Rezacova, Using the fractions skill score to assess the relationship between an ensemble QPF spread and skill. *Atmos. Res.* **94**, 684–693 (2009)
- M. Zimmer, H. Wernli, Verification of quantitative precipitation forecasts on short time-scales: a fuzzy approach to handle timing errors with SAL. *Meteorol. Z.* **20**, 95–105 (2011)
- E. Zsótér, Recent developments in extreme weather forecasting. *ECMWF Newsl.* **107**(107), 8–17 (2006)
- E. Zsótér, F. Pappenberger, D. Richardson, Sensitivity of model climate to sampling configurations and the impact on the Extreme Forecast Index. *Meteorol. Appl.* (2014). <https://doi.org/10.1002/met.1447>



Verification of Short-Range Hydrological Forecasts

Katharina Liechti and Massimiliano Zappa

Contents

1	Introduction	954
2	Verification of Ensemble Streamflow Nowcasts	955
2.1	Varying Meteorological Input: Taking into Account the Input Uncertainty	956
2.2	Varying Parameterization: Taking into Account the Parameter Uncertainty	957
3	Operational Hydrological Ensemble Prediction System (HEPS)	960
3.1	Correlation	961
3.2	Frequency Bias	961
3.3	Value Score	964
3.4	Brier Skill Score	964
3.5	Rank Histograms	968
3.6	Conditional Bias: Reliability Plots	969
3.7	Performance Overview	971
4	Flood Peak and Flood Timing Verification	971
5	Conclusions	973
	References	974

Abstract

For the mitigation of floods and flashfloods, operational nowcast and forecast systems are crucial. This chapter provides practical illustrations of the verification of hydrological ensemble prediction systems with a temporal horizon of up to 5 days.

Section 2 shows the application of two ensemble approaches for discharge nowcasts. The results show that both ensemble approaches have added value compared to deterministic nowcasts.

Section 3 presents the evaluation of an operational flood forecasting system. The system is run with the two deterministic COSMO-2 and COSMO-7 weather

K. Liechti (✉) · M. Zappa

Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

e-mail: liechti@wsl.ch; zappa@wsl.ch

forecasts and with the probabilistic COSMO-LEPS weather forecast. The evaluation with several skill scores suggests that decisions that need to be taken with a lead time of 1 day and more should be based on the ensemble forecast.

Ensemble forecasts can be difficult to interpret. Section 4 provides a helpful tool for the estimation of flood peak timing and magnitude based on probabilistic forecasts.

Keywords

Real-time experiment · Hydrological forecast · Short range · Skill score · Forecast verification

1 Introduction

In contrast to deterministic predictions, probabilistic ensemble predictions contain information on the uncertainty of the prediction. The uncertainties that are represented by the ensemble can stem from the model parameterization, the model structure, and/or the input data, depending on the setup of the ensemble prediction system. To assess if this information on uncertainty has an added value for the forecasters and users, the ensemble predictions have to be compared to the deterministic predictions. As a measure for direct comparison of deterministic and ensemble predictions, the Brier skill score (BSS) is used here. The BSS is the mean squared error of the prediction for a predefined threshold normalized by a reference prediction like climatology (Wilks 2006). Other metrics like the correlation coefficient, the frequency bias, and the value score are deterministic measures which need the prior reduction of the ensemble to its median or mean, for example. For more detailed evaluation of probabilistic predictions, three additional metrics are used here. The rank histogram shows if the ensemble has an appropriate degree of dispersion. The reliability diagram (conditional bias) is a powerful visualization that sets the observed relative frequency of an event in relation to the forecast probability of that event. Thus the reliability diagram shows if there exist conditional biases in the prediction. The value score gives information on the economic value of a prediction system for a user with a specific cost/loss ratio. The detailed descriptions of the performance measures used in this chapter can be found in Wilks (2006).

For all examples in this chapter, the hydrological model PREVAH was used.

PREVAH is semi-distributed, working with hydrological response units. For the presented examples, the model is run at hourly time steps and on a 500 m grid. The meteorological variables needed to drive the model are air temperature, wind speed, water vapor pressure, global radiation, sunshine duration, and precipitation. Detailed descriptions of the model can be found in Vivenioli et al. (2009a, b, and c).

This chapter is divided into three main parts. Sections 2 and 3 will assess the added value of the ensemble approach versus the deterministic approach. The focus of Sect. 2 is on nowcasts. The objective here is to assess the added value for peak flow predictions when including the uncertainties stemming from the model

parameterization and from the meteorological input. Thus, two different kinds of ensemble nowcasts are compared to deterministic nowcasts.

Section 3 presents an operational hydrological forecast system with a lead time of up to 5 days. The system includes two deterministic model chains as well as a probabilistic model chain, driven by the COSMO-LEPS ensemble weather forecast, which consists of 16 members (Marsigli et al. 2005). The performance of the three different forecast chains are compared with different skill scores.

Section 4 presents a tool that helps to interpret ensemble forecasts. The tool is called Peak-Box and helps to interpret ensemble forecasts (spaghetti plots) for predicted flood situations, in terms of both peak flood magnitude and timing.

In the presented examples, the performance of each forecast product is computed by comparing it to the discharge observation. However, if no discharge observation is available, the simulated flow, driven by precipitation measurements or estimates, can be used as a reference instead of the observation. If effects from systematic bias of the hydrological model are to be excluded from the analysis, it is recommended to use the simulated flow as reference (Jaun et al. 2008).

2 Verification of Ensemble Streamflow Nowcasts

The objective of this chapter is to assess the added value for peak flow predictions when including the uncertainties stemming from the meteorological input and from the model parameterization. These uncertainties are included using ensembles. When working with one specific hydrological model, like in the presented examples (Sects. 2.1, 2.2), there are basically two approaches to produce ensembles. The meteorological input to the model and the model parameterization can be varied (Liechti et al. 2013b).

The focus of this chapter lies on very short-term forecasts, which are generally called nowcasts (Mandapaka et al. 2012). In this chapter, streamflow nowcasts are forced by operationally available precipitation measurements or estimates. Therefore their maximum lead time equals to the response time of the considered catchment.

It is self-evident to question the purpose of ensemble streamflow nowcasts with a lead time that does not exceed the response time of the catchment. At first running an ensemble streamflow nowcast seems to be overkill. For smaller catchments with short response times, streamflow nowcasts do not help much for flood preparedness. However, continuous ensemble streamflow nowcasts can be a valuable source for ensembles of initial conditions of the hydrological model for the initialization of NWP-driven forecasts (Liechti et al. 2013a; Zappa et al. 2011). But for catchments with a response time of a few hours and more, even streamflow nowcasts can provide information that may help to take decisions on mitigation measures. Streamflow nowcasts are thus particularly interesting in flood situations.

The following experiment is set up for the Verzasca catchment in southern Switzerland. The Verzasca is a mountainous catchment with steep slopes and shallow soils. The valley is only sparsely populated and affected by human impact; thus, damage potential is low. Frequent convective and orographic precipitation

events often lead to flash floods (Panziera and Germann 2010). The response time of the basin in terms of peak discharge is in case of very local thunderstorms between 2 and 3 h. At the gauging station, the catchment area sums up to 186 km². Shortly after the gauge, the river flows into a retention lake which is used for hydropower production. The available meteorological information from ground-based stations by MeteoSwiss and other providers is rather dense with an estimated density of one rain gauge per 100 km² in southern Switzerland (e.g., Andres et al. (2016)).

The data used for this analysis is taken from continuous hourly streamflow nowcasts generated between May 2007 and November 2010. To avoid the snow and snowmelt season, only the months May to November were considered. This is also the season where the relevant thunderstorms occur.

All streamflow nowcasts are compared to the observed discharge and their performance compared to each other. The focus of this analysis is on peak discharge; thus, the actual data used for the performance analysis are the daily maxima of the observation, of each ensemble member, and of the deterministic streamflow nowcasts. This means that the analysis includes 856 days.

2.1 Varying Meteorological Input: Taking into Account the Input Uncertainty

In this first part of the experiment, a deterministic streamflow nowcast driven by radar Quantitative Precipitation Estimate (QPE) is compared to a probabilistic streamflow nowcast driven by a radar ensemble. Both products are used at a temporal resolution of 1 h. The deterministic radar QPE (referred to as RAD) has a spatial resolution of 1 km. The radar ensemble introduced by Germann et al. (2009) consists of 25 members and has a spatial resolution of 2 km, in the following it is referred to as REAL (radar ensemble generator designed for usage in the Alps using LU decomposition). The members of the REAL ensemble result from the sum of the current weather radar image and a stochastic perturbation field. This methodology allows to account for the residual space-time uncertainties of the atmosphere (Germann et al. 2009).

The performance of the streamflow nowcasts in predicting the exceedance of a predefined threshold can be calculated using various skill scores. Here, the Brier skill score (BSS) is used (Wilks 2006). The BSS is a widely used summary skill score that allows direct comparison of deterministic and ensemble predictions and For every day, the maximum discharge value of the observation, of each ensemble member, and of the deterministic nowcast is taken. Then it is checked for each of these 27 values if it exceeds a predefined threshold value or not. For the observation and the deterministic forecast, the answer will be yes or no. For the radar ensemble nowcast, the answer will be the percentage of the ensemble members exceeding the threshold, i.e., the nowcast probability to exceed the threshold. The thresholds tested here correspond to the 60%, 70%, 80%, 90%, and 95% quantile of the observed daily maximal discharge during the months May to November from the years 1989 to 2010.

The performance analysis with the Brier skill score shows that the peak discharge nowcasts benefit from using the radar ensemble REAL (Fig. 1). The BSS values for

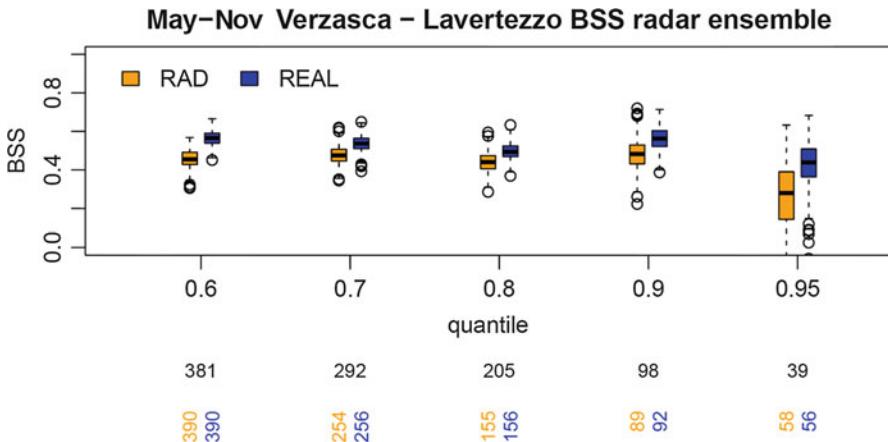


Fig. 1 Brier skill score for the deterministic streamflow nowcast driven by the radar QPE (RAD) and for the ensemble streamflow nowcast driven by the radar ensemble (REAL). The uncertainty estimation of the BSS values results from a bootstrap with 500 random samples. Horizontal numbers below the graph show the number of days in which an exceedance of the threshold quantile was observed. Vertical numbers show for each nowcast type the number of days that was simulated to be above the threshold quantile (for the ensemble nowcasts, the number of ensemble members exceeding the threshold was divided by the ensemble size)

the ensemble are higher than those for deterministic streamflow nowcasts using the normal radar QPE as input. This superiority gets more pronounced for higher threshold quantiles.

Values exceeding high quantiles are per definition scarce; therefore, it is recommended to apply a bootstrap method to get confidence bounds and thus a feeling for the accuracy of the estimated skill score (Efron 1979; Wilks 2006). In the presented examples, 500 random samples of nowcast-observation pairs of daily maxima were drawn with replacement from the 856 days included in the study period. For each of these 500 samples, the BSS was calculated. With the resulting 500 BSS values, the confidence bounds were calculated. Here, the number of bootstrap samples is set to 500 because more samples did not result in a significant difference of the score values. However, the number of bootstrap samples should be chosen for each study separately, generally the more the better. In Figs. 1 and 2, it can be seen that the confidence intervals for the BSS values are larger for the highest quantile, indicating that BSS values estimated from fewer data points are more uncertain.

2.2 Varying Parameterization: Taking into Account the Parameter Uncertainty

Usually hydrological models are calibrated and then used with the parameter set that performed best during the verification period. However, as no model structure perfectly represents the processes to be modeled, it can be argued that no such best

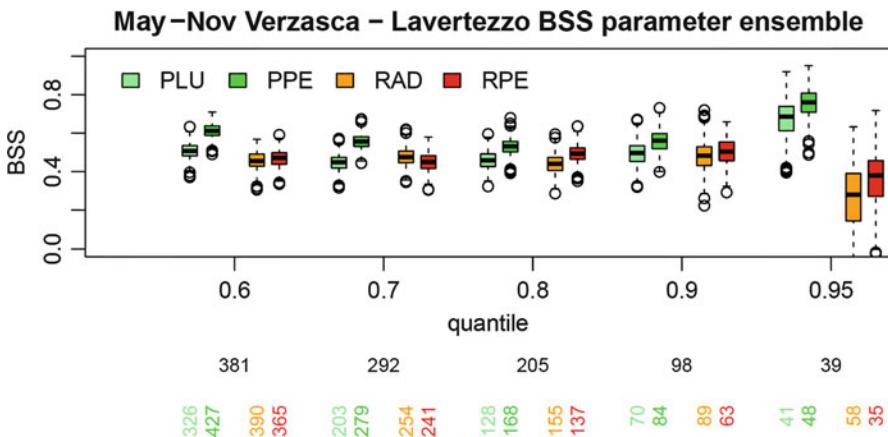


Fig. 2 Brier skill score for the deterministic nowcast driven by the interpolated rain gauge data (PLU) and driven by radar QPE (RAD) and for the ensemble nowcasts resulting from using 26 different parameter sets in combination with the interpolated rain gauge data (PPE) and in combination with the radar QPE data (RPE). The uncertainty estimation of the BSS values results from a bootstrap with 500 random samples. Horizontal numbers below the graph show the number of days in which an exceedance of the threshold quantile was observed. Vertical numbers show for each nowcast type the number of days that was simulated to be above the threshold quantile (for the ensemble nowcasts, the number of ensemble members exceeding the threshold was divided by the ensemble size)

parameter set exists. This fact can be met by the concept of equifinality (Beven and Freer 2001). Several equally likely parameter sets can be searched with a Monte Carlo (MC) experiment (Kuczera and Parent 1998; Vrugt et al. 2008). This approach was followed in this second part of the example. The seven most important parameters for conditioning the precipitation input and the surface runoff generation were allowed to change randomly during an MC experiment. Then all the MC runs were ranked according to an objective function consisting of the Nash-Sutcliffe efficiency (Nash and Sutcliffe 1970) and the sum of weighted absolute errors (Lamb 1999; Viviroli et al. 2009a). Finally, 26 parameter sets were chosen randomly around the 95% ranking. A more detailed description of this approach can be found in Zappa et al. (2011). These 26 parameter sets were then used to run the hydrological model with deterministic precipitation data from radar QPE and from interpolated rain gauge data, respectively. This resulted in two ensemble streamflow nowcasts that both account for the parameterization uncertainty of the hydrological model. The ensemble driven by radar QPE data is labeled RPE, and the ensemble driven by interpolated rain gauge data is labeled PPE hereafter. The performance of these parameter ensemble nowcasts is then compared to the performance of their deterministic counterparts. These deterministic streamflow nowcasts are the result of running the hydrological model with the default parameterization usually used for the Verzasca catchment in combination with the radar QPE (RAD) and the interpolated rain gauge data (PLU).

The performance analysis based on the Brier skill score shows also here that ensemble nowcasts generally outperform the deterministic nowcasts (Fig. 2).

PPE outperforms PLU, and interestingly it can be seen that the nowcasts driven by interpolated rain gauge data outperform the nowcasts driven by radar data. The two main reasons for this are that the hydrological model was calibrated using interpolated rain gauge data and that there is a rain gauge within the catchment. If the hydrological model would have calibrated with radar data or if there were no rain gauges within the catchment, this result might have been different (Liechti et al. 2013b).

RPE is generally better than RAD but does not reach the performance of REAL (Fig. 1). In mountainous regions there are many uncertainties involved in radar QPEs, and thus it is not surprising that REAL, which attempts to account for these uncertainties, performs better than RAD and RPE.

If the rank histograms of the two radar-based ensemble streamflow nowcasts REAL and RPE are compared (Fig. 3), it can be seen that both are underdispersed. However this is less pronounced for REAL, especially on the high quantile. Underdispersed forecasts are typical in such basins with quick response to precipitation. Postprocessing would be needed to be applied to overcome this (e.g., Bogner et al. (2016), (Hemri et al. 2013)).

However, radar ensembles like REAL are not widely available yet, and thus it is still difficult to account for the impact of the radar QPE uncertainty on the

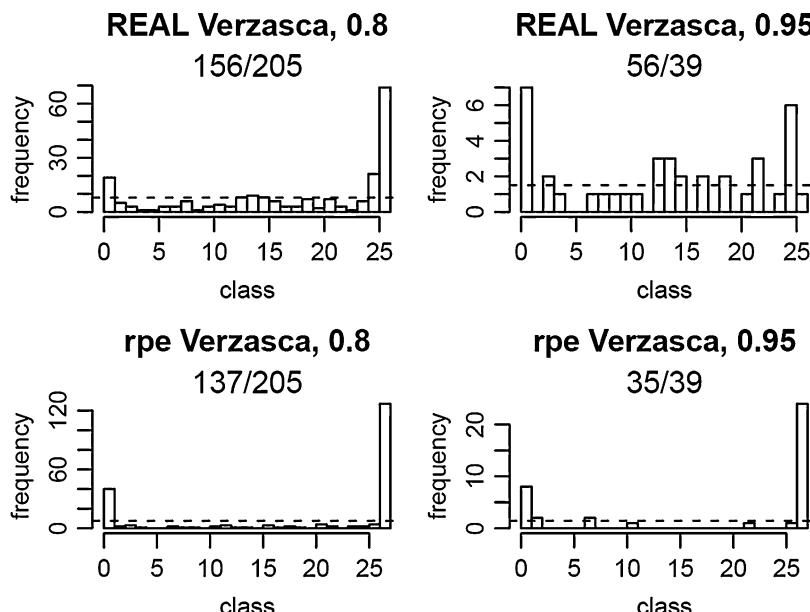


Fig. 3 Rank histograms for nowcasts driven by REAL and by the radar QPE in combination with the 26 different parameter sets (RPE). The number of days the ensemble and the observation exceeded the threshold quantile (0.8 and 0.9) are given in the subtitles

hydrological prediction. However, producing a hydrological ensemble using several parameter sets in combination with the deterministic radar QPE seems to be a reasonable alternative to account at least for some of the hydrologic uncertainty sources, namely, the hydrological parameter uncertainty.

3 Operational Hydrological Ensemble Prediction System (HEPS)

The verification of operational HEPS is shown here using the hydrological forecasting system for the Sihl river in Zurich, Switzerland (Addor et al. 2011; Badoux et al. 2010; Zappa et al. 2010).

This system was set up to help protect the city of Zurich from flood damage caused by the Sihl river. The flood damage potential of the city is high as a lot of infrastructure was built on the alluvial fan of the Sihl river during the last century. It is estimated to about five billion Euro. This information may seem irrelevant first, but it is relevant if the economic value of the system is assessed. The pre-alpine Sihl catchment has an area of 336 km², thereof 46% first flows into a reservoir lake that is used for hydropower production but can also act as a retention basin. It is possible to draw down the reservoir lake as a preventive measure if a major event is expected (Zappa et al. 2015). However, it is to be considered that this action needs a lead time of 1 to 3 days and that if the predicted event does not occur, the authorities have to pay the spilled water to the operator of the hydropower plant.

For this operational forecast system, the hydrological model PREVAH is driven by three different numerical weather prediction (NWP) models. These are the two deterministic models COSMO-2 and COSMO-7 and the probabilistic model COSMO-LEPS, which has 16 members (Fundel et al. 2010; Marsigli et al. 2005). The models differ in their spatial resolution, lead time, and update cycle (Table 1).

The system is operational since 2007. A major change in the system in late 2009 was the increase in spatial resolution of the meteorological models. The critical period for the Sihl river lasts from March to October, including the snowmelt season, the thunderstorms in summer, as well as long-lasting precipitation events in summer and autumn. During these months the emergency response team of the authorities is on duty. So, even though the forecast system runs continuously over the whole year, for the analysis presented here, only the relevant months (March to October) from each year (2010 to 2014) are considered. A period of 2 weeks from summer 2014 that includes an event resulting from an artificial drawdown of the reservoir lake was

Table 1 Properties of numerical weather prediction (NWP) systems used to drive the PREVAH model. Initialization times refer to the runs of the NWP used in the hydrological forecast chain

Model	Spatial resolution	Initialization	Lead time	Member
COSMO-2	2.2 km	00, 03, 06, ... UTC	24 h	1
COSMO-7	6.6 km	00, 06, 12 UTC	72 h	1
COSMO-LEPS	7 km	12 UTC	132 h	16

excluded from the sample. This is justified as the forecast otherwise gets wrongly penalized. The observation data is available in hourly time steps for the whole period.

The system is evaluated here with several scores, from simple to more complex. It is one possible set of scores and can be extended or changed if desired (Brown et al. 2010). The correlation and the frequency bias give a broad first impression about the behavior of the system in a deterministic way. Such an approach has been particularly useful for slowly introducing end users to more complex skill scores. The Brier skill score is used to directly compare the performance of deterministic and ensemble forecasts to predict the exceedance of a predefined threshold. The value score gives information about the economical usefulness of the forecast system for users with a specific cost/loss ratio. Then the rank histogram and the conditional bias (reliability plot) are used to get more insight into the characteristics and performance of the ensemble forecast.

3.1 Correlation

For a first overview of the data, one can look at the correlation of the daily mean values from observed and forecast time series. For COSMO-LEPS only the ensemble median is considered here. From Fig. 4 (middle and right panel), it can be seen that the data scatters more with increasing lead time. This is to be expected as the forecasts get more uncertain with lead time. With an R^2 of 0.86, COSMO-2 and COSMO-7 outperform the COSMO-LEPS ensemble median on the first forecast day. Thereafter COSMO-LEPS ensemble median outperforms COSMO-7 on forecast days two and three.

3.2 Frequency Bias

The frequency bias sets the predicted frequency of a binary event (here the exceedance of a predefined discharge threshold) in relation to the observed frequency of an event of the same kind (number of predicted threshold exceedances/number of observed threshold exceedances). A bias of one means that the event was equally often predicted as it was observed. Bias higher or lower than one indicates a relative over- and underforecasting. The frequency bias, however, does not say anything about temporal coincidence of observed and predicted events.

The frequency bias for the deterministic COSMO-2- and COSMO-7-based forecasts and for the median of the COSMO-LEPS-based ensemble forecast is presented here as a function of lead time and for different flood quantiles (50%–98%).

For the Sihl forecasting system, the peak flows are the most relevant. Therefore, only thresholds higher than the observed median hourly discharge of the forecast period were considered for the analysis. However, these thresholds are still very low compared to the discharge values that are of interest for the purpose

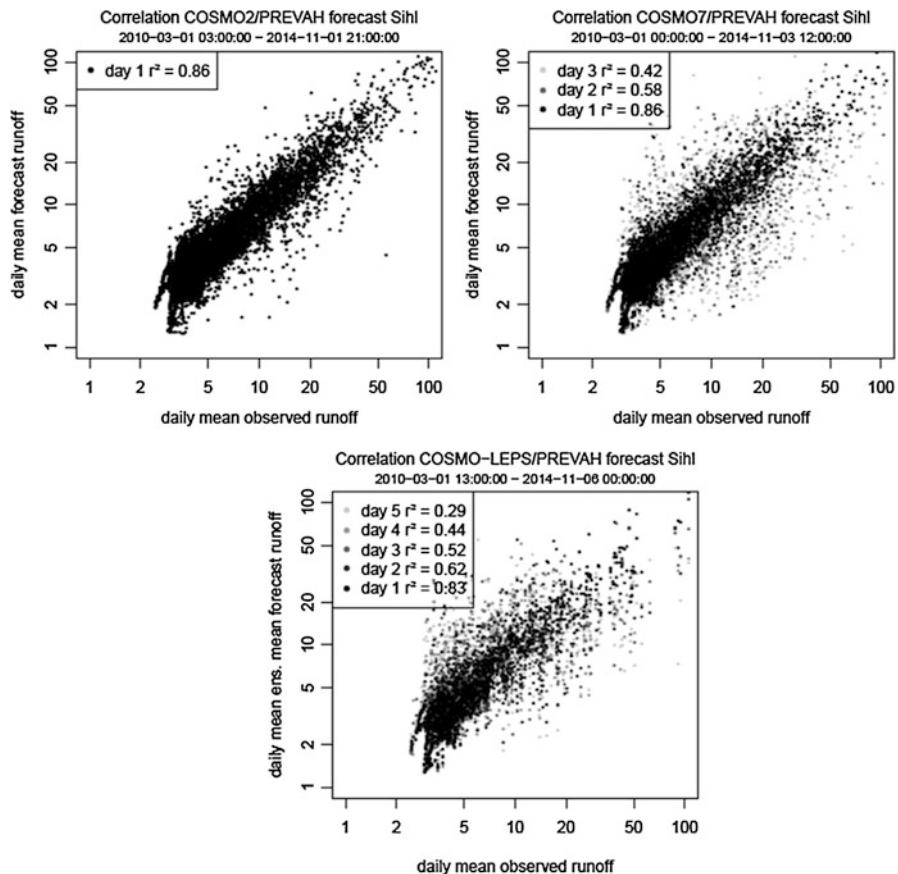


Fig. 4 Correlation between forecast and observed mean daily discharge. Dots are colored by lead time. For the COSMO-LEPS model chain, the forecast mean daily discharge was calculated on the basis of the ensemble mean

of this flood forecasting system. The 98% quantile of the observed hourly discharge is around $40 \text{ m}^3/\text{s}$ (max is around $190 \text{ m}^3/\text{s}$), whereas the first warning level defined by the stakeholders is at $100 \text{ m}^3/\text{s}$. So the results for the frequency bias presented here are useful information about the general tendency of the system and not about the behavior of the system during flood conditions. Also the daily cycle that is visible in the bias over the 132 h of lead time (Fig. 5) results from small-scale daily fluctuations in the discharge that are superposed in periods with high discharge; thus, these fluctuations are irrelevant with respect to flood forecasting. So the information about the frequency bias presented here as a function of lead time and for different flood quantiles is a general one. Figure 5 is rather blue, corresponding to a frequency bias below 1, which indicates that the forecasts more often underpredict the discharge on the tested thresholds.

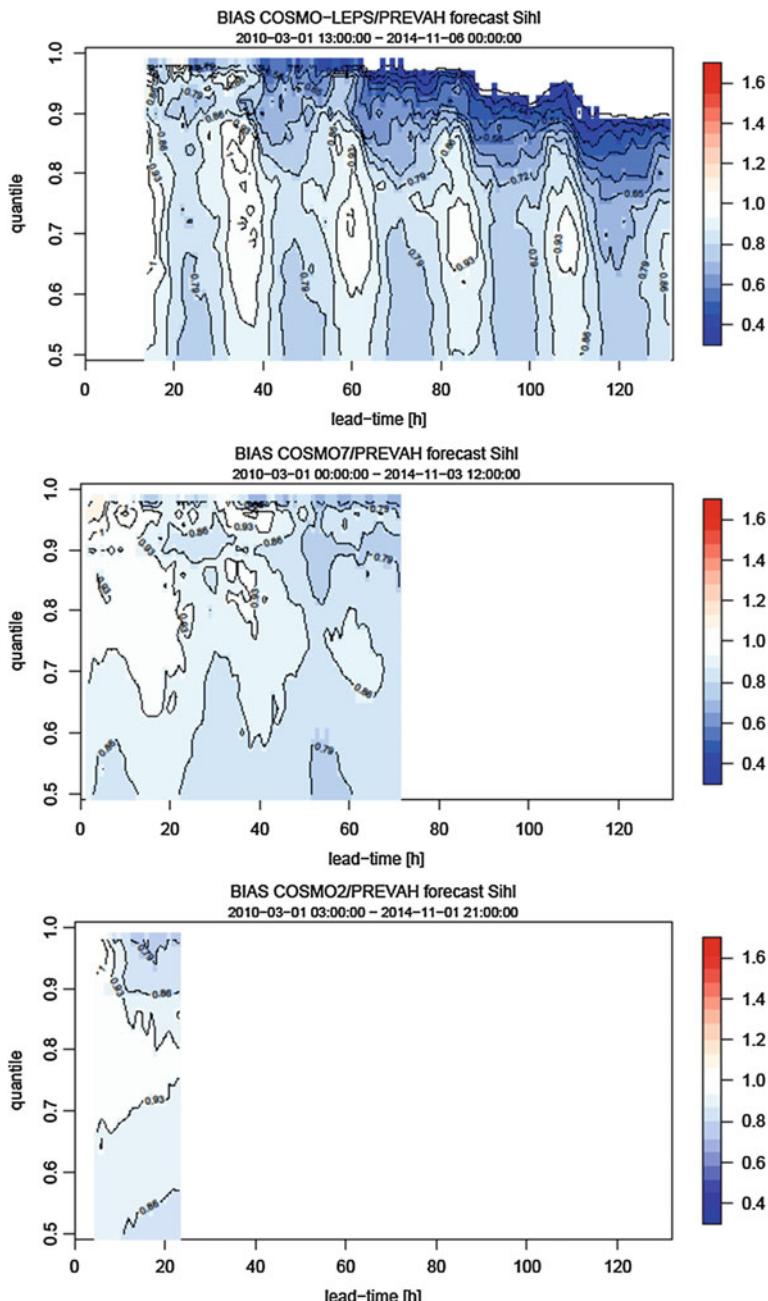


Fig. 5 Frequency bias for the COSMO-LEPS median (*top*), COSMO-7 (*center*), and COSMO-2 (*bottom*). Values below 1 indicate underforecasting. 1 means no bias. Frequency bias for COSMO-LEPS median at high lead times and high quantiles is below 0.3 and not colored

3.3 Value Score

The value score is a measure for the economic value of a forecast system. It is dependent on the cost-loss ratio of the specific user. The cost is the amount of money the user has to pay for the prevention of damages through an event; here these are the costs for the hydrological forecast system. The loss is the amount of money the end user loses if an event takes place without having taken any preventive measures. Meaningful cost/loss ratios vary between zero and one. For cost/loss ratios higher than one, a preventive measure does not pay off. End users with a cost/loss ratio close to zero will suffer big damages in case of an event.

An end user should take preventive measures as soon as the forecast probability of an event is higher than the cost/lost ratio of the end user. If the forecast system reaches a value score higher than zero, the end user can profit from the system. A value score of one means that the events were perfectly forecast and the end user only has to pay the cost for the preventive measure.

The Sihl forecast system was analyzed for the 80%, 90%, and 95% discharge quantiles for different lead times and cost/loss ratios ranging from 0 to 1 (Fig. 6). The value score for the Sihl forecast system is particularly good for users with a low cost/loss rate and also for long lead times. As mentioned in the introduction of Sect. 3, this is the case for the authorities of the city of Zurich, and thus they can profit from the HEPS.

3.4 Brier Skill Score

The Brier skill score has been proven to be a suitable measure to compare deterministic and ensemble forecasts. Unlike deterministic measures it is not necessary to reduce the ensemble to its median. However, it can be recommended to use the ensemble median as an additional deterministic forecast to assess the amount of additional skill that results from the ensemble spread only. The comparison of the score from the ensemble median with the scores for other deterministic forecasts, here COSMO-2 and COSMO-7, is straightforward and shows the minimal gain that can be achieved by the ensemble forecast. This may also be used to communicate with end users that are still skeptic to ensemble forecasts.

For flood forecasts the magnitude and timing of peak flows are most important. Depending on the available data one can work with the daily maxima of forecasts and observations (or reference simulation), or if data is available at, for example, hourly time steps, it is possible to choose a moving window for which the maximum values are extracted. In this way it is possible to choose an appropriate tolerance in timing for the peak flow forecasts. However, it is to be considered that in most cases, new model runs are not available at every time step. This means that forecasts from different lead times are mixed with the moving window method. The advantage of the moving window method above the daily maxima method is, however, that it avoids classifying an event as a missed event in cases where the forecast and the observed peak are close together but on different calendar days.

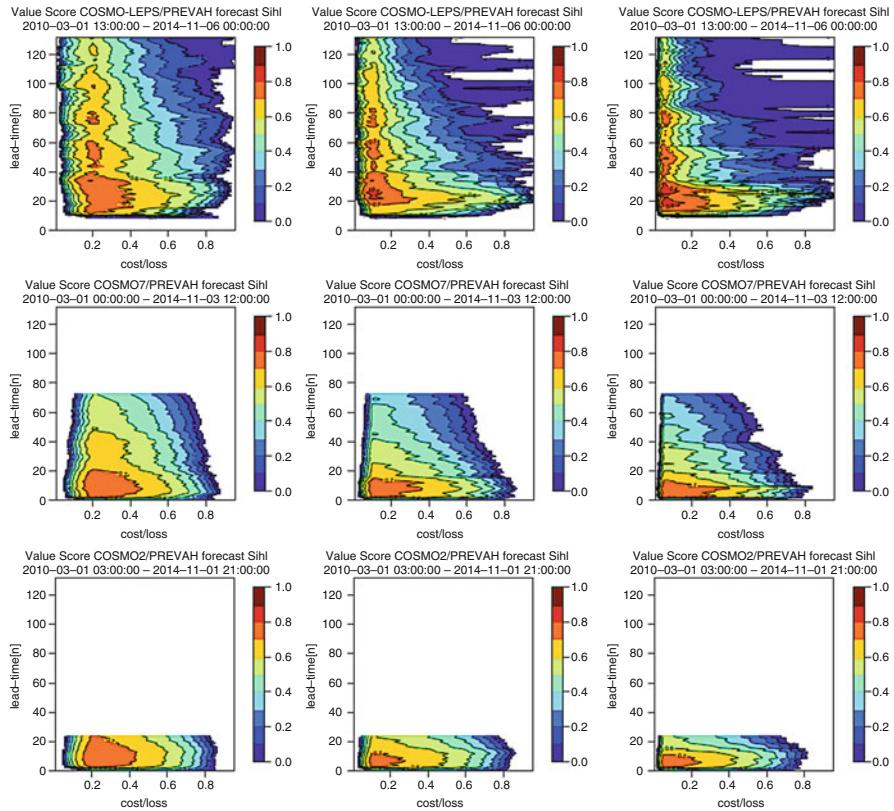


Fig. 6 Value score of forecasts based on COSMO-LEPS, COSMO-7, and COSMO-2 for different lead times (y-axes) and cost/loss rates (x-axes). Displayed are the results for events with a 20% (left), 10% (center), and 5% (right) probability of occurrence

For the analysis of the Sihl forecasting system, a moving window of 13 h was chosen. The tested thresholds are equal to the 80%, 90%, 95%, and 99% quantiles of the climatology. Where the climatology here corresponds to the observed 13 h maxima from the months March to October of the years 2007–2014. To estimate the uncertainty of the BSS values, a bootstrap was applied, as described in Sect. 2.1.

Because of the 13 h moving window, there are no data points for the first 6 h of the forecast. For better comparison of the BSS from the three forecast chains (COSMO-2, COSMO-7, and COSMO-LEPS), the first BSS value of each chain corresponds to the BSS value for the seventh hour of available hydrological forecast (Fig. 7). Due to the short computation time, for COSMO-2 and COSMO-7 this is exactly the seventh forecast hour. For COSMO-LEPS, however, this is the 17th forecast hour, because the first 10 h of the forecast (12–21 UTC) have already passed until the hydrological forecast is completed. So the underlying forecast of the COSMO-LEPS forecast chain is much older once the hydrological forecast is completed. Nonetheless, the COSMO-LEPS forecast chain performs remarkably

well compared to the deterministic forecasts (Fig. 7). The BSS of all forecast chains decays with increasing lead time. However, the decay is slower for the ensemble forecast chain. The BSS of the ensemble forecast is positive over the whole forecast period of 5 days and also for very high thresholds (99% – quantile). The performance of the ensemble median is mostly as good as or better than the performance of the

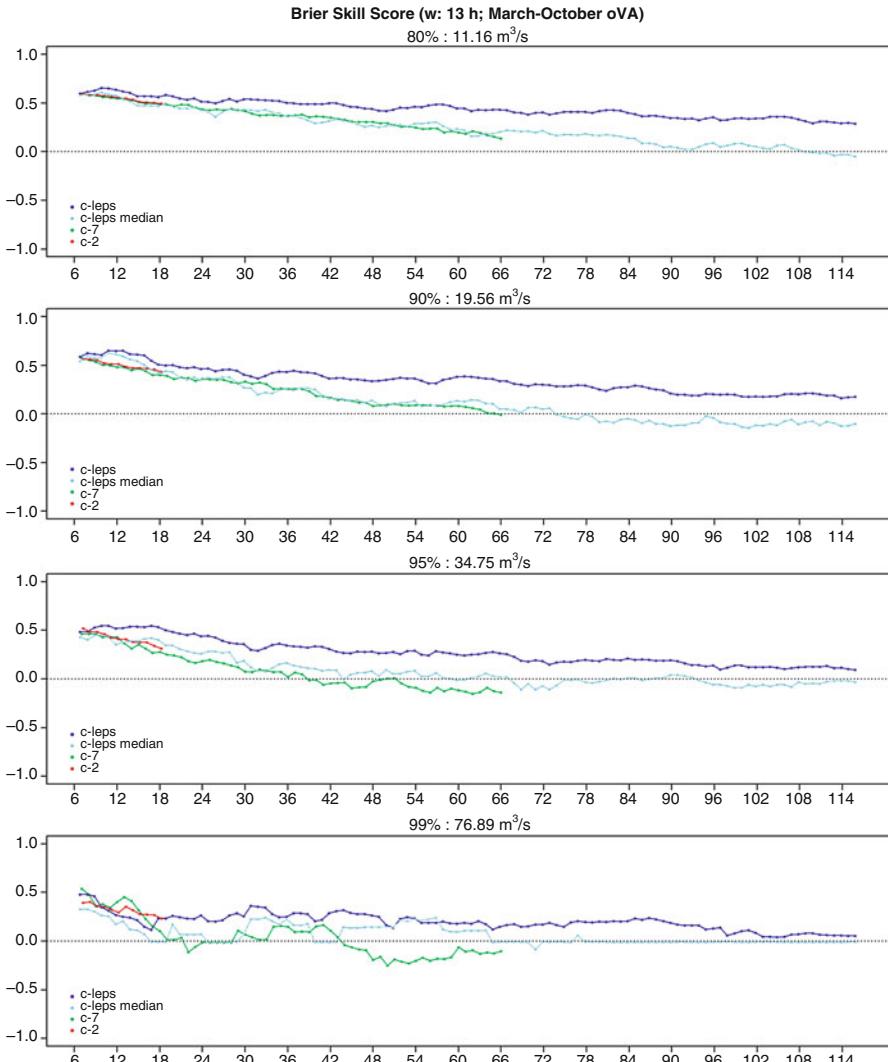


Fig. 7 Brier skill score. Values were estimated by applying a bootstrap with 500 iterations. Points displayed refer to the median of the bootstrap. Thresholds relate to 13 h max values from the months March to October from the period of 2007 to 2014. Labels on x-axes refer to the center of the 13 h running window

COSMO-7 forecast chain and additionally provides longer lead time. Also, the BSS values for the COSMO-2 forecast chain is as good as or better than the values of the COSMO-7 chain. This means that for the forecast system presented here, the COSMO-7 chain is redundant, and the forecasters and users should base their decisions on COSMO-2 and COSMO-LEPS.

The 95% confidence bounds resulting from the bootstrap show that the uncertainty in the estimated BSS values increase with increasing threshold (Fig. 8).

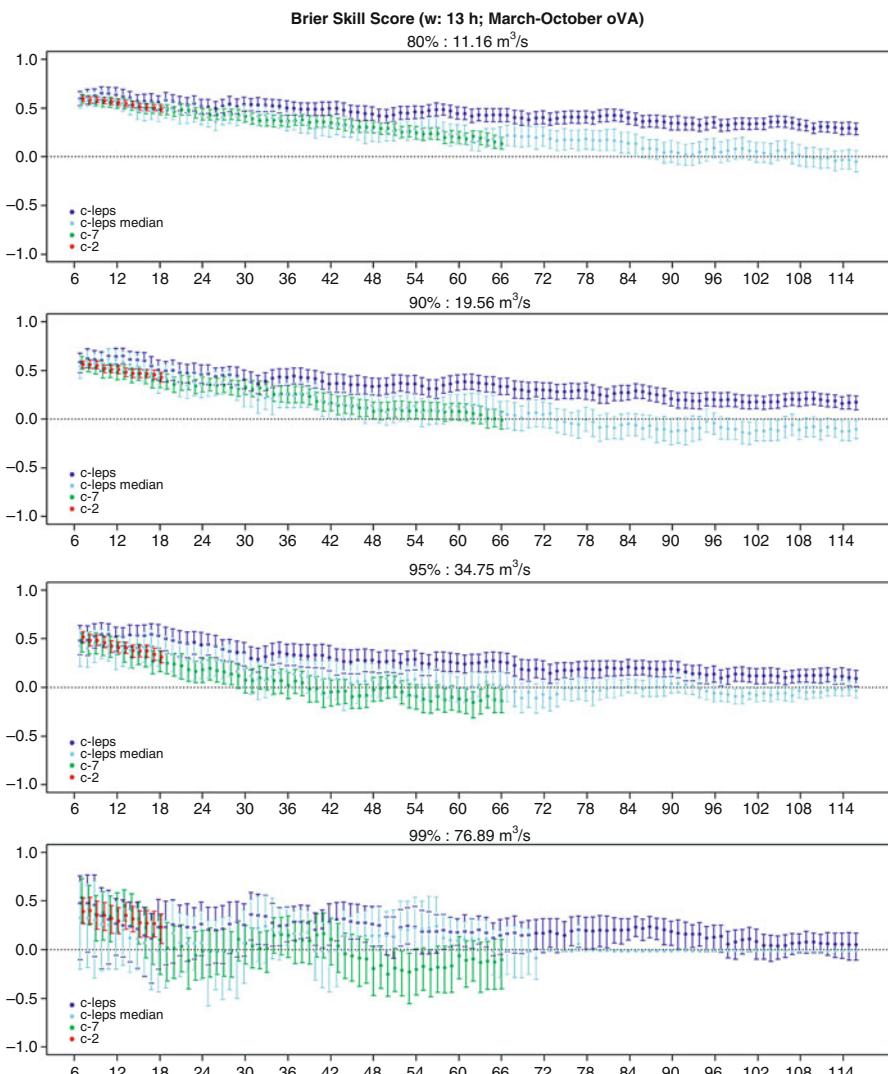


Fig. 8 Same as Fig. 7, additionally showing the 95% confidence intervals of the estimated BSS, based on a bootstrap with 500 iterations

However for the ensemble forecast, the confidence bounds do not cross the zero BSS line for tested thresholds up to the 95% quantile. This shows that, in contrast to the deterministic forecast, the ensemble forecasts have skill (in terms of Brier score) for the whole forecast horizon.

3.5 Rank Histograms

The COSMO-LEPS ensemble consists of 16 members, so ideally each rank of the rank histogram would be populated with a relative frequency of 1/17. This is not the case for the ensemble forecasts of the Sihl forecast system. The first and last rank are populated much more often, indicating an underdispersion of the ensemble (Fig. 9). Also, the histogram is skewed toward the highest rank indicating an

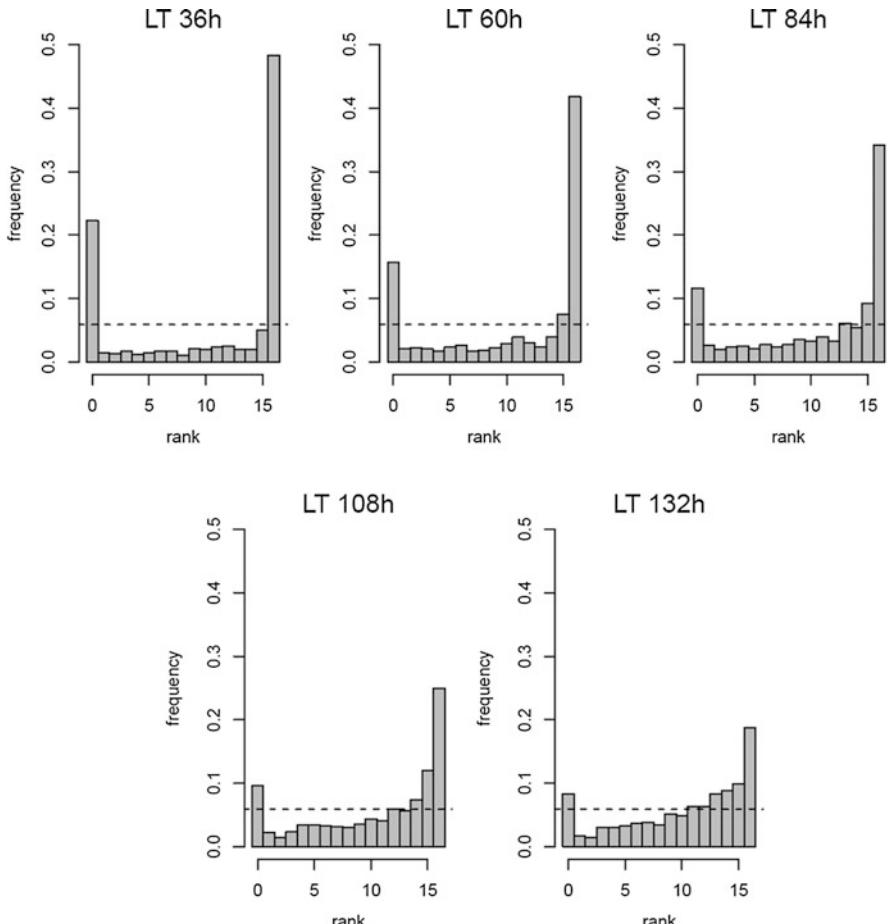


Fig. 9 Rank histograms for forecast days 1–5 (corresponding to lead time 36, 60, 84, 108, 132 h)

underforecasting bias. This is an overall analysis, thus the results are dominated by low-flow periods. In periods with little or no precipitation, the ensemble members rapidly converge, and thus often all ensemble members are above or, as seen from Fig. 9, even more often below the observation.

3.6 Conditional Bias: Reliability Plots

Reliability plots, also called attribute diagrams, are used to analyze ensemble forecasts. Compared to scalar summary-scores like the BSS, the reliability plots are much more comprehensive as they show the full joint distribution of forecasts and observations for events of a specific magnitude. They are created by binning all forecast-observation pairs according to the forecast probability (Wilks 2006).

The basic features of the reliability plot are the diagonal 1:1 line, the no resolution line, the no skill line, and of course the reliability curve.

The diagonal 1:1 line is the line for which the forecast probability and the observed frequency are equal. The no resolution line is the line that marks the climatological frequency (the observed mean frequency) of the event in question. The reliability curve, the red line on Fig. 10, consists of the points resulting from plotting the forecast probabilities of the bins against the corresponding observed frequencies. From the position of the points on the plot information about forecast reliability and resolution can be gained. The reliability is the mean squared vertical distance of the points on the reliability curve to the 1:1 line. The resolution is the mean squared vertical distance of the points on the reliability curve to the no resolution line. As long as the absolute value of the resolution term is bigger than the reliability term the forecast adds to a positive skill. The no skill line marks the line on which reliability and resolution are equal. All points that fall into the shaded area add positively to the forecast skill.

The position of the reliability curve also exhibits information about conditional and unconditional bias of the forecast. If the overall slope of the reliability curve differs from the slope of the 1:1 line a conditional bias exists. If the reliability curve is parallel but below or above the 1:1 line an unconditional bias (systematic bias) exists and the forecast is said to be overforecasting and underforecasting, respectively.

To complete the reliability plot the relative frequencies of the forecasts in the forecast bins (refinement distribution) are written beside the points on the reliability curve. Alternatively, this information can be plotted in a histogram. It gives information about the forecast's sharpness and the confidence the forecaster can have in the estimated reliability curve. It illustrates how well the forecast can distinguish situations that are different from the climatology.

The reliability plots for the Sihl forecast system (Fig. 10) indicate that the higher the tested discharge threshold and the higher the forecast probability of exceeding this threshold, the more an overforecasting bias is present. However, this result has to be taken with caution, because the bins at high thresholds and high forecast probabilities are very scarcely populated. Nonetheless, the reliability plots show that the forecasting system has skill for the tested thresholds.

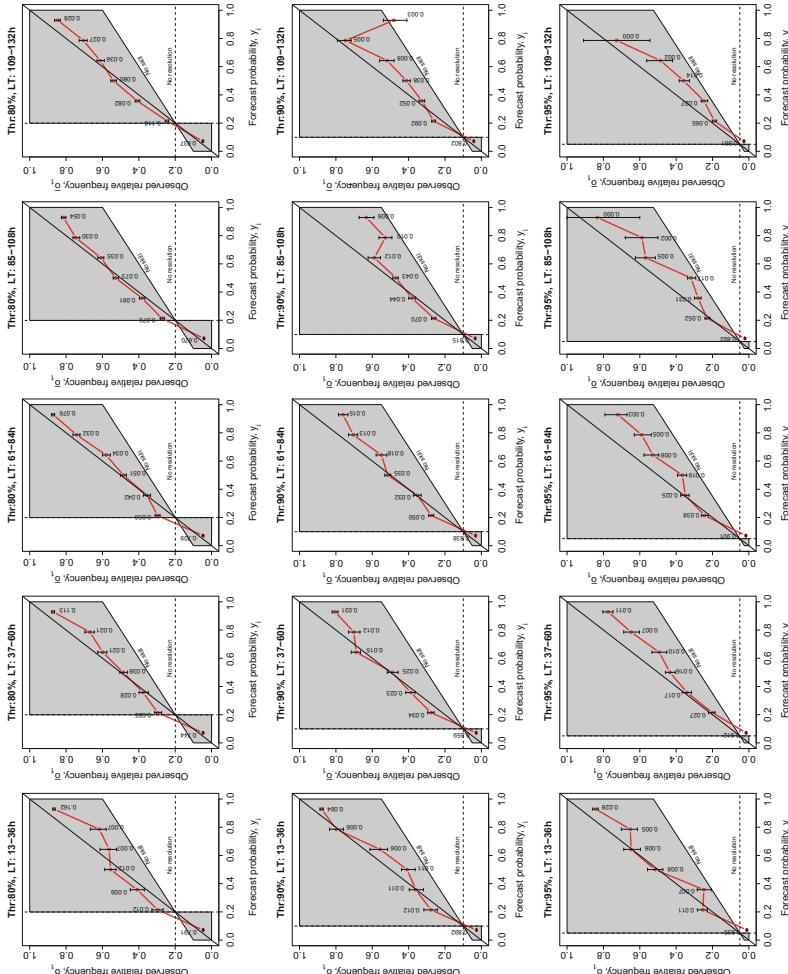


Fig. 10 Forecast probability (x-axis) and observed frequency (y-axis) of the COSMO-LEPS-based forecasts for seven different probability classes. The numbers next to the estimated frequencies correspond to the relative frequency of events in the classes. The tiles are sorted by bins of lead times (columns) and discharge thresholds (rows). The thresholds correspond to the 80% (ca. 8.7 m³/s), 90% (ca. 14 m³/s), and 95% (ca. 21.5 m³/s) quantile

3.7 Performance Overview

When it comes to the interpretation of the case study presented in this chapter, the following main points can be listed:

- Correlation: On the first forecast day, the daily mean runoff forecast is better when derived using the COSMO-2- and COSMO-7-weather forecast. COSMO-LEPS outperforms COSMO-7 on forecast days two and three.
- Frequency bias: For the tested discharge thresholds, the forecasts are rather underpredicting the discharge.
- Brier skill score: The BSS values achieved by the ensemble forecast are positive also for long lead times and high thresholds. The skill of the ensemble forecast in terms of Brier score is generally better than the skill of the deterministic forecasts.
- Rank histogram: The rank histograms for the ensemble forecast of the Sihl forecast system indicate an underdispersion of the ensemble and an underforecasting bias.
- Value score: The end user with a low-cost loss rate would benefit from the information provided by the ensemble forecasts.
- Reliability plots: The higher the tested threshold and the higher the forecast probability of exceeding this threshold, the more an overforecasting bias can be seen from the reliability plot.

For the lead times relevant for the hydrological forecast system for the Sihl river in Zurich, the COSMO-LEPS-based ensemble forecasts are clearly preferred over the deterministic forecasts. Of course there are other measures that can be applied to evaluate a HEPS (Brown et al. 2010; Wilks 2006), and these might be more appropriate to evaluate other aspects than the here addressed question on the usefulness of ensemble predictions. Prior to each verification effort, the forecaster needs to carefully select verification metrics that are meaningful to the current application and not redundant in describing a given aspect of the forecast quality. However, if the main findings do not change, it can be recommended to keep things as simple as possible. This can also ease the communication of the results to the end users.

4 Flood Peak and Flood Timing Verification

In Sects. 2 and 3, it was shown that it pays off to work with the ensemble approach. Ensemble forecasts provide valuable information about the uncertainty of the forecast. However, the interpretation of ensemble forecasts can be somewhat difficult, especially when the ensemble spread is high, as it is often the case when forecasting flood events. Therefore Zappa et al. (2013) developed the “Peak-Box.” The Peak-Box is a simple but efficient tool to ease the interpretation of ensemble forecasts and to support decision-making when facing a forecast flood event. For decision makers the crucial question is how high and when the flood will be. The Peak-Box extracts

from the ensemble summary information to best address this question. The Peak-Box is defined as the best estimate of the timing and magnitude of a forecast flood event. It frames the discharge peaks of all members of the ensemble forecast and then takes the median in timing and magnitude of these peaks.

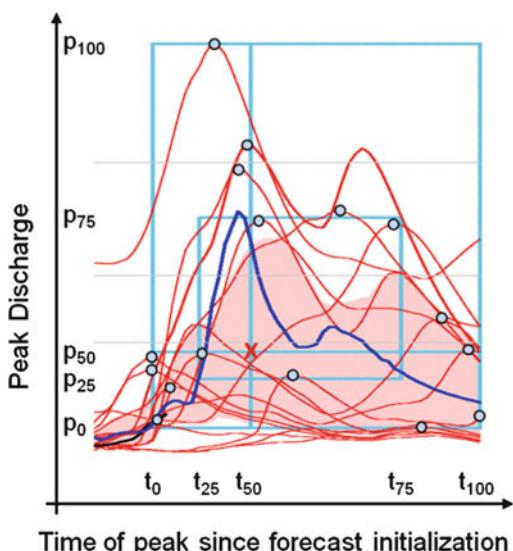
The steps to get the Peak-Box are the following (Fig. 11):

- Mark the highest peak of each member.
- Draw boxes limited by quantiles in timing and magnitude of the ensemble member peaks.
- “Best estimate”: Intercept of median in peak timing and in peak magnitude (x).

This tool has been proven to be useful for flood peak estimation in the operational use of HEPS (Zappa et al. 2013), with some obvious limitations in the assessment of peak flow timing in case of multi-peak forecast events. It is also a good tool to show people the advantage of ensemble forecasts, and it can be related to the work of Francis Galton, who established the concept of the wisdom of the crowds (Galton 1907). It means that for a specific question the collective answer of a group of individuals is trusted more than the answer of a single expert. Translated to hydrological forecasting and referring back to the leading question of this chapter, this would mean that ensemble forecasts (group of individuals) should be preferred over deterministic forecasts (single expert) because they should be more trustworthy.

To test this a game was played with 162 persons, all experts in the field of flood forecasting. They were given spaghetti plots of four ensemble hindcasts each consisting of 16 members for a flood event. Their task was to estimate the timing and magnitude of the peak that was observed for that event. Then for each estimate

Fig. 11 Illustration of how to draw the Peak-Box. Small circles indicate the peaks of each ensemble member. The crossline indicates the median in peak timing and magnitude. The inner rectangle frames the interquartile of peak timing and magnitude, and the outer rectangle frames the whole range of peak timing and magnitude



of an expert, the distance in timing and magnitude to the observed peak was calculated. The same was done for the best estimate from the Peak-Box. The result was that none of the experts made a peak estimate better than the Peak-Box, and only four of the experts scored equal to the Peak-Box. So the peak estimate of the ensemble was better than 158 out of 162 experts.

However, if the median peak estimate of all experts was considered, this expert median was only slightly inferior compared to the Peak-Box estimate. Only nine of the experts made a better estimate than the expert median. Therefore this example supports the concept of the wisdom of the crowds and shows the value of working with ensembles instead of deterministic forecasts. Even the median of an ensemble should generally be preferred over a deterministic forecast. This game is freely available from the HEPEX resources web page (<http://hepex.irstea.fr/resources/>). An analysis of the game is presented as a blogpost at <http://hepex.irstea.fr/heps-challenges-the-wisdom-of-the-crowds/>. The Peak-Box is a good tool to summarize the information in an ensemble when it comes to flood forecasts.

5 Conclusions

The examples in Sect. 2 showed that ensembles generated with different parameter sets as well as ensembles generated from varying meteorological input have an additional value compared to deterministic nowcasts, when it comes to peak flows. The ensembles resulting from applying different model parameterization showed a significant underdispersion. This was somewhat less pronounced for the ensembles generated from varying input, i.e., radar ensemble. The combination of the two approaches would certainly ease the issue with the underdispersion (Zappa et al. 2011). A general recommendation cannot be given, but for every system, the preferred approach has to be found based on data resources and computational resources.

In Sect. 3 the three model chains COSMO-2, COSMO-7, and COSMO-LEPS of the HEPS for the Sihl river were analyzed and compared. A basic comparison between forecast and observed daily means showed that the ensemble forecast scores better than the deterministic forecast beyond day one of lead time. A general investigation of the frequency bias indicates that the system is rather underforecasting the discharge. However a more detailed analysis of the ensemble forecast based on the conditional bias showed that the higher the tested threshold of peak discharge and the higher the forecast probability of exceeding this threshold, the more an overforecasting bias can be seen.

Even though the COSMO-LEPS ensemble forecasts are 10 h older than the deterministic COSMO-2 and COSMO-7 forecasts when they get available for the end user, they perform remarkably well in comparison to the deterministic forecasts. The BSS of the COSMO-LEPS ensemble forecast is positive over the whole forecast period of 5 days and also for very high thresholds (99% – quantile). For the 80%, 90%, and 95% thresholds, the skill of the ensemble forecast outperforms the deterministic forecasts, while for the 99% threshold, this is the case after about

18 h. Because flood mitigation for the Sihl river in Zurich needs lead times of 1 day and more, the COSMO-LEPS-based ensemble discharge forecasts are much more valuable than the deterministic forecasts.

Section 4 finally presented the Peak-Box, a tool for the interpretation of ensemble forecasts of floods. It summarizes the information about timing and magnitude of the different ensemble members and provides a “best guess” for the forecast event’s peak flow timing and magnitude. Additionally it can be used as a verification measure with respect to flood peak timing and magnitude predictions (Zappa et al. 2013).

Generally it was shown that there is added value in the application of ensemble predictions. To which extent this applies depends certainly on the problem at hand, but for flood-related problems, it is highly recommended to work with ensemble forecasting systems. Nevertheless, this chapter demonstrated how systematic and robust verification could help forecasters and users in the analysis of (ensemble) prediction performance for specific situations and therefore maximize the value of the particular forecast system.

References

- N. Addor, S. Jaun, M. Zappa, An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* **15**(7), 2327–2347 (2011)
- N. Andres, G. Lieberherr, I.V. Sideris, F. Jordan, M. Zappa, From calibration to real-time operations: an assessment of three precipitation benchmarks for a Swiss river system. *Meteorol. Appl.* **23**(3), 448–461 (2016)
- A. Badoux et al., IFKIS-Hydro Sihl: Beratung und Alarmorganisation während des Baus der Durchmesserlinie beim Hauptbahnhof Zürich. *Wasser Energ. Luft* **102**(4), 309–320 (2010)
- K. Beven, J. Freer, Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *J. Hydrol.* **249** (1–4), 11–29 (2001)
- K. Bogner, K. Liechti, M. Zappa, Post-processing of stream flows in Switzerland with an emphasis on low flows and floods. *Water* **8**(4), 115 (2016)
- J.D. Brown, J. Demargne, D.-J. Seo, Y. Liu, The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Software* **25**(7), 854–872 (2010)
- B. Efron, 1977 Rietz lecture – bootstrap methods – another look at the jackknife. *Ann. Stat.* **7**(1), 1–26 (1979)
- F. Fundel, A. Walser, M.A. Liniger, C. Frei, C. Appenzeller, Calibrated precipitation forecasts for a limited area ensemble forecast system using reforecasts. *Mon. Weather Rev.* **138**(1), 176–189 (2010)
- F. Galton, The wisdom of crowds. *Nature* **75**(1949), 450–451 (1907)
- U. Germann, M. Berenguer, D. Sempere-Torres, M. Zappa, REAL – Ensemble radar precipitation estimation for hydrology in a mountainous region. *Q. J. Roy. Meteorol. Soc.* **135**(639), 445–456 (2009)
- S. Hemri, F. Fundel, M. Zappa, Simultaneous calibration of ensemble river flow predictions over an entire range of lead times. *Water Resour. Res.* **49**(10), 6744–6755 (2013)
- S. Jaun, B. Ahrens, A. Walser, T. Ewen, C. Schär, A probabilistic view on the August 2005 floods in the upper Rhine catchment. *Nat. Hazards Earth Syst. Sci.* **8**(2), 281–291 (2008)

- G. Kuczera, E. Parent, Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. *J. Hydrol.* **211**(1–4), 69–85 (1998)
- R. Lamb, Calibration of a conceptual rainfall-runoff model for flood frequency estimation by continuous simulation. *Water Resour. Res.* **35**(10), 3103–3114 (1999)
- K. Liechti, L. Panziera, U. Germann, M. Zappa, The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrol. Earth Syst. Sci.* **17**, 3853–3869 (2013a)
- K. Liechti, M. Zappa, F. Fundel, U. Germann, Probabilistic evaluation of ensemble discharge nowcasts in two nested Alpine basins prone to flash floods. *Hydrol. Process.* **27**(1), 5–17 (2013b)
- P.V. Mandapaka, U. Germann, L. Panziera, A. Hering, Can Lagrangian extrapolation of radar fields be used for precipitation nowcasting over complex Alpine orography? *Weather Forecast.* **27**(1), 28–49 (2012)
- C. Marsigli, F. Boccanera, A. Montani, T. Paccagnella, The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlinear Processes Geophys.* **12**(4), 527–536 (2005)
- J.E. Nash, J.V. Sutcliffe, River flow forecasting through conceptual models (1), a discussion of principles. *J. Hydrol.* **10**, 282–290 (1970)
- L. Panziera, U. Germann, The relation between airflow and orographic precipitation on the southern side of the Alps as revealed by weather radar. *Q. J. Roy. Meteorol. Soc.* **136**(646), 222–238 (2010)
- D. Viviroli, H. Mittelbach, J. Gurtz, R. Weingartner, Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part II: Parameter regionalisation and flood estimation results. *J. Hydrol.* **377**(1–2), 208–225 (2009a)
- D. Viviroli, M. Zappa, J. Gurtz, R. Weingartner, An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. *Environ. Model. Software* **24**(10), 1209–1222 (2009b)
- D. Viviroli, M. Zappa, J. Schwanbeck, J. Gurtz, R. Weingartner, Continuous simulation for flood estimation in ungauged mesoscale catchments of Switzerland – Part I: Modelling framework and calibration results. *J. Hydrol.* **377**(1–2), 191–207 (2009c)
- J.A. Vrugt, C.J.F. ter Braak, M.P. Clark, J.M. Hyman, B.A. Robinson, Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation. *Water Resour. Res.* **44**, W00B09 (2008)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences* (Elsevier, Amsterdam, 2006), p. 627
- M. Zappa et al., IFKIS-Hydro Sihl: Ein operationelles Hochwasservorhersagesystem für die Stadt Zürich und das Sihltal. *Wasser Energ. Luft* **102**(3), 238–248 (2010)
- M. Zappa, S. Jaun, U. Germann, A. Walser, F. Fundel, Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmos. Res.* **100**(2–3), 246–262 (2011)
- M. Zappa, F. Fundel, S. Jaun, A ‘Peak-Box’ approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. *Hydrol. Process.* **27**(1), 117–131 (2013)
- M. Zappa et al., Crash tests for forward-looking flood control in the city of Zürich (Switzerland). *Proc. IAHS* **370**, 235–242 (2015)



Verification of Medium- to Long-Range Hydrological Forecasts

Luc Perreault, Jocelyn Gaudet, Louis Delorme, and Simon Chatelain

Contents

1	Introduction	978
2	Probabilistic Hydrological Forecasts at Hydro-Québec (HQ)	979
2.1	Midterm Hydrological Forecast	980
2.2	Long-Term Hydrological Forecast	982
2.3	Case Studies	982
3	Assessing Hydrological Probabilistic Forecasts: Statistical Framework	984
4	Estimation of Scores	990
5	Medium-Term Hydrological Forecasts Verification: Application to Real Data	993
6	Using a Production Planning Model to Evaluate Long-Term Hydrological Forecast Value	998
7	Conclusion	1008
	References	1011

Abstract

Hydrological forecasting is crucial for hydropower production and risk management related to extreme events. Since uncertainty cannot be eliminated from such a process, forecasts should be probabilistic in nature, taking the form of probability distributions over future events. However, verification tools adapted to probabilistic hydrological forecasting have only been recently considered. How can such forecasts be verified accurately? In this chapter a simple theoretical framework proposed by Gneiting et al. (2007) is employed to provide a formal guidance to verify probabilistic forecasts. Some strategies and scoring rules used

L. Perreault (✉) · J. Gaudet (✉) · L. Delorme (✉)
IREQ Hydro-Québec Research Institute, Varennes, QC, Canada
e-mail: perreault.luc@ireq.ca; gaudet.jocelyn@ireq.ca; delorme.louis@ireq.ca

S. Chatelain (✉)
McGill University, Montreal, QC, Canada
e-mail: simon.chatelain@mail.mcgill.ca

to measure the performance of hydrological forecasting systems, namely, Hydro-Québec, are presented. Monte Carlo simulation experiments and applications to a real archive of operational medium-range forecasts are also presented. An experiment is finally performed to evaluate long-range hydrological forecasts in a decisional perspective, by employing hydrological forecasts in a stochastic mid-term planning model designed for optimizing electricity production. Future research perspectives and operational challenges on diagnostic approaches for hydrological probabilistic forecasts are given.

Keywords

Probabilistic forecasting · Hydrological forecasts · Proper scoring rules · Skill scores · Estimation · Multivariate verification · Energy score · Economic value of forecasts

1 Introduction

The purpose of forecasting is to support informed decision-making. Inflow forecasts are issued to help operate reservoirs, assess resource capabilities, evaluate risks, and determine pricing of hydro-energy. Hydrological forecasts are typically provided using hydrological models driven by some estimates of future weather; flows observed during the previous time period, a.k.a. persistence forecasts; or average flows, a.k.a. climatology forecasts. Hydropower companies and agencies make considerable efforts to provide accurate hydrological forecasts for various lead times, using physically based conceptual and statistical hydrological models, weather forecasts, and both deterministic and probabilistic techniques. The ability to provide reliable and accurate medium- and long-range hydrological forecasts is fundamental for the effective operation and management of water resource systems. For instance, at Hydro-Québec, daily hydrological forecasts serve as input to stochastic decision-making models designed for optimizing electricity production while avoiding spillage and inundation. Very important decisions therefore depend upon these forecasts.

Advanced forecast verification measures are critical to the assessment of operational impacts of inflow events (Weber et al. 2006; Perreault 2013; Alfieri et al. 2014). In fact, it gives objective information on various aspects and conditions of the forecast quality, which should benefit the forecasters and users in their decision-making process. Insufficient knowledge of the forecast skill eventually translates into uncertainty on the level of risk adopted into operations and may lead to a lack of confidence in the forecasts and suboptimal decisions. To communicate forecast skill to decision-makers and forecasters but also to establish benchmarks of forecasting systems, a variety of performance measures have been developed. The evaluation of probabilistic forecasts is a challenging task since it involves the comparison of two different quantities: functions (the probability distributions of the forecasts) and real values (the observations), which are not directly comparable. This situation complicates the verification framework and the interpretation of verification results and is a

contributing factor explaining in part the wide variety of performance measures that exists.

In this chapter, we address both the assessment of the forecast quality and the assessment of the forecast value. We first focus on the use of statistical verification measures and skill scores to assess the quality of midterm probabilistic hydrological forecasts. Then, we go a step further by presenting an experiment to evaluate the economic value of long-term hydrological forecasts using a stochastic planning model specifically designed for the optimization of Hydro-Québec's (HQ) electricity production. HQ ranks among the world's largest electric companies. For more than 50 years, Hydro-Québec has generated, transmitted, and distributed nearly all the electricity consumed in Québec, now totaling more than 176 TWh a year. It has an installed capacity of 36,912 MW. It also sells power on wholesale markets in northeastern North America. Hydrological forecasting is thus a central activity for this public company.

Section 2 briefly describes Hydro-Québec's hydrological forecasting system and the case studies considered herein. In Sects. 3 and 4, we explain the verification statistical framework adopted at Hydro-Québec and the metrics in use and address the important problem of the estimation of scoring rules. Section 5 focuses on the application of verification tools to actual midterm forecasts, while Sect. 6 is dedicated to the economic value of long-term forecasts. We end the chapter with some conclusions and a number of questions about the use and development of forecast verification.

2 Probabilistic Hydrological Forecasts at Hydro-Québec (HQ)

One of the practical purposes of hydrology is to make inflow forecasts for the future. Hydrological processes, models, and variables associated with forecasting future hydrological states all contain a wide part of uncertainty. Therefore, hydrological forecasts should be probabilistic and expressed in the form of a probability distribution, the predictive distribution (Dawid 1984), in order to properly represent the phenomenon and what we know about it. Cloke and Pappenberger (2009) and Krzysztofowicz (2001) eloquently presented the need, the significance, and the unavoidable nature of probabilistic hydrological forecasts in today's world.

A wide variety of methodologies are used to produce probabilistic hydrological forecasts. The most common approaches are either based on hydrological models and a number of explanatory variables for the future states of the hydrometeorological system; another family of approaches uses statistical models. Hydro-Québec actual inflow forecasting system relies on both methodologies. On the one hand, daily midterm hydrological forecasts are produced for more than 90 basins by using a conceptual hydrological model and the approach of the extended streamflow prediction similar to the one proposed by Day (1985). On the other hand, once a year in December, a statistical model is employed to generate long-term spring volume forecasts for some major watersheds. Here, midterm hydrological forecasts

refer to a period ranging from 2 to 30 days; long-term hydrological forecasts refer to a period exceeding 30 days.

2.1 Midterm Hydrological Forecast

HQ's daily hydrological forecasts use as inputs a pseudo-probabilistic weather forecast derived from a perturbed deterministic weather forecast for the first 9 days. Historical error of the deterministic meteorological prediction (integrated at the basin scale) is used to create several scenarios from the available deterministic forecast. To extend the hydrological forecast beyond this meteorological forecast lead time, historical weather series are used as multiple possible inputs in the hydrological model. The end result is a hydrological forecast that extends at least 30 days in the future and possesses 189 flow series or members as of 2014 (Major changes are currently made to the actual HQ's hydrological forecast system. In particular, work is under way to integrate ensemble weather forecasts for mid-term hydrological predictions). Those scenarios are viewed as ensemble hydrological forecasts. While they do provide a distribution of forecasts, the distributional hypotheses are often hard to define and variable in time. This adds some challenges to the problem of evaluating the performance of hydrological probabilistic forecasts using statistical diagnostic tools.

From these members, distribution information, such as quantiles with certain probabilities of exceedance, can be estimated on the basis of some probability distributions. However, the assumption on the form of the probability distribution is, more often than not, wrong. By construction and hypothesis, we would expect a Pearson III type of distribution. The last several years of our archives of daily inflow forecasts rather show that for a lead time of 1–5 days, our forecasts usually can be approximated by standard symmetric (60% of the time) and asymmetric distributions (30%) or a mixture of distributions (10%). There is a shift for longer lead time, where we have a roughly equal amount of standard asymmetric and mixtures of distributions and very few symmetric distributions. Figure 1 presents typical forecast distributions, for a flow volume of 7 days, for one of HQ's large watershed. The wide variety of distributions is immediately visible upon examination of the figure. Bimodal (and sometimes trimodal) types of distributions, at first thought to be the results of a system malfunction, correspond to meteorological forecasts where there is a probability of heavy precipitation (or warm temperatures causing snowmelt) and a probability of light precipitation (or cold temperatures). They truly represent the future expected state of our hydrological system and should show a high level a forecasting skill, if the probabilities are properly represented. What is of concern for us, in terms of verification, is that we do not know what the true distribution is, and we do not know ahead of time what the distribution will look like.

Two challenges emerge from that situation for forecast verification. First, in the context of hydropower production, verification diagnostic tools must correctly reward a hydrological forecasting system which produces such detailed predictive densities, for instance multimodal predictive distributions. To our view, their

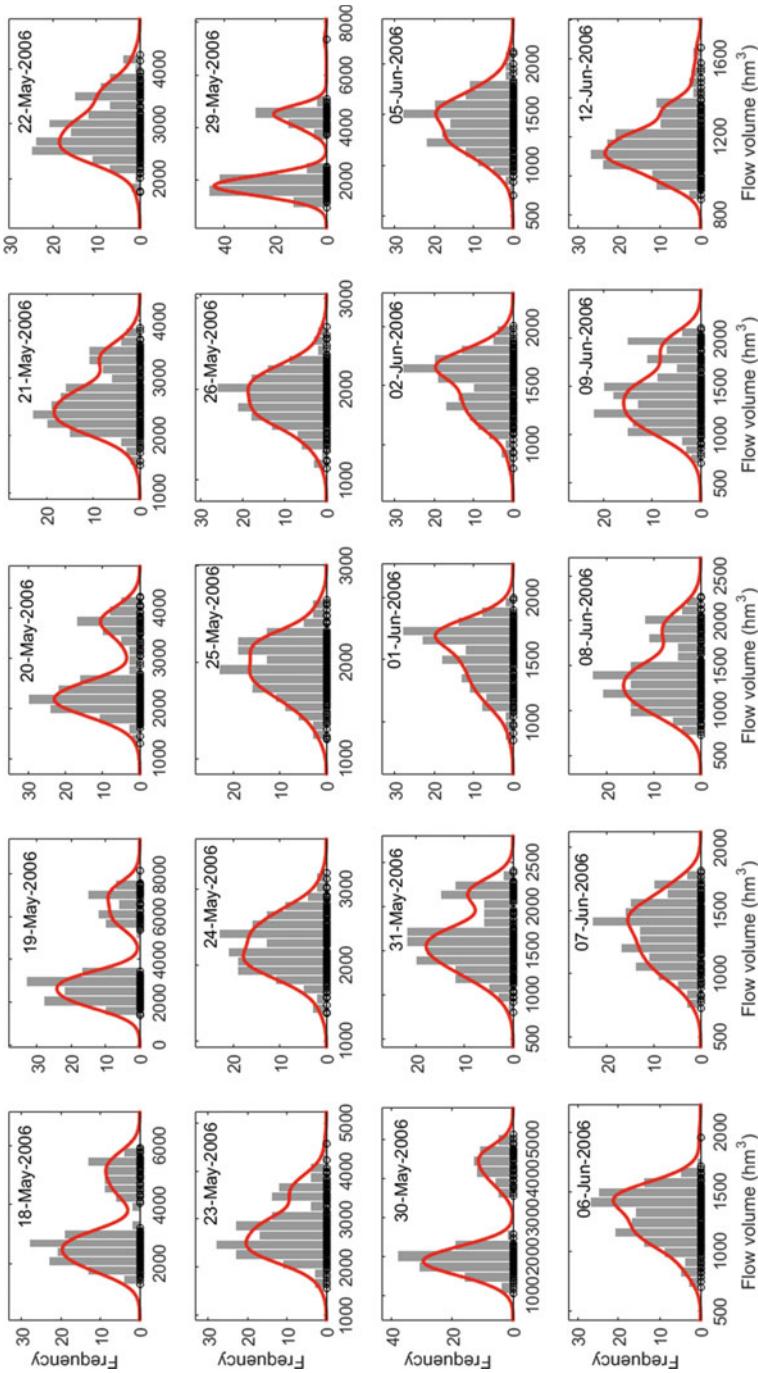


Fig. 1 Some typical forecast predictive distributions on a large watershed for the total flow volume of the next 7 days

sensitivity to the shape of the distribution is clearly an important property. We will expand on that problem in Sects. 3 and 5. Secondly, for a given scoring rule, we must make assumptions about the predictive distributions when calculating verification scores and metrics if a parametric estimation method is used. The choice of the estimation method is then an important issue. We will come back to that problem in Sect. 4.

2.2 Long-Term Hydrological Forecast

Other forecasting approaches to produce inflow forecasts are truly probabilistic and rely on a variety of statistical tools to issue a forecast. The most common approaches are based upon a probabilistic model such as a linear regression model, ARIMA, etc. Predictive distributions are generally known probability laws with such approaches, as they are consequent with the fundamental hypotheses. In this context, the verification problem is generally easier to tackle, as one is working with known family of distributions and has to make fewer hypotheses.

At HQ, a statistical approach is used to produce in December a probabilistic forecast of the aggregated May–July (MJJ) spring streamflow for several basins. In fact, an ability to issue a MJJ streamflow forecasts 5 months in advance is of particular interest to water planners and managers at HQ with respect to hydropower generation. The May–July aggregated flows count for about 50% of the total annual flow volume in the basins and are composed of melted winter snowpack and spring precipitation, with the prior winter months December–April being low-flow months. Thus knowledge in December of the future MJJ reservoir inflows can be used for making decisions about future releases during the winter contributing to a more proactive water management that may prove useful in extreme dry or wet years. The approach is based upon a linear regression model using the principal components of exogenous measures of atmospheric circulation inferred from the National Centers for Environmental Prediction/National Center for Atmospheric Research reanalysis project (see Sveinsson et al. 2008; Perreault et al. 2007). In the following, we will refer to this statistical approach as the ATM model.

2.3 Case Studies

The case study developed in this paper to illustrate the use of verification tools concerns recent medium-term hydrological forecasts produced daily at Hydro-Québec for the Manicouagan watershed, a major hydropower system. As illustrated in Fig. 2, the Manicouagan watershed is subdivided into five subcatchments for which forecasts are produced every day: Petit Lac Manicouagan, Manic-5, Manic-3, Toulnustouc, and Manic-2. The Manicouagan watershed is located in northeast of the province of Quebec, Canada. This water resource system consists of two hydropower plants with reservoirs in parallel (Manic-5 and Toulnustouc) and three downstream run-of-river hydropower plants (Manic-3, Manic-2, and Manic-1). The



Fig. 2 Manicouagan watersheds

total installed capacity is 6202 MW, which is about 17% of HQ's total capacity. In operating the system, generation planners face a variety of decisional problems. Two of those, common to every installation, are safety and the respect of environmental laws and regulations. For the two upstream watersheds, with large reservoirs, the other main concerns are those of long-term energy planning and optimization and

efficient releases for the operation of the run-of-river plants, given the inflows on those subbasins. For the three run-of-river plants, the issue is an efficient scheduling, given the inflows on the watersheds and the upstream releases. It is quite clear that the future state of inflows plays a major role on the decisions that will be made and the efficiency of the operations. The need for probabilistic hydrological forecasts is thus self-evident.

Our illustration of the assessment of the hydrological forecast value is specific to the statistical long-term MJJ spring prediction for the Churchill Falls basin where a single power plant, not owed by Hydro-Québec, is installed (Fig. 3). The total capacity is 5 428 MW (15% of HQ's total capacity). This watershed served as a test basin to explore the potential use of atmospheric circulation variables for making seasonal streamflow forecasts for high-latitude basins on the Québec-Labrador peninsula.

Given the nature of the forecasts presented in this section, how can we adequately evaluate probabilistic medium- to long-range hydrological forecasts? How can we compare and rank competing hydrological forecasting approaches? In the following sections, we will try to propose a framework that tackles the operational challenges raised here which ensue from the way the forecasts are being constructed and used to optimize electricity production. We will enlighten the issues such as the unknown forecasts and observations distributions, metric choice, uncertainty in score evaluation, multivariate verification, and the communication of verification information. Note that probabilistic forecasts based on historical hydrographs, denoted by HIST, are used as baseline forecasts to evaluate the performance of HQ medium- to long-range hydrological predictions, in particular when using skill scores.

3 Assessing Hydrological Probabilistic Forecasts: Statistical Framework

We employ the simple theoretical framework proposed by Gneiting et al. (2007) to formalize the problem of assessing hydrological probabilistic forecasts quality. At time t nature chooses a probability distribution G_t , which is considered as the true inflow generating process. On the other hand, the “forecaster” produces a hydrological forecast in the form of a predictive distribution denoted F_t or a sample from F_t (ensemble forecasts). Of course, the “true” distribution G_t is in practice never known but an observation x_t drawn from it is available. Figure 4 illustrates the framework proposed by Gneiting et al. (2007).

As mentioned by the authors, the diagnostic approach for such a forecast faces a challenge, in that the forecasts take the form of probability distributions, whereas the observations are real valued. The evaluation of probabilistic hydrological forecasts thus needs particular types of diagnostic tools. Of course, it is impossible to evaluate a single probabilistic forecast, and the size T of the forecast archives plays a major role in determining the significance of the result.

Before considering verification tools such as scoring rules, we must formally define what can be considered as a good probabilistic forecast. Gneiting et al. (2007) assert that the goal of probabilistic forecasting is to “maximize the sharpness of the

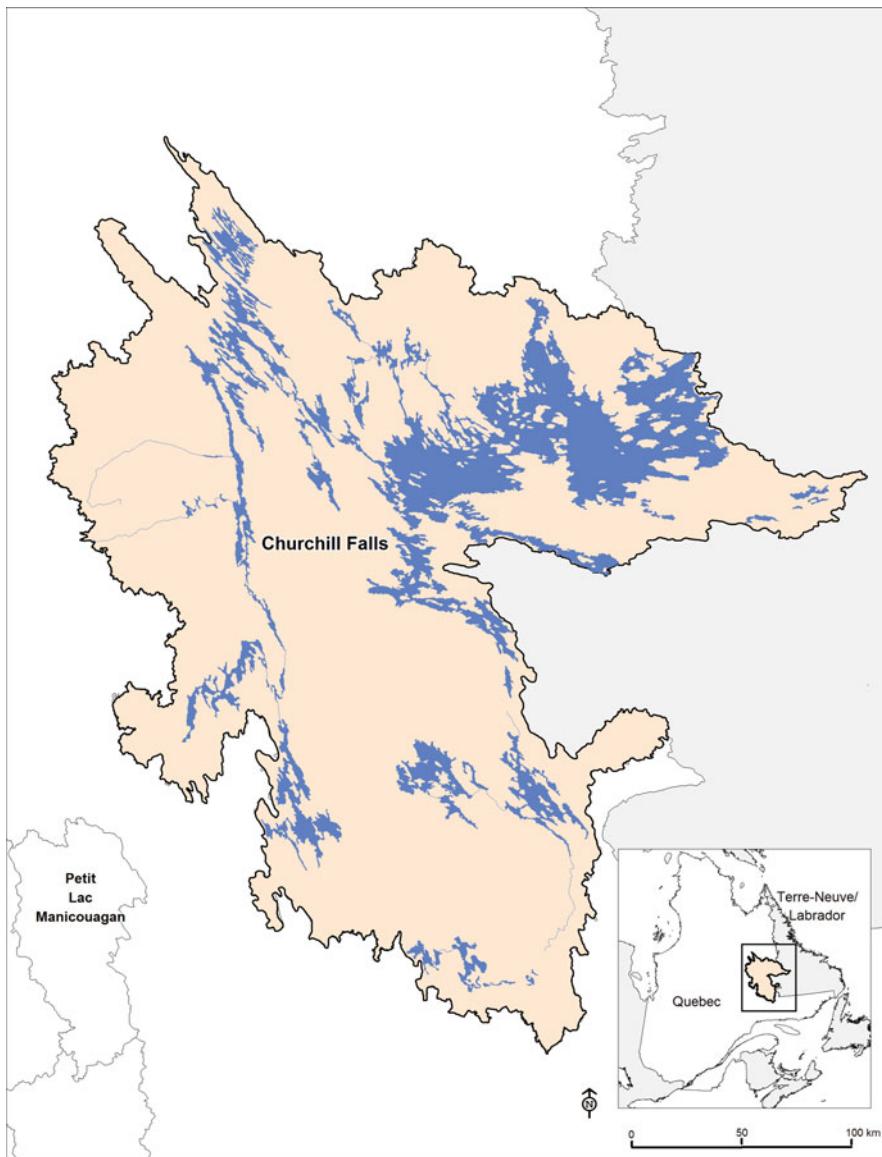


Fig. 3 Churchill Falls watershed

predictive distributions subject to calibration.” Here, calibration refers to the statistical consistency between the predictive distribution and the observation, while sharpness refers to the concentration of the predictive distribution. The sharper the forecast, the higher the information value it will provide if, of course, it is reliable, i.e., well calibrated.

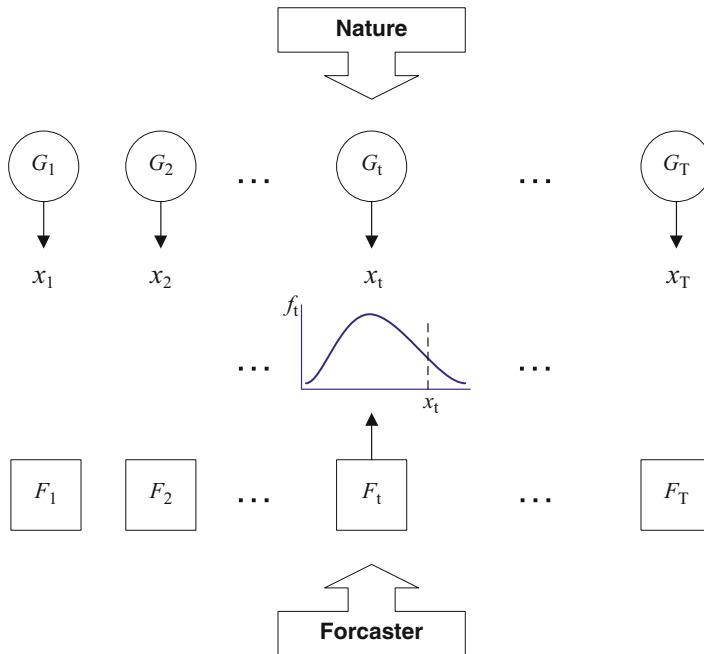


Fig. 4 Framework proposed by Gneiting et al. (2007)

Scoring rules, denoted here by $S(F,x)$, address calibration and sharpness simultaneously. Therefore, they are interesting summary measures of predictive performance. Of course, only proper scores must be considered to evaluate probabilistic forecasts. To avoid hedging, the forecaster's probability assignment to a future inflow event should be completely independent of the particular verification system. This is guaranteed if the scoring rules in use are proper. In short, a proper score will always prefer a probabilistic forecast if it is, in fact, more accurate (Brocke and Smith 2007).

In decision theory, finance, and meteorology, scoring rules have been used intensively for many years to evaluate probabilistic forecasts. A number of scores have been proposed in these domains. However, verification of probabilistic forecasts using formal scoring rules and sophisticated diagnostic tools has drawn the attention of hydrologists quite recently. Nevertheless, the use of standard scoring rules is becoming widespread in hydrology, for instance, in hydropower companies where daily inflow forecast is crucial to optimize planning of electricity production capacity (Électricité de France, BC-Hydro, Hydro-Québec, etc.; see Weber et al. 2006; Perreault 2013). Besides deterministic scores such as the Nash-Sutcliffe coefficient and graphical diagnostic tools such as the reliability diagram, the CRPS is probably the most used scoring rule. It is often the only score suited to evaluate probabilistic forecasts that is considered in number of verification experiments (see, for instance, Alfieri et al. 2014; Brown et al. 2014). The fact that the CRPS is

Table 1 Mathematical definition of scoring rules with their respective loss function

Scoring rules	Mathematical expression	Behavior
Logarithmic	$S(F_t, x_t) = -\log[f_t(x_t)]$	
Quadratic	$S(F_t, x_t) = \int_{-\infty}^{\infty} [f_t(y)]^2 dy - 2f_t(x_t)$	
Spherical	$S(F_t, x_t) = \frac{-f_t(x_t)}{\left[\int_{-\infty}^{\infty} [f_t(y)]^2 dy \right]^{1/2}}$	
CRPS	$S(F_t, x_t) = \int [F(y) - \mathbb{I}_{[x_t, +\infty)}(y)]^2 dy$	
Quantile	$S(q_t, x_t) = (x_t - q_t)(\mathbb{I}_{(-\infty, x_t]}(qt) - \alpha)$	

reported in the units of the observations and that it reduces to the mean absolute error if F is a deterministic forecast, thereby allowing a direct comparison between probabilistic and deterministic forecasts, may explain its popularity. However, the hydrologists should not limit their analysis to the use of a single score. A verification analysis could benefit from additional scoring rules. In what follows, we briefly present some scores that, to our knowledge, are rarely considered in the evaluation of medium- to long-range hydrological forecasts.

Mathematical definitions of several scoring rules in use at Hydro-Québec for univariate continuous predictive distributions are presented in Table 1: the logarithmic (ignorance score), quadratic, spherical, and CRPS scores. The quadratic (Selten 1998) and spherical (Jose 2009) scoring rules are currently used in finance and economics. Quantile scoring rules, which are rarely considered in hydrology, are also presented in Table 1. We find this type of scoring rules very useful since

some hydraulic planning models use as input not all the predictive distribution but often predictive quantile estimates. The expression of the simplest quantile scoring rule is given. As mentioned by Gneiting et al. (2007), the econometric literature refers to this scoring rule as the tick or the check loss function. This score is also well known for its application in quantile regression.

Note that f_t stands here for the predictive density function and $\mathbb{I}_A(y)$ is the well-known indicator function defined as

$$\mathbb{I}_A(y) = \begin{cases} 0 & \text{if } y \notin A \\ 1 & \text{if } y \in A \end{cases}. \quad (1)$$

To illustrate how these scoring rules differ, we present in Table 1 a graph which shows their behavior for different possible values of the observation x_t , assuming a normal predictive density function.

The logarithmic score penalizes the “forecaster” very severely when the observation x_t is situated in the extremities of the predictive density. This is very clear in the corresponding graph of Table 1, where the value of the score quickly tends to infinity when the observation is far from the center of the predictive distribution. The quadratic and spherical score functions are bounded. Selten (1998) recommends that the quadratic score should, for that reason, be favored compared with the logarithmic score. According to this author, the latter would be too sensitive to the situations where the observation is situated far in the tails of the predictive density. However, that remains debatable. In fact, the rejection of a scoring rule should take into account not only its mathematical behavior but also the decisional contexts for which forecasts are produced. In some critical situations, namely, face to the possible occurrence of harmful extreme events, we believe that severe scoring rules which are sensitive to the shape of the predictive distribution must be included in a verification system. In a recent comparative study with quadratic and spherical scores, Bickel (2007) finds several advantages to use the logarithmic score and other sensitive scoring rules. Finally, the CRPS is known to be less sensitive to the shape of predictive densities than many other scoring rules. Its expression is in fact related to the so-called robust M-estimators (Huber 1981). This “robustness” property is often viewed as an advantage. To our view, this may be questionable in our decisional context.

As mentioned, the scoring rules presented in Table 1 are designed to evaluate a univariate predictive distribution. However, good hydrological forecasts should reproduce spatial coherence for basins within a given region. Preserving the spatial dependence structure of inflows becomes crucial when the hydrological forecasts are used to plan hydroelectric production of a large set of sites such as the park of equipment of Hydro-Québec (more than 90 sites). Verifying the performance marginally (i.e., only the forecasts issued for each individual site) is then clearly not enough. This issue becomes very important since statistical post-processing of hydrometeorological probabilistic forecasts is gaining popularity. Do these statistical approaches preserve the spatial coherence and the dependence structure between

variables? How can we adequately evaluate a multivariate probabilistic forecast (here a multisite hydrological forecast)? One option is to use a multidimensional generalization of the CRPS, the energy score (Gneiting et al. 2008). Given an ensemble forecasts $\mathbf{y}_1, \dots, \mathbf{y}_n$ for a vector-valued quantity that takes values in \Re^d and the corresponding observation \mathbf{x} (for instance, ensemble hydrological forecasts for d basins), then the nonparametric estimator of the energy score proposed by Gneiting et al. (2008) is given by

$$S(F, \mathbf{x}) = \frac{1}{n} \sum_{j=1}^n \|\mathbf{y}_j - \mathbf{x}\| - \frac{1}{2n^2} \sum_{i=1}^n \sum_{j=1}^n \|\mathbf{y}_i - \mathbf{y}_j\|, \quad (2)$$

where $\|\cdot\|$ denotes the Euclidian norm.

The energy score is proper. In a multivariate setting, this property of a scoring rule gains extra meaning. Indeed a multivariate distribution is composed of marginal distributions tied together by a dependence structure. This means that a proper multivariate score, such as the energy score, will reward a forecast in which the marginal distributions are well predicted individually as well as in relation to each other.

To illustrate this, we conduct a Monte Carlo experiment with the energy score on copulas, which describes the dependence structure exactly with marginal effects eliminated (Genest and Favre 2007). We use two different predictions and compare the distributions of the average scores for bivariate Clayton copula observations. One prediction takes into account the dependence structure (DEP), while the other treats the marginal effects as independent via the independence copula (IND). This way both forecasts perfectly predict the marginal distributions which are uniform, but one wrongly assumes that the two variables are independent, while the other one uses the correct copula. As we can see in Fig. 5, the distribution of the average energy score is lower for the forecasts that picked the right dependence structure. The difference is more significant as the Kendall rank correlation coefficient (commonly referred to as Kendall's tau coefficient) increases. Note that it has been pointed out recently by Pinson and Tastu (2013) that the ability of the energy score to detect incorrectly specified correlations can be limited. Scheuerer and Hamill (2015) have then proposed an alternative multivariate family of scoring rules based upon the concept of variogram which are “more discriminative with respect to the correlation structure.”

This simple experiment demonstrates that a multivariate forecast is not simply the sum of its univariate parts. In meteorology and hydrology, different physical measures have strong dependence structures which a forecast should accurately recreate. The use of proper multivariate scores such as the energy score or the more recent variogram-based scoring rules is recommended as it penalizes incoherent predictions in the marginal distributions as well as misrepresentations of the dependence structure. As mentioned earlier such scoring rules can also be used to verify that statistical post-processing of hydrometeorological forecasts does not alter the relationship between different variables it is affecting.

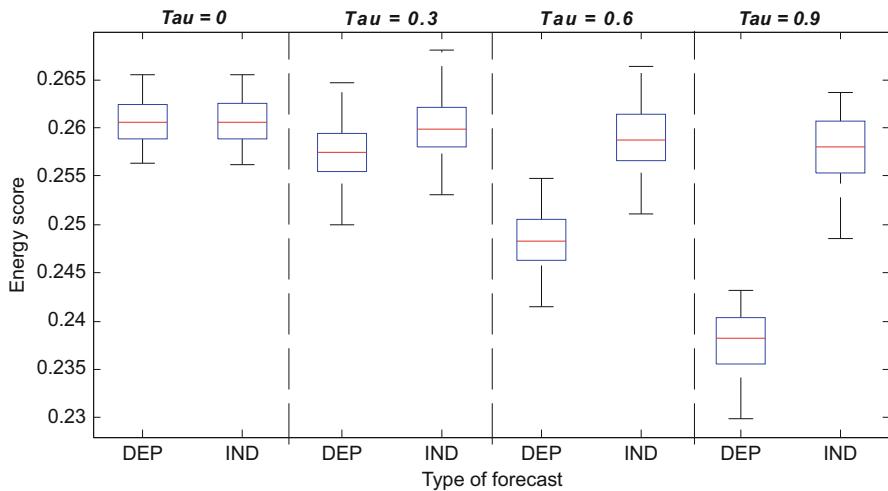


Fig. 5 Distributions of average energy scores obtained with an independence assumption (*IND*) and with the correct dependence structure (*DEP*) of a bivariate Clayton copula, as Kendall's tau varies. Kendall's tau is a measure of association between variables

4 Estimation of Scores

Very seldom do hydrological forecasts actually consist of an explicitly defined predictive probability distribution function. Instead we might produce different scenarios loosely based on quantiles or deterministic forecasts whose initial values are perturbed with random noise. This leaves us with several options to estimate a given score of a forecast. For instance, shall we consider a parametric approach or a more robust nonparametric method to estimate the CRPS? These approaches can lead to different estimated score values and therefore may have an impact on the following conclusions made about the forecast system. This issue becomes very important when the forecasting process produces predictive distributions with various shapes such as those presented in Fig. 1.

To illustrate this problem, we performed a simulation study aimed at comparing different parametric and nonparametric estimation methods for the CRPS. We pick nature's probability distribution G and produce a sample y_1, \dots, y_n as well as an observation x from that distribution. The true score, which is the best one on average, can be calculated and compared to an estimator based on the ensemble y_1, \dots, y_n . For the parametric methods, we study the effect of misjudging the real probability distribution G in different cases such as asymmetry or overdispersion. The nonparametric estimators do not suffer from that issue but might not be appropriate in certain situations, notably for small sample sizes.

In this study we focus on comparing different nonparametric estimators $S_{np,k}$ of the CRPS as well as a parametric estimator S_p numerically computed using a Monte

Carlo approximation (Robert and Casella 2000). Note that Taillardat et al. (2016) have given theoretical and analytical formulas for the CRPS for several probability distributions.

The first nonparametric estimator is based on the representation of the CRPS with expectations (Gneiting and Raftery 2007) and has the following simple expression:

$$S_{np,1}(F, x) = \frac{1}{m} \sum_{i=1}^m |y_i - x| - \frac{1}{2m} \sum_{i=1}^m |y_i - y'_i|, \quad (3)$$

where $\mathbf{y} = (y_1, \dots, y_m)$ and $\mathbf{y}' = (y'_1, \dots, y'_m)$ denote two independent samples of size m from the predictive distribution F . In practice, we have to split the ensemble y_1, \dots, y_n into two subsamples of equal length. This estimator was used, namely, by Friederichs and Thorarinsdottir (2012).

We also propose to use another nonparametric estimator which appears natural when considering the integral representation of the CRPS, where we swap the predictive distribution function F for the empirical distribution function F_n :

$$S_{np,2}(F, x) = \int [F_n(z) - \mathbb{I}_{[x, +\infty)}(z)]^2 dz, \quad (4)$$

where

$$F_n(z) = \frac{1}{n} \sum_{i=1}^n \mathbb{I}_{(-\infty, z]}(y_i). \quad (5)$$

The third nonparametric estimator from Naveau et al. (2014) is based on U-statistics and has the following form:

$$S_{np,3}(F, x) = x + \frac{1}{n} \sum_{i=1}^{2n} (y_i - x) \mathbb{I}_{(-\infty, y_i)}(x) - 2\hat{\mu}_n, \quad (6)$$

where

$$\hat{\mu}_n = \frac{1}{2n(2n-1)} \sum_{1 \leq i < j \leq 2n} \max(y_i, y_j) \quad (7)$$

Figure 6 presents the RMSE of the estimators presented above in a few selected cases described in Table 2. These figures illustrate the relative performance in the case of a “perfect” prediction and over- and underdispersion as well as a case of asymmetrical law predicted to be symmetric as it might be the case with inflow or rainfall, for example.

We can observe that if the sample size is large enough, the nonparametric estimators are not significantly worse than the parametric estimators if the predictive family is correctly chosen (here a normal), except for $S_{np,1}$ (Fig. 6a). In the case of a

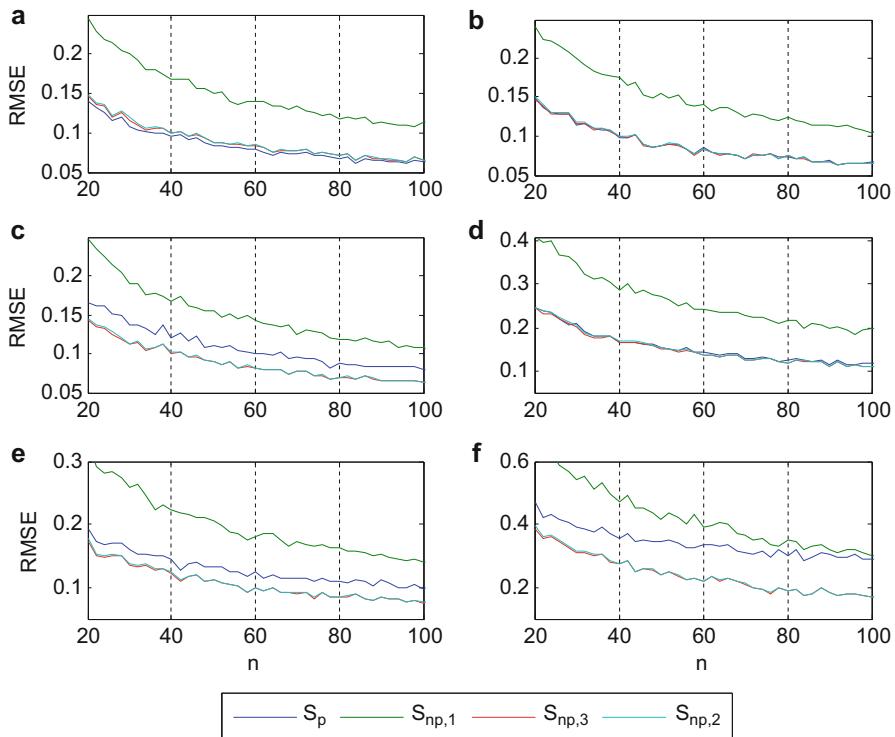


Fig. 6 RMSE of various parametric (S_p) and nonparametric ($S_{np,1}, S_{np,2}, S_{np,3}$) estimators of the CRPS for the different predictive situations of Table 2 using ensembles

Table 2 Selected cases for the comparison of three estimators for the CRPS

Selected cases	Nature G	Predictive F	Comment
a	Normal	Normal	Ideal forecast
b	Normal	Logistic	Overdispersed predictive density
c	Normal	Laplace	Overdispersed predictive density
d	Logistic	Normal	Underdispersed predictive density
e	Laplace	Normal	Underdispersed predictive density
f	Gamma	Normal	Wrong (symmetric) predictive density

drastic misjudged probability distribution (Laplace vs. normal or gamma vs. normal), the nonparametric estimators $S_{np,2}$ and $S_{np,3}$ are clearly better choices as expected (Fig. 6c, e, f). This makes these estimators more reliable if the choice of a predictive family is difficult as it is the case for midterm hydrological forecasts produced at Hydro-Québec. Note that estimator $S_{np,1}$ used by Friederichs and Thorarinsdottir (2012) consistently underperforms. This is not surprising since the sample has to be split in two independent subsamples.

We therefore recommend to consider either the nonparametric methods $S_{np,2}$ or $S_{np,3}$ to evaluate the CRPS. In the following section, estimator $S_{np,3}$ developed by Naveau et al. (2014) is used.

It is important to point out that the results obtained in this section are not particular to medium- to long-range forecasts nor they are to hydrological data. Our conclusions and recommendations are valid for any forecasted univariate random variable.

5 Medium-Term Hydrological Forecasts Verification: Application to Real Data

Hydro-Québec is currently proceeding to an intensive verification experiment of its hydrometeorological forecasts archive. The use of verification scoring rules is illustrated with recent medium-term hydrological forecasts produced. We concentrate on the Manicouagan watershed, a major hydropower system (see Fig. 2). As an illustration, Fig. 7 presents 30-day hydrological forecasts produced on the 25th of April 2013 for the Manic-5 subbasin. We can observe in particular that the dispersion of the hydrological forecasts for the first 10 days is quite low. Also, multimodal predictive distributions seem to emerge after a 5-day forecast.

Figures 8, 9, 10, 11, and 12 illustrate the performance of HQ's 2013 medium-term hydrological forecasts for the Manic-5 watershed for the overall year. Forecasts are available for 233 days. The volumes considered a range from 1 to 30 days of cumulated outputs. PIT histograms and average univariate scores are presented for HQ's predictions and also for the probabilistic forecast based on historical data

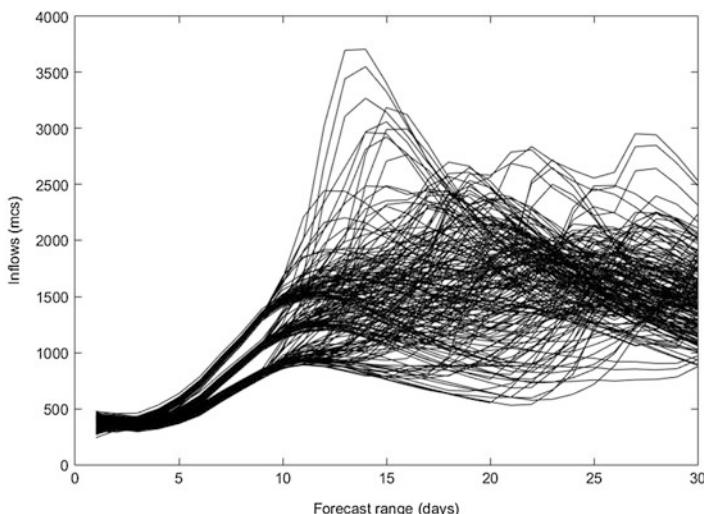


Fig. 7 Hydro-Québec's forecasts produced on the 25th of April 2013 for Manic-5 watershed

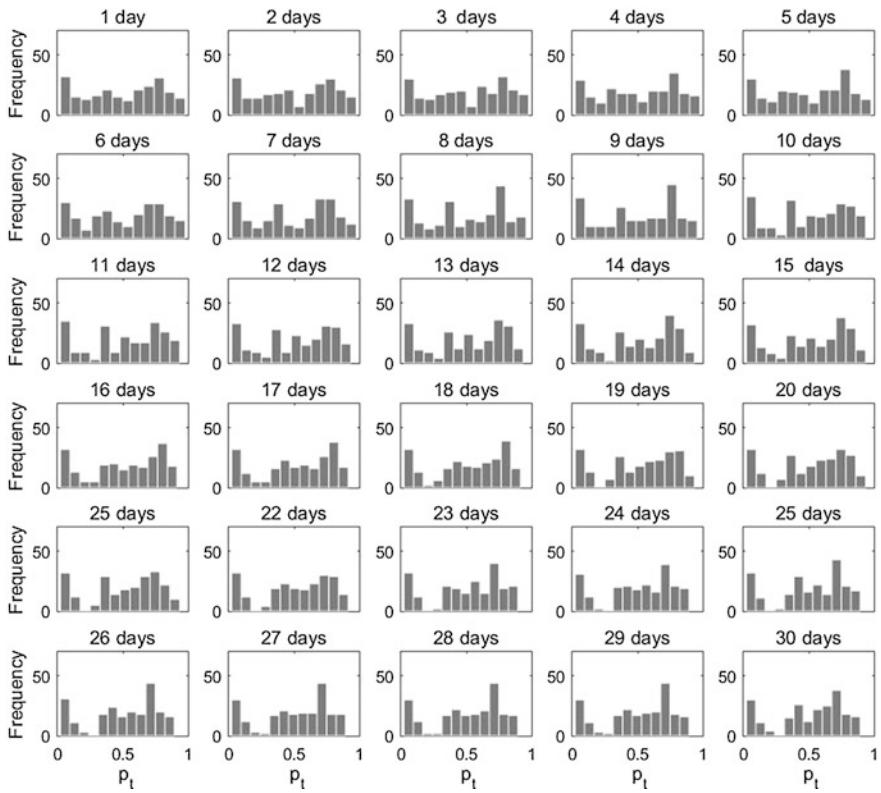


Fig. 8 PIT histogram based on the overall year for 1–30-day forecasted volumes produced in 2013 for Manic-5 watershed using historical inflows (HIST)

(HIST). The scoring rules considered are the CRPS, logarithmic, quadratic, and spherical scores (see Table 1). The scoring rules and the PIT values were estimated using nonparametric approaches (as mentioned in the previous section, $S_{np,3}$ was used to estimate the CRPS).

The PIT histograms for the overall year are first presented in Figs. 8 and 9, respectively, for HIST forecasts and HQ's predictions.

The HIST forecast seems well enough calibrated. However, we would expect a better result (more uniform PIT histograms). It is probably due to the fact that the year 2013 is quite particular in terms of hydrological behavior. In fact, its hydrograph is different from the one observed historically (see Fig. 10). The year 2013 distinguishes itself more particularly by a hasty spring flood and low summer and fall inflows.

The PIT histograms for HQ's predictions (Fig. 9) suggest unbiased but strongly underdispersed 1–7-day forecasts. This is consistent with Fig. 7 where low dispersion can be observed for the prediction of the first 10 days. As the forecast horizon increases, it becomes biased: the inflow is systematically underestimated.

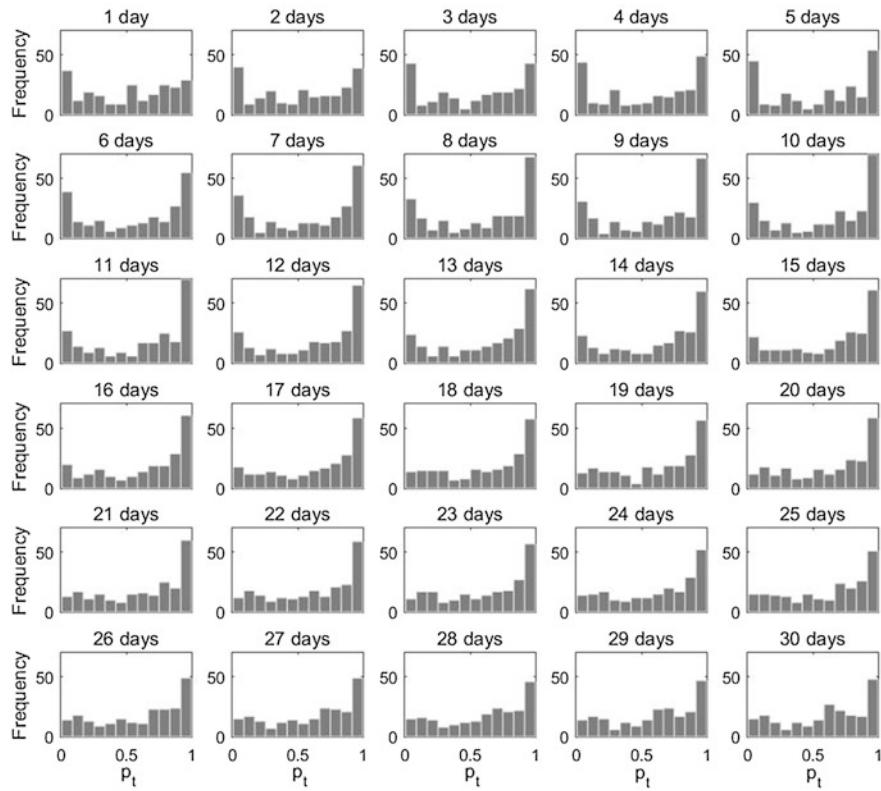


Fig. 9 PIT histogram based on the overall year for 1–30-day forecasted volumes produced in 2013 for Manic-5 watershed using Hydro-Québec’s forecasting system

The performance of HQ’s predictions for watershed Manic-5 measured with the four scoring rules averaged for the overall year is presented in Fig. 11. The values are reported as skill scores with respect to the HIST forecast. Note that the comparison of two competitive forecast systems based on skill scores (using one forecast set as a baseline) is much more straightforward than the visual comparison of the PIT histograms presented earlier.

According to the CRPS and the spherical score, HQ’s inflow forecasts outperform the historical approach (HIST forecast) for all forecast horizons even though they are biased and underdispersed (see PIT histograms of Fig. 9). The CRPS shows a clear advantage for HQ’s predictions (from 40% to 75%). The spherical score is much less generous (from 10% to 30%). The logarithmic and quadratic scoring rules are severe and more sensitive to the very inaccurate forecasts. These scoring rules bring a different perspective to the verification analysis. Unlike the CRPS and the spherical score, these measures give preference to the HIST forecast for certain ranges. The logarithmic score indicates a “skill” for HQ’s prediction which seems limited to the first 6 days and suggests that the HIST forecast should have been used for 2013

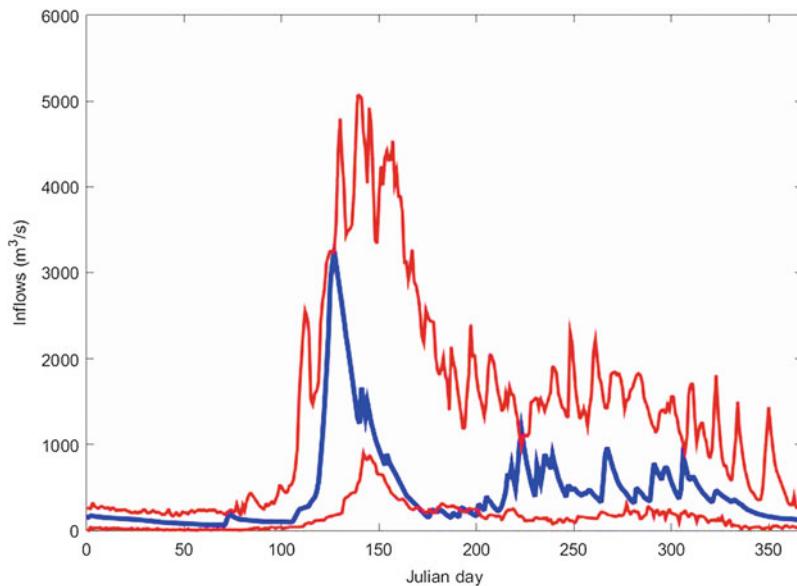


Fig. 10 Year 2013 for Manic-5 (hydrograph in blue) compared to past historical years (daily minimum and maximum in red). Year 2013 distinguishes itself more particularly by a hasty spring flood and low summer and fall inflows

beyond this range. This result is very useful for the forecasters whom can retrospectively characterize the hydrological events that may have been inaccurately forecasted in 2013. This could not have been done if only the CRPS have been used.

The year was then split into four according to the seasonality of the volumes of output water and based upon the expertise of HQ's hydrologists. The performance of HQ's predictions measured with the four scoring rules averaged for the spring and fall seasons are presented, respectively, in Figs. 12 and 13. HQ's forecasts are available for only 60 days for each season. As for the overall year, the values are reported as skill scores with respect to the HIST forecast.

According to all the four scoring rules, HQ's hydrological forecasting system outperforms systematically the HIST prediction in spring (Fig. 12). The CRPS is the least severe showing skill score values between 40% and 80%. It is in the fall that HQ's forecasting system has performed poorly in 2013 for Manic-5 watershed (Fig. 13). This can be observed for all of the four scoring rules.

Since the sample size can be very limited when the scores are averaged over a season (only 60 forecasts in 2013 for spring and fall), it is essential to take into account the uncertainty in the estimation of scoring rules. A standard bootstrap technique was used here to construct confidence intervals for each expected score. The graphs of Fig. 14 present the average CRPS as well as 90% confidence intervals for each season of 2013. Note the fact that the CRPS has the same magnitude as the studied variable explains the increasing trends in the graph as the volumes are cumulated over more days. Also note that the uncertainty increase with the number

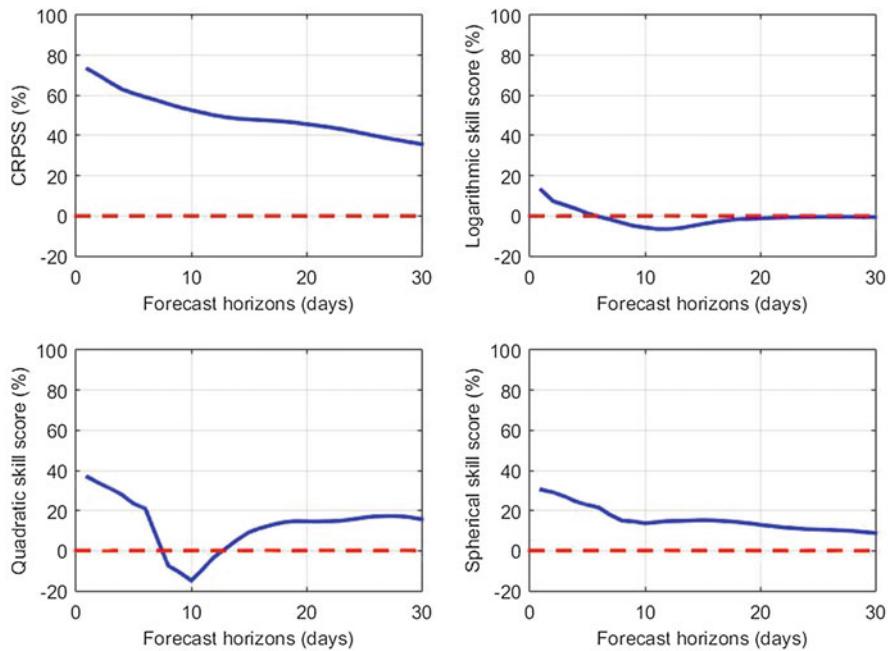


Fig. 11 CRPSS, logarithmic, quadratic, and spherical skill scores obtained for 1–30-day forecasted volumes produced in 2013 for the *overall year* of Manic-5 watershed using Hydro-Québec's forecasting system and historical inflows (HIST) as the reference forecasts

of cumulated volumes is explained by the increased variation due to summing as well as the decrease in number of predictions available, as it was chosen to not spill over into other seasons. For this particular year, one can conclude that Hydro-Québec's predictions performed significantly better than a probabilistic historical prediction in the summer for all lead times and in the spring for forecast horizons less than 25 days.

The next graphs present the results of a similar study conducted on the hydrological forecasts for the whole Manicouagan hydropower system which includes five interconnected drainage basins: Petit Lac Manicouagan, Manic-5, Manic-3, Toulnustouc, and Manic-2 (see Fig. 2). The energy score is used as a generalization of the CRPS to multivariate predictions. No transformation of the data is necessary since the Euclidean norm in the score function will naturally penalize the larger prediction errors in individual basins. This multivariate score allows to measure global performance of forecasts for systems with complicated dependence structures. In this example we can conclude that HQ's hydrological forecast for the five basins of the Manicouagan hydropower system is more effective globally than a probabilistic historical forecast in all seasons but fall for horizons longer than 10 days (see Figs. 15 and 16). As mentioned in Sect. 2, to extend the hydrological forecast beyond the meteorological forecast lead time (9 days), historical weather series are used as

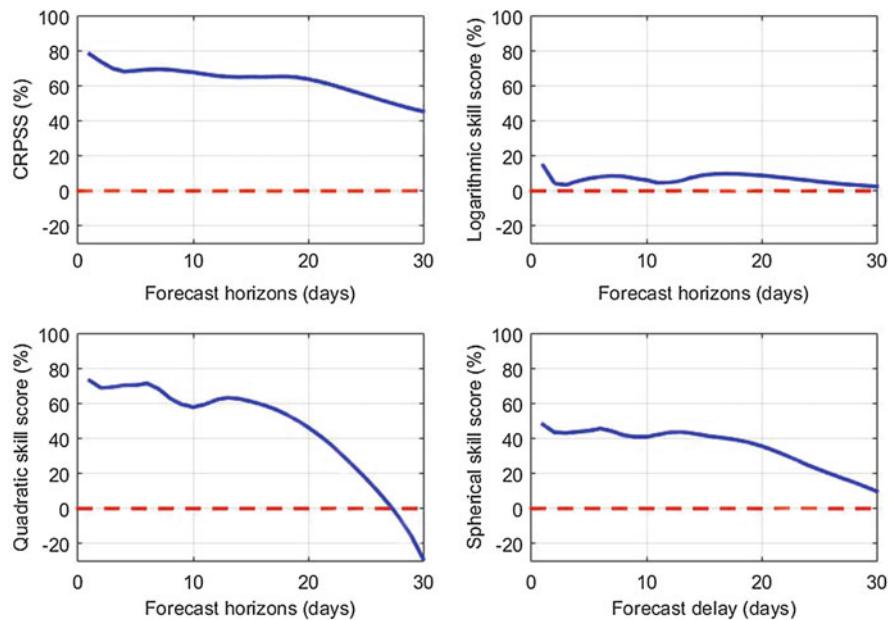


Fig. 12 CRPSS, logarithmic, quadratic, and spherical skill scores obtained for 1–30-day forecasted volumes produced in 2013 for the *spring season* of Manic-5 watershed using Hydro-Québec's forecasting system and historical inflows (HIST) as the reference forecasts

multiple possible inputs into a hydrological model. The model acts as a filter producing therefore underdispersed forecasts. This is particularly noticeable in the fall where strong autumn inflows are produced by convective heavy rain. This partly explain why HQ's forecasting system shows less skill in the fall compared to forecasts based upon historical hydrographs (HIST).

6 Using a Production Planning Model to Evaluate Long-Term Hydrological Forecast Value

Statistical scoring rules and graphical tools are essential for a detailed evaluation of the strengths and the weaknesses of a hydrological forecast system. They allow to measure the relative importance of the various components and to identify which one needs to be improved. Sophisticated statistical verification analyses such as the one presented in the previous sections are particularly useful for hydrologists and scientific decisions. But, what about the decision-makers who are usually not familiar with quantities such as the CRPS? Even if considerable efforts are granted to communicate the results of a verification experiment, it may still be useless for water planners. Therefore, it could be worth it to evaluate hydrological forecasts in a more decisional perspective.

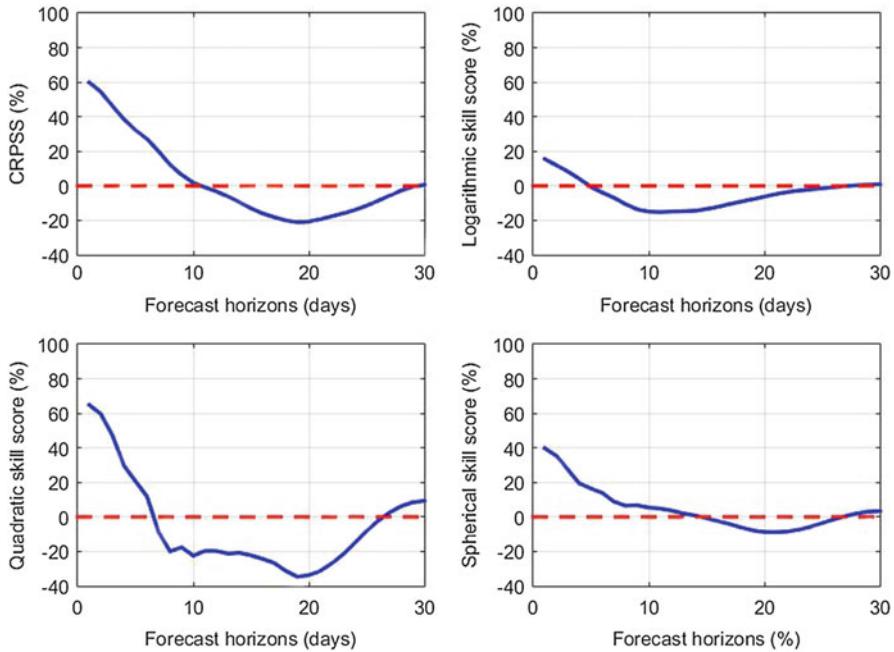


Fig. 13 CRPSS, logarithmic, quadratic, and spherical skill scores obtained for 1–30-day forecasted volumes produced in 2013 for the *fall season* of Manic-5 watershed using Hydro-Québec's forecasting system and historical inflows (HIST) as the reference forecasts

This section takes one step further by not only considering statistical verification scoring rules, but by employing hydrological forecasts in a stochastic midterm planning model designed for optimizing electricity production. This will help to evaluate the validity of economic models as verification tools. In this experiment, we consider the specific application of the evaluation of long-term forecast value for the aggregated May–July (MJJ) spring streamflow produced in December using the ATM model (see Sect. 2.2). Again, the HIST forecasts are used as the baseline approach (probabilistic forecasts based on MMJ historical inflow observations). The performance of the ATM model is measured in terms of its ability to adaptively forecast one-step ahead the aggregated MJJ flow volume during the 1961–2004 forecast evaluation period. Figure 17 illustrates for the Churchill Falls watershed (see Sect. 2.3) the one-step-ahead adaptive forecast of aggregated MJJ flows for the 1961–2004 period using model ATM and the CRPSS (skill score) for each forecast. The one-step-ahead forecasts (in red) match the observed data (in blue) quite well. The average CRPSS obtained is 26% for this basin, indicating that model ATM improves the MJJ forecasts produced in December compared to model HIST.

The main objective of hydraulic production midterm planning is to determine for each week of a planning horizon of 1–2 years an optimal plan for the use and storage of water in system reservoirs while satisfying a known weekly demand in electricity.

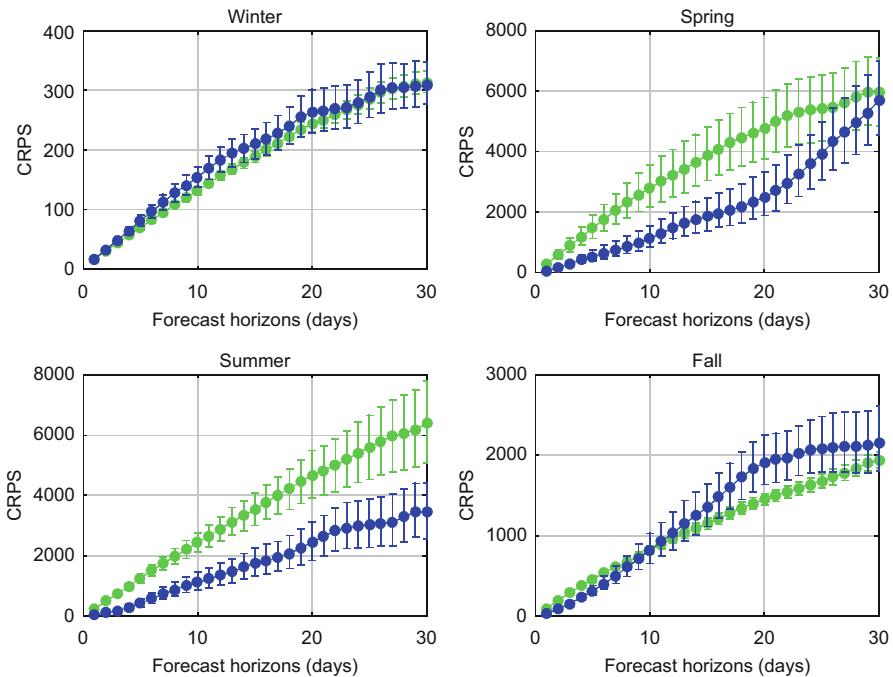


Fig. 14 CRPS scores along with a 90% confidence interval for 1–30-day forecasted volumes produced in 2013 for each season of Manic-5 watershed using Hydro-Québec’s forecasting system (in blue) and the HIST probabilistic forecast using historical observations (in green)

The maximization of water value in reservoirs at the end of the planning horizon is the main component of the cost function to minimize. This function also includes energy purchase costs and sale revenues, as well as thermal production costs. The main decision variables may be subdivided into two main categories. The first one is related to the hydraulic part of the system, e.g., reservoirs, hydraulic plants, spillways, and river segments (a segment is delimited by two successive reservoirs). The second category is related to a simplified energy transit network, composed of demand geographical zones, transit links, energy markets, and thermal plants. Each plant and market is associated with a given zone and hence contributes directly to its demand. Midterm planning is subject to constraints such as energy demand to fulfill in each zone, for each week, and bounds on water discharge. The “optimal” solution of the planning problem is the solution that respects all constraints while maximizing a cost function which includes the value of stored water in the reservoirs at the end of the horizon and energy sale revenues, thermal plant generation costs, and energy purchase costs.

In order to improve flood control, the planning model is coupled with an inflow scenario tree so that the stochastic nature of inflows is taken into account. This kind of modeling approach, which is based on stochastic programming theory (see Birge and Louveaux 1997), is an extension of deterministic production planning tools

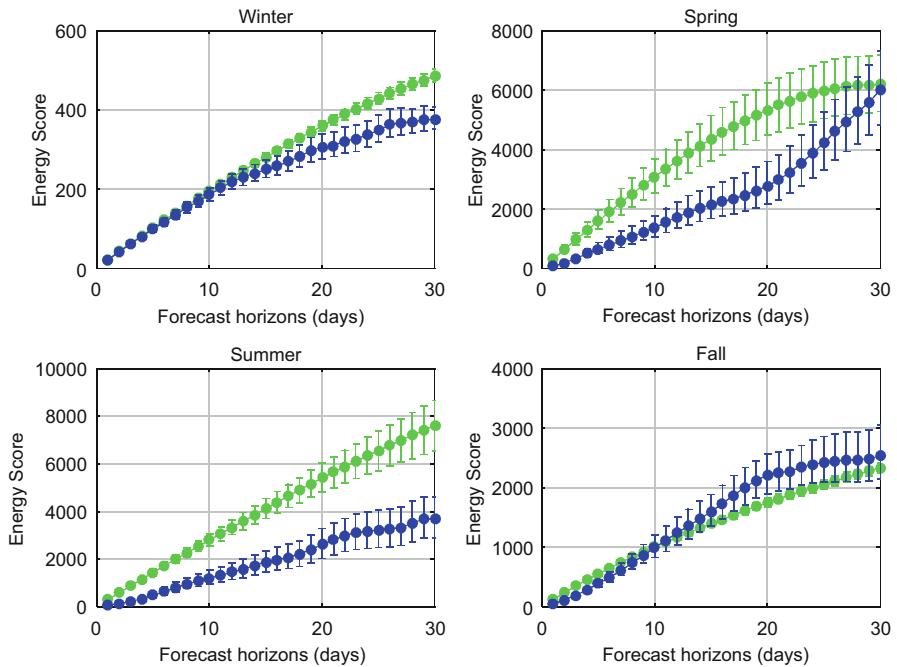


Fig. 15 Energy scores along with a 90% confidence interval for 1–30-day forecasted volumes produced in 2013 for each season of Manicouagan hydropower system using Hydro-Québec’s forecasting system (*in blue*) and the naïve probabilistic forecast using historical observations (*in green*)

already in use at Hydro-Québec. One of the main hypothesis underlying deterministic production planning models is the fact that natural inflows are supposed to be known over the entire planning horizon, which is a serious drawback of such models. In order to minimize the risk of obtaining bad solutions, a “mean” inflow scenario is often used, based on historical data which (hopefully) represent all the spectrum of possibilities. To handle explicitly several inflow scenarios simultaneously, an inflow tree sub-model can be integrated into deterministic models. Such a tree is illustrated in Fig. 18.

In this example, the tree, which covers the planning horizon from left to right, is composed of two parts, called stages. The first stage is represented as a unique branch: it is the deterministic part of the model. Only one inflow scenario is considered: it is assumed here that the planner knows pretty well in advance the inflows during the first weeks of the planning horizon. The second stage starts at the end of the first stage and covers the remaining of the horizon. It is represented as a sub-tree which has as many branches as there are inflow scenarios: it represents the stochastic part of the model. At each branch is associated a probability: the greater the probability for a given scenario, the more important the cost of the solution for the corresponding branch. The scenario tree is considered as whole in the solution

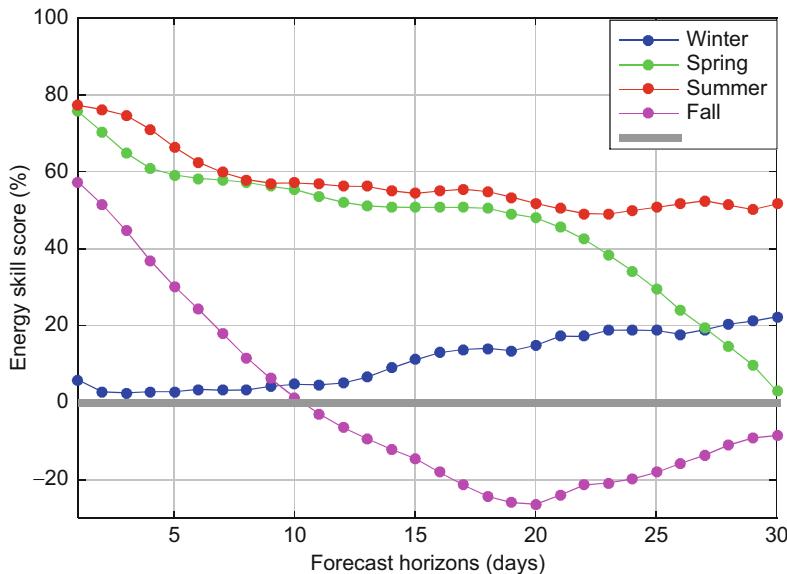


Fig. 16 Energy scores express in skill score for 1–30-day forecasted volumes produced in 2013 for each season of Manicouagan hydropower system using Hydro-Québec's forecasting system and historical inflows (HIST) as the reference forecasts

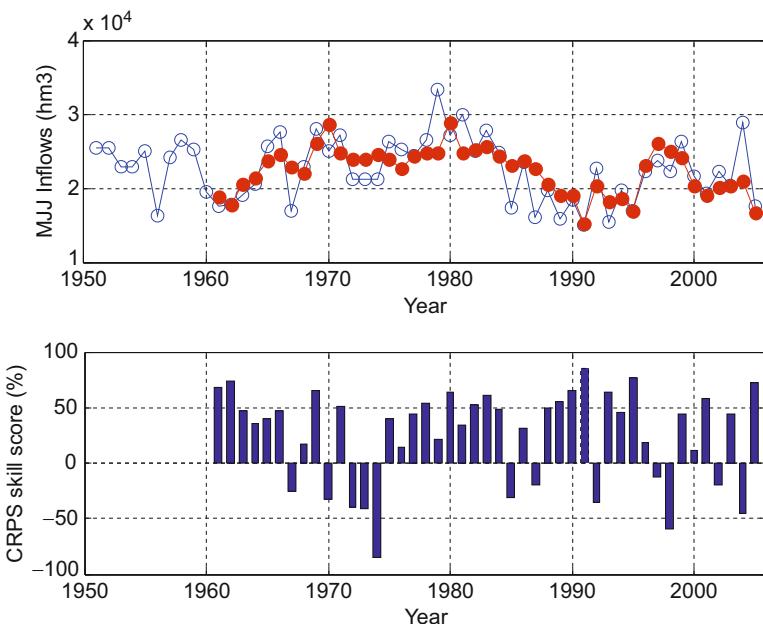


Fig. 17 One-step-ahead adaptive forecast of aggregated MJJ flows for the 1961–2000 period using model ATM (in red); CRPS (skill score) for each forecast compared to model HIST

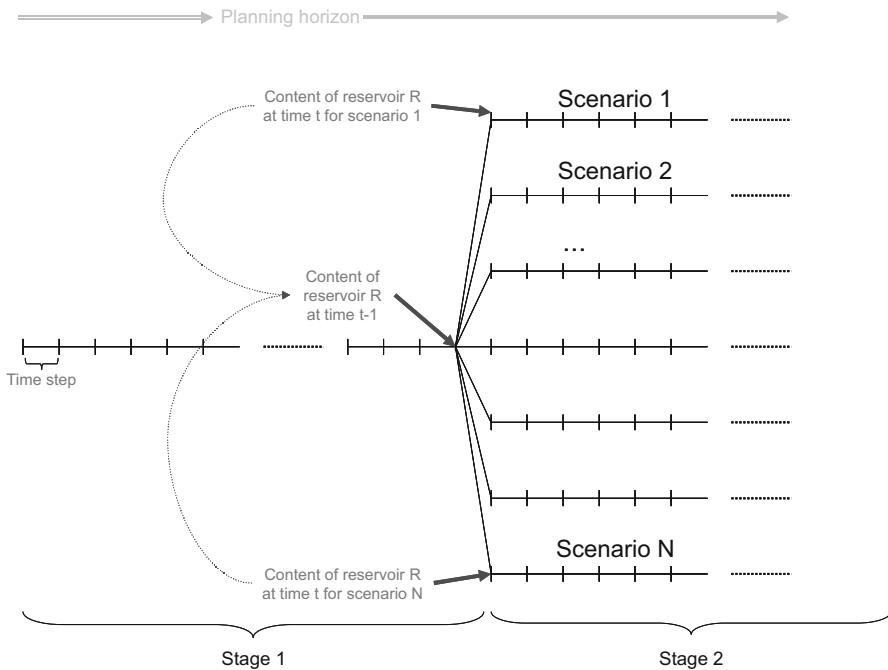


Fig. 18 Inflow scenario tree with two stages

process. Each branch defines a production planning subproblem, and all these subproblems depend on the state of the system at the end of the first stage: hence, they must be solved simultaneously. Thus, for example, the content of a given reservoir R at time t for scenario 1 (see Fig. 18) depends on its content at time t-1 (which is unique because it belongs to the first stage) via the reservoir hydraulic state equation. In the same way, its content at time t for scenario N depends on the content at time t-1 with the only difference being that the reservoir inflow value used in the hydraulic state equation is drawn from the Nth inflow scenario. The objective function of the model is then the sum of the objective function related to the first stage and the weighted sum (using the probabilities) of the objective functions related to the branches of the second stage.

The quality of the solution will depend on:

- The inflow scenarios: ideally they must represent the whole spectrum of the inflow expected.
- The probabilities: they should be derived from an inflow forecasting model.

Our calculation method focuses on the latter: in short, we would like to know if we can improve the solution while using more accurate inflow forecasting models, and if so, is this improvement significant?

To estimate the cost impact of using the ATM model to predict the total flood volume 5 months in advance (in December), we simulate the production midterm planning process. This simulation aims to show the use of a statistical forecasting model based on atmospheric variables to make better use of energy purchase and sale opportunities during the year, such as to avoid some costly purchases, to plan preventive purchases at minimal cost, and to benefit from additional sales during the following summer.

The simulation is based on the stochastic production midterm planning model presented above. The planning process takes place on the 1st of December as it is currently done at Hydro-Québec. The tree is composed of two stages with weekly time steps:

- The deterministic part (i.e., the first stage) begins at the end of November and ends just before the beginning of the flood period, which begins in May in the simulated area: note that in December, the planner has a good knowledge of the inflows during the winter season. For the purpose of our simulations, we could have reduced the length of this first stage, but since they are based on the current planning methodology used at Hydro-Québec, we have decided to use a 5-month first stage.
- The stochastic part (i.e., the second stage) starts at the beginning of the spring flood period and covers the remaining of the horizon, which ends at the end of October.

The inflow scenarios are derived from historical inflow data, which covers 54 years (from 1951 to 2004): 1-year inflow data defines one inflow scenario. It is assumed that the last 54 years are representative of the future hydrologic patterns, at least for the next 10 years. The planning process simulation is composed of two phases, called planning exercises, which correspond to two given “planning times” during the horizon. The first exercise is held at the end of November (the beginning of the planning horizon). Its goal is to determine the “optimal” level of reservoirs at the beginning of May and to settle the energy sales and purchases in winter (i.e., during each week of the first stage). The second exercise is held at the beginning of May, when the flood amplitude is fully predictable. It is assumed that the reservoir levels in May are those generated by the first planning exercise. Again here energy sales and purchases are generated for each week of the second stage, given a minimal bound on the energy total stock (in reservoirs) at the end of the planning horizon.

This simulation can be run for each inflow scenario (i.e., for each year):

- Given an inflow year, which is considered as the “planning year,” we must first generate a set of weights (i.e., probabilities) for each inflow scenarios (of the second stage) using an appropriate inflow forecasting model. The inflows for the first stage are those of the planning year.
- The first planning exercise then generates the initial conditions for the second exercise, for which the inflows become fully known during the second stage: they correspond to those of the planning year.

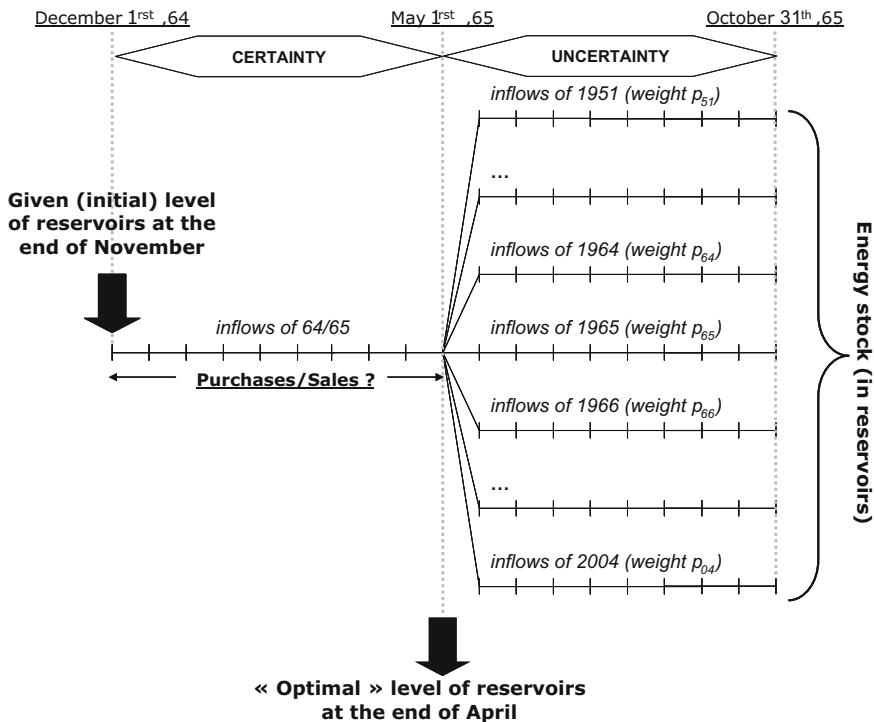


Fig. 19 The production planning process simulation for years 64–65: first planning

- For each simulation, the same operation bounds, market structures, energy demand, and reservoir initial levels (at the end of November) are used. The only differences are the inflow data.

As an illustration, the two next figures present the simulation for the two particular years 1964 and 1965 (Figs. 19 and 20).

To estimate the cost impact related to the use of forecasts based upon atmospheric variables (model ATM), two simulations are performed for each planning year:

- The “mean simulation or climatology simulation” results from the forecast obtained using only the historical MJJ spring forecasts (model HIST). We assume that all inflow scenarios have the same occurrence probability. They are then given the same weight for all historical scenarios.
- The “atmospheric simulation” takes into account the forecast obtained by the ATM statistical model. Historical scenarios are weighted to update the forecast HIST with respect to the ATM forecasts based upon atmospheric variables. These weights are based upon the ratio of the predictive densities of forecasts HIST and ATM (see Fig. 21). For instance, if the model ATM predicts small MJJ inflows, higher weights are given to the low-inflow historical scenarios. This approach is

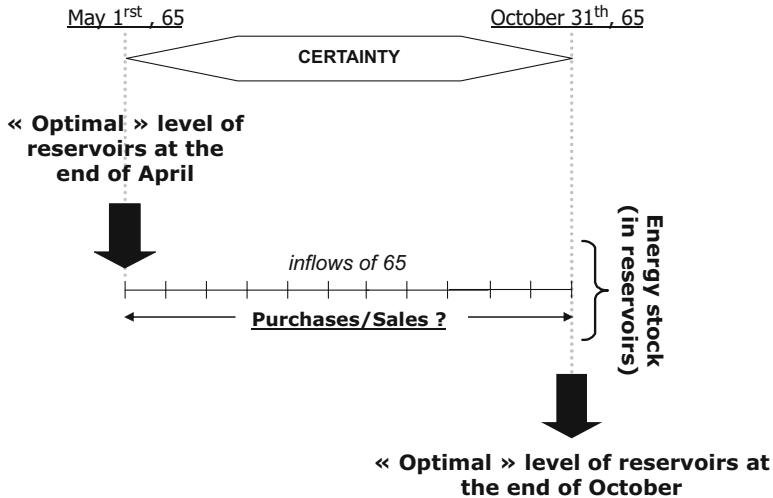


Fig. 20 The production planning process simulation for years 64–65: second planning exercise

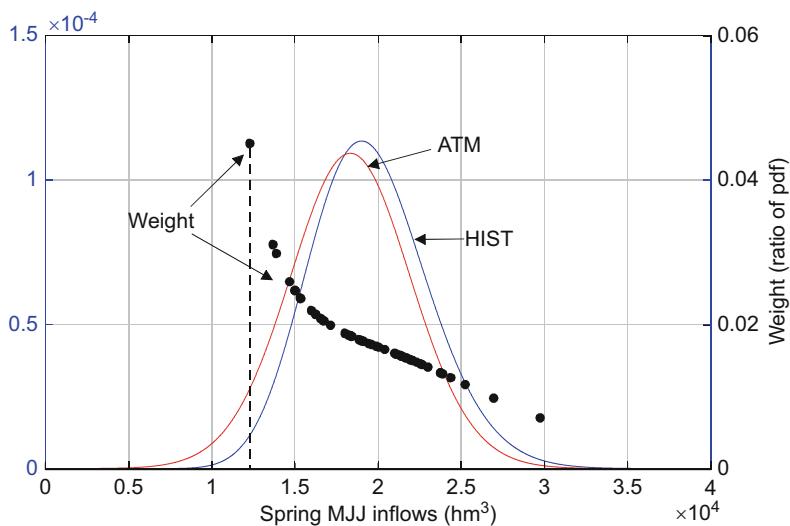


Fig. 21 Weights based upon the ratio of the predictive densities

inspired by the method of importance sampling (see Robert and Casella 2000, Chap. 3).

The difference between the objective function values obtained from each of the two simulations for a given planning year gives the “net gain” implied by the use of an inflow forecasting model based on atmospheric variables. Note that for certain

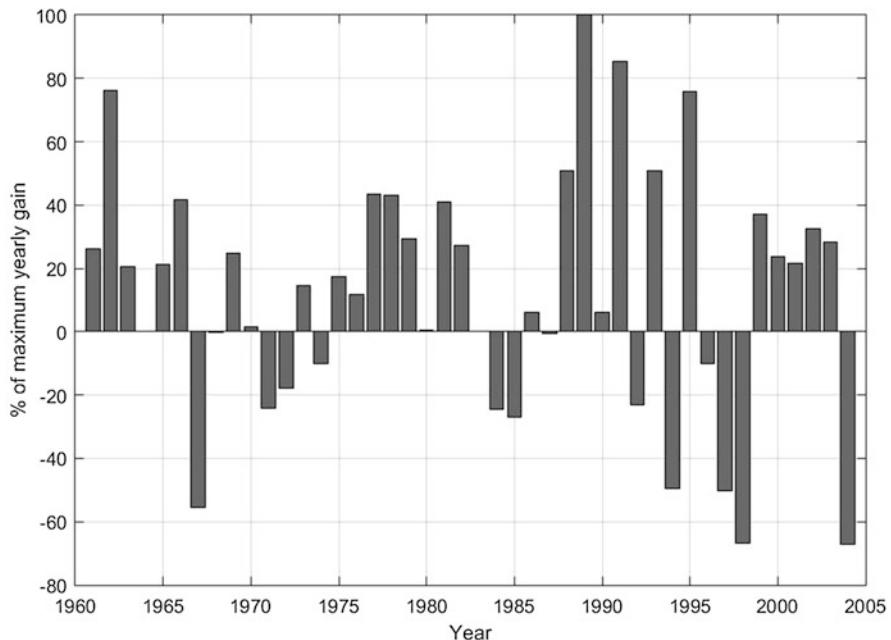


Fig. 22 Net gain implied by the use of inflow forecasting model based on atmospheric variables for Churchill Falls valley (%)

years, this gain may be negative, i.e., the historical mean may have a better predictive value: for such years, it may be concluded that atmospheric variables are not sufficient to explain the inflow amplitude. By combining these differences over all planning years, we can calculate a mean and a standard deviation which can be considered as a measure of the quality of the inflow estimations provided by the forecasting model used.

We have experimented with our methodology using historical data from Hydro-Québec main network and energy market structure as they were in 2003. We have focused our tests on inflow forecasting for Churchill Falls valley for the years 1961–2004. As our model can only take into account one set of scenario weights and because these weights differ significantly from one site to another, we had to limit our study to a given site. Figure 22 presents the experiment results. It shows a convincing argument for Hydro-Québec to keep on enhancing their probabilistic long-term forecasting system. Retrospectively, high positive financial returns would have been obtained in average if the Churchill Falls hydropower complex could have been managed during the last 43 years on the basis of a more evolved probabilistic seasonal forecasting system for inflows. The yearly net gains, which take into account water value in the reservoirs at the end of each year, energy sale revenues, and energy purchase cost, are most of the time positive. The expected net gain over all years is commercially significant.

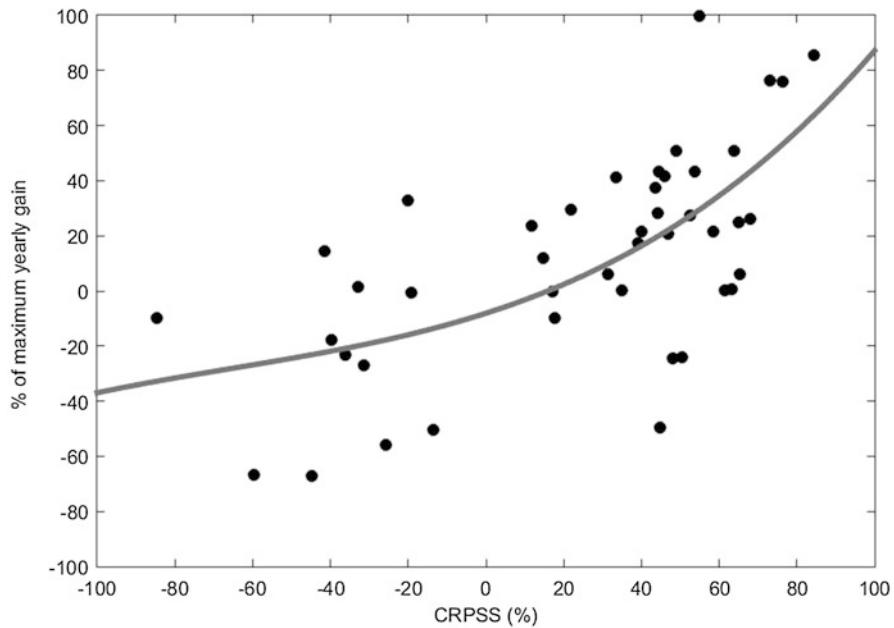


Fig. 23 Relation between the net gains and the corresponding values of the CRPSS

This result is coherent with those obtained with the CRPS (average skill score of 26%). In fact, Fig. 23 presents a scatter plot which shows a significant relation between the output of the optimization model and the corresponding values of the CRPSS presented in Fig. 17. The Kendall rank correlation coefficient amounts to 0.45.

7 Conclusion

In this chapter, we shared some of our experience with mid- and long-term hydrological forecasts verification.

First, we cannot reinforce enough the necessity of using proper metrics as mentioned by many authors, namely, Gneiting and Raftery (2007) and Bröcker and Smith (2007), if we want to obtain a reliable and true evaluation of a forecast's quality.

The second important issue we want to bring is that when facing operational hydrological forecasts, we often do not know what is the real statistical distribution of the forecast (the predictive density function). This is in part because we use a hydrological model, and there is some expertise from a forecaster in issuing the forecast. In fact, forecasters may modify the output forecasts “on the fly” by

changing input observations or initial states. Hypotheses are thus not well known and therefore predictive distributions not well specified. We showed that the true distribution of the forecast may not be what was expected: we may be issuing trimodal forecasts when we think we are producing forecasts with a Pearson type of distribution. Oftentimes, we also do not know the observations' distribution. These facts should be accounted for, because when using scores, we make assumptions about these distributions, and when calculating score values, we usually have to account for some distribution. The use of nonparametric estimation methods helps to alleviate that problem, as was shown.

One more difficulty with the evaluation of forecast quality is the choice of a metric. We clearly showed that different scores have different behavior and evaluate the same forecast differently. Since scores have different designs and behaviors, such a result was expected. A forecast quality analysis will then depend in part on the choice of the metrics. If one chooses an insufficient variety of scores, results can be misleading and can lead to incorrect conclusions about a system's performance or needs of improvement. One should note that often, a single score is used in practice, generally the CRPS. As shown in the applications, the CRPS can lead to different conclusions than with other scoring rules. This may be explained by the fact that the CRPS is not sensitive to the shape of predictive densities and therefore may not detect some problems with the forecasting system. We should reinforce that efficient and useful forecast verification will rely on a variety of scoring rules and diagnostic tools.

When assessing operational forecasts, one more problem that we may face is the relative shortness of the forecast-observation archive time series. It is critical to work and evaluate uniform phenomena; that means that when working with series exhibiting seasonal behaviors, we should divide our sample into uniform subsamples, reducing the series length. That brings some problems into play. Our score evaluations will be affected by the number of observations in the series. Adding or removing one observation can affect the mean score greatly. Some thoughts should then be given to the length of our records before making any inference about our forecast verification method or forecast performance. It is critical to ensure that we use a sufficiently long and uniform archive of observations from the same distribution.

In order to compare scores together (of different forecast periods, or tools, or methodologies) one should not rely on the estimated score values only but should also consider the uncertainty around the evaluated score value. This is especially true with short series with a limited number of observations, where the evaluation of a mean value can exhibit some volatility. We cannot emphasize enough the importance of computing some sort of confidence interval around any score that is evaluated, a practice that is surprisingly lacking. Comparing two values without these confidence intervals is again not a recommended practice and, once again, can lead one to make incorrect decisions based on insufficient information.

One more challenge in forecast verification relates to the evaluation of more extreme conditions. By definition, an extreme event or condition is not common. We will then often be left with very small sample sizes, with the associated

aforementioned problems. In addition, some scores perform poorly in the evaluation of extreme value forecasting. For instance, Naveau et al. (2014) and Taillardat (2017) demonstrated that the CRPS is in fact not suitable for the evaluation of extreme value type of predictive distributions (GEV, generalized Pareto). Hedging or weighting strategies can be employed, but it can lead to improper or degenerative scores (Gneiting and Ranjan 2011). The approach proposed by Taillardat (2017) might be a solution to that problem.

Another important issue for the evaluation of hydrological forecasts relates to multivariate problems, such as evaluating the quality of forecasts for a series of watersheds, where there is a spatial dependence structure. This situation will often present itself to operators that have to deal with decision-making at more than one site, where the decision at one site may have an impact at a different site for which a decision is needed. In such situations, one should use an adapted metric to properly consider the dependence structure. The use of a typical univariate score will not provide accurate information about forecast performance. In this chapter, we illustrated the use of a generalization of the CRPS to multivariate forecasts, the energy score. Such an approach should be part of best practices in forecast evaluation in hydrology.

Finally, assessment of the forecast value is complementary to the forecast quality evaluation since it is specific to an application, and it is critical for users to help them understand the benefits of using forecasts in their decision making process. A planning model was used to evaluate the value of specific forecasts and to communicate their skill. It is clear that forecast verification is a complex science. If one does the exercise properly, it is expected that a number of metrics will be used, each associated with confidence intervals. The amount of information that will be available is quite large and will often give a complicated picture of forecast quality, with apparently conflicting information. For a trained expert, such a situation is interesting, as all the needed information will be available. However, we know from experience that in trying to communicate the value of a forecast based on its performance, such a plethora of information often will drown the message and the nonexpert will be left with no useable indication about forecast performance. With that in mind, we used a generation planning optimization model with real data to conduct an experiment. We found that the output of the optimization model is well correlated, on an observation by observation basis, with the CRPS evaluation of the same data set of real inflow forecasts and observations. In addition, the mean CRPS skill score shows real skill in the new experimental forecasting method, and our economical evaluation indicated that the new experimental forecasting method would lead to significantly positive expected returns. That enabled us to simplify the message to higher management: instead of presenting complex statistical values with little meaning, we presented a single value of expected return. In our opinion, this highlights two issues in addition to the statistical issues already discussed. The first is the importance and the difficulty of communicating forecast quality information in a way that renders that exercise useful. The second related issue is the need for some sort of a

synthetic scoring metric. The world of hydrological forecast verification is still open to exploration and propositions from the community.

References

- L. Alfieri, F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, P. Salamon, Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.* **517**, 913–922 (2014)
- J.E. Bickel, Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decis. Anal.* **4**, 49–65 (2007)
- R.B. Birge, F. Louveaux, *Introduction to Stochastic Programming*. Springer Series in Operations Research (Springer, New York, 1997)
- J. Bröcker, L.A. Smith, Scoring probabilistic forecasts: the importance of being proper. *Weather Forecast.* **22**, 382–388 (2007)
- J.D. Brown, M. He, S. Regonda, L. Wu, H. Lee, D.J. Seo, Verification of temperature, precipitation, and streamflow forecasts from the NOAA/NWS hydrologic ensemble forecast service (HEFS): 2 streamflow verification. *J. Hydrol.* **519**, 2847–2868 (2014)
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009)
- A.P. Dawid, Present position and potential developments: some personal views: statistical theory: the prequential approach. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984)
- G. Day, Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan. Manag.* **111**, 157–170 (1985)
- P. Friederichs, T. Thorarinsdottir, Forecast verification scores for extreme value distributions with an application to peak wind prediction. *Environ. Sci. Technol.* **23**, 579–594 (2012)
- C. Genest, A.-C. Favre, Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **12**, 347–368 (2007)
- T. Gneiting, A.E. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007)
- T. Gneiting, F. Balabdaoui, A.E. Raftery, Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **69**, 243–268 (2007)
- T. Gneiting, L.I. Stanberry, E.P. Grimit, L. Held, N.A. Johnson, Assessing probabilistic forecasts of multivariate quantities, with applications to ensemble predictions of surface winds. *Test* **17**, 211–235 (2008)
- T. Gneiting, R. Ranjan, Comparing density forecasts using threshold- and quantile-weighted scoring rules. *Journal of Business & Economic Statistics* **29**, 411–422 (2011)
- P.J. Huber, *Robust Statistics* (Wiley, New York, 1981)
- V.R. Jose, A characterization for the spherical scoring rule. *Theor. Decis.* **66**, 263–281 (2009)
- R. Krzysztofowicz, The case for probabilistic forecasting in hydrology. *J. Hydrol.* **249**, 2–9 (2001)
- P. Naveau, R. de Fondeville, D. Cooley, H. Benveniste et al., Scores (CRPS), inference and extremes. *Séminaire Statistique des Sommets de Rochebrune*, 30 Mar–4 Apr 2014
- L. Perreault, *Vérification de prévisions hydrologiques probabilistes – Version 2*. Technical Report IREQ-2013-0149, Institut de recherche d'Hydro-Québec (2013)
- L. Perreault, R. Garçon, J. Gaudet, Modelling hydrologic time series using regime switching models and measures of atmospheric circulation. *La Houille Blanche* **6**, 111–123 (2007)
- P. Pinson, J. Tastu, *Discrimination Ability of the Energy Score*. Technical Report, Technical University of Denmark (2013)
- C. Robert, G. Casella, *Monte Carlo Statistical Methods* (Springer, New York, 2000)
- M. Scheuerer, T.M. Hamill, Variogram-based proper scoring rules for probabilistic forecasts of multivariate quantities. *Mon. Weather Rev.* **143**, 1321–1334 (2015)
- R. Selten, Axiomatic characterization of the quadratic scoring rule. *Exp. Econ.* **1**, 43–62 (1998)

- O.G.B. Sveinsson, U. Lall, V. Fortin, L. Perreault, J. Gaudet, S. Zebiak, Y. Kushnir, Forecasting spring reservoir inflows in Churchill Falls basin in Quebec Canada. *J. Hydrol. Eng.* **13**, 426–437 (2008)
- M. Taillardat, O. Mestre, M. Zamo, P. Naveau, Calibrated ensemble forecasts using quantile regression forests and ensemble model output statistics. *Mon Weather Rev.* **144**, 2375–2393 (2016)
- M. Taillardat, Méthodes Non-Paramétriques de Post-Traitement des Prévisions d'Ensemble. PhD Thesis (2017)
- F. Weber, L. Perreault, V. et Fortin, Measuring the performance of hydrological forecasts for hydropower production at BC Hydro and Hydro-Québec, in *Proceeding of the 18th Conference on Climate Variability and Change*, AMS, Atlanta, 30 Jan–2 Feb 2006



Application of Hydrological Forecast Verification Information

Kevin Werner, Jan S. Verkade, and Thomas C. Pagano

Contents

1	Introduction	1014
2	What Constitutes Good Forecast Verification?	1016
3	Who Can Use Verification Information?	1016
4	Verification for Model Developers and System Designers	1017
4.1	Role in the Forecast Process	1017
4.2	Actions Stakeholders Can Take to Improve the Forecast Process	1017
4.3	Context and Constraints	1018
4.4	Verification Needs	1019
5	Verification for Operational Forecasters	1020
5.1	Role in the Forecast Process	1020
5.2	Actions Stakeholder Can Take to Improve the Forecast Process	1020
5.3	Context and Constraints	1021
5.4	Verification Needs	1021
6	Verification for Forecast Users	1023
6.1	Role in the Forecast Process	1023
6.2	Actions Stakeholder Can Take to Improve the Forecast Process	1024

K. Werner (✉)

National Weather Service, National Oceanic and Atmospheric Administration, Salt Lake City, UT, USA

e-mail: kevin.werner@noaa.gov

J. S. Verkade (✉)

Deltares, Delft, The Netherlands

Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, The Netherlands

Delft University of Technology, Delft, The Netherlands

e-mail: jan.verkade@deltas.nl; j.s.verkade@tudelft.nl

T. C. Pagano (✉)

Bureau of Meteorology, Melbourne, VIC, Australia

e-mail: tompagan@bom.gov.au

6.3	Context and Constraints	1024
6.4	Verification Needs	1025
7	Verification for System Administrators	1027
7.1	Role in the Forecast Process	1027
7.2	Actions Stakeholder Can Take to Improve the Forecast Process	1027
7.3	Context and Constraints	1027
7.4	Verification Needs	1028
8	Recommendations and Conclusions	1029
	References	1030

Abstract

Verification studies and systems often focus solely on the exercise of verifying forecasts and not on the application of verification information. This chapter discusses the potential for application of hydrological forecast verification information to improve decision-making in and around the forecast process. Decision-makers include model developers and system designers, forecasters, forecast consumers, and forecast administrators. Each of these has an important role in decisions about forecasts and/or the application of forecasts that may be improved through use of forecast verification. For each, we describe the role, the actions that could be taken to improve forecasts or their application, the context and constraints of those actions, and needs for verification information. Consistent with other studies and assessments on forecast verification, we identify the need for a routine forecast verification system to archive data, plan for operations, measure forecast performance, and group forecasts according to application. Further, we call on forecast agencies and forecast consumers to use forecast verification as a routine part of their operations in order to continually improve services and to engage others to use forecast verification to improve decision-making.

Keywords

Hydrological forecasting · Forecast verification · Decision making

1 Introduction

Good forecast verification only acquires value through its ability to improve the effectiveness of the forecasting systems and users' decisions (Stanski et al. 1989). The present chapter explores the link between hydrological forecast verification, i.e., the process of quality-assessing hydrological forecasts, and improving the forecasting systems and users' decisions.

Quality assessment evaluates the fitness for use of the forecasts by affected stakeholders, such as natural resource managers and people at risk. Verification focuses on the accuracy-related aspects of quality. The reasons for verification fall broadly into three categories: administrative, economic, and scientific (Brier and

Allen 1951; Stanski et al. 1989; Welles et al. 2007). **Administrative** reasons include the justification of the cost of implementation of or improvement of a forecasting system to whoever bears the costs of that system (often ultimately the taxpayer). **Economic** reasons are the expected benefits accrued to a stakeholder, through the use of the forecasts. **Scientific** verification includes the identification of strengths and weaknesses of a forecast product in order to define research and development that will lead to improvements in the forecasts. This can impact the systems for modeling physical processes (e.g., rainfall to runoff, hydrodynamic routing), the forecasters operating these systems, or both.

Once made aware of verification and its application, forecast users maintain a strong appetite for verification information (Hartmann et al. 1999). Increased reporting of past forecast performance was ranked as the highest of 23 development priorities in surveys of users – members of European Union member states’ national and subnational hydrological forecasting agencies – of the European Flood Awareness System (Wetterhall et al. 2013). Considering that its cost and complexity are much lower than traditional investments, such as improving physical model representations (Pagano et al. 2014), it is remarkable how frequently this investment is not made.

Despite the benefits, there is a growing, but we believe still relatively underdeveloped, culture of forecast verification in operational hydrology (Welles et al. 2007; Welles and Sorooshian 2009), with some notable exceptions that have developed since (Bureau of Meteorology 2015; Demargne et al. 2009). For example, it took 80 years of seasonal water supply forecasting in the Western USA before the first scientific verification of those operational forecasts was published (Pagano et al. 2004).

The lack of hydrologic forecast verification is not for a lack of verification methods. See, for example, Jolliffe and Stephenson (2012), Murphy (1993), Wilks (2011), WWRP/WGNE Joint Working Group on Forecast Verification Research (2015) and the references therein. The literature is also rich on specific analyses that can be performed including, for example, measures-oriented and distributions-oriented verification (e.g., Bradley et al. 2004; Murphy 1997) and conditional verification (Bradley and Schwartz 2011). In addition, there are software packages available to facilitate the computation of these measures (see Pocernich (2012) for a recent discussion thereof).

Verification metrics must be aligned with the goals of the verification activity – ultimately, the improved effectiveness of the forecasts. The present chapter explores the link between computing the verification metrics and improving the forecast process and products, on which the scientific literature is nearly silent. The chapter is structured accordingly. In Sect. 2, the general attributes of a good verification are introduced. The effectiveness of the verification depends on the intended audience and so Sect. 3 contains an overview of verification users. Each category of user and their roles in forecasting, their relevant decisions, and their verification needs are then addressed in subsequent sections. The chapter ends with some general conclusions as well as calls for action.

2 What Constitutes Good Forecast Verification?

The qualities of good forecast verifications are largely similar to the qualities of good forecasts (World Meteorological Organization 2013) and, more generally, good information (Wang and Strong 1996). These include aspects such as production, credibility, accuracy, transmission, and messaging:

- Production pertains to the act of producing verification information: is the verification easy to generate and is it produced routinely?
- Credibility refers to how the verification is perceived by users: is it honest, impartial, and unprejudiced?
- Accuracy pertains to how correct the verification is, technically speaking: have the verification measures been calculated correctly, and has their uncertainty/statistical significance been accurately quantified?
- Transmission refers to how the verification gets to the users: is the verification available soon after the forecast is made? Are evaluations for the public distributed freely and are easily accessible?
- Finally, messaging pertains to how the verification is framed for the user: is the verification clear and easy to understand? Is the verification in meaningful units and is it expressed in users' terms? Is it relevant and specific to a given user? Is it meaningful to those using it?

The primary theme of these aspects is “know and serve your users effectively.” Who is the verification for and what is that person’s motivation? What information do they need and what are their limitations? These issues are explored in the remainder of this chapter.

3 Who Can Use Verification Information?

A typology of users of verification information starts with a conceptualisation of how verification information is used to improve the effectiveness of forecast products. As a starting point, we use the “forecast – decision – response” model. Here, “forecast” is modeled as a single process containing multiple subprocesses including model development and real-time forecasting. The outcome of that process – a forecast – is communicated to a user so as to inform a forecast-sensitive decision. Depending on the purpose of the forecasting system, the response can vary from setting the height of a water gate to warning an at-risk community against an impending flood to making a conscious choice to do nothing. In some – but not all – cases, the decision affects the observed outcome (e.g., attempts to prevent the flood succeed).

Figure 1 depicts the role of verification in the “forecast – decision – response” process as well as the adjacent “model development” and “observation” processes. Verification is often done by comparing model outcomes, forecasts, or forecast-sensitive decisions with observations. The information about the quality of these is fed

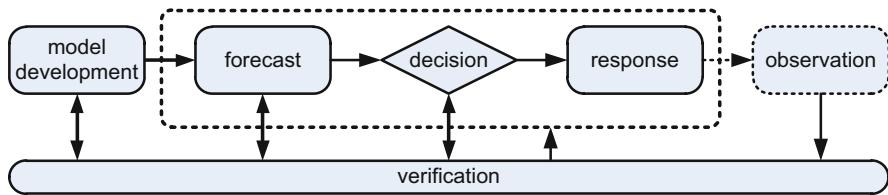


Fig. 1 Role of verification in the process and forecast application information flow

back to model developers, forecast–decision–response system designers, forecasters, and/or decision-makers. Ultimately the goal is for these actors to improve the forecast and its application based on verification. This is described in subsequent chapters.

4 Verification for Model Developers and System Designers

4.1 Role in the Forecast Process

Developers are those who contribute to the creation and improvement of systems used to produce forecasts (Pappenberger et al. 2015a). This is distinct from forecasters (described in the next section) who are the intended users of the forecast production systems. Developers may include research scientists and trained personnel within forecasting agencies but also those in the wider community, such as academics and consultants.

A typical forecasting system may include one or more sources of guidance, such as statistical or dynamical models of runoff. Models are articulations of the scientific community’s views about how natural systems behave (Pagano et al. 2014). Dynamical hydrological models typically contain generalized laws relating precipitation, snow melt, and evaporation to runoff but also contain observed parameters derived from catchment characteristics (e.g., catchment area) and conceptual parameters that are tuned to local observed data.

Forecasting systems may include chains of models, such as one dynamical model predicting future rainfall forcing a different dynamical model predicting runoff, followed by a statistical model to reduce forecast biases or quantify uncertainty.

4.2 Actions Stakeholders Can Take to Improve the Forecast Process

There are a variety of ways in which modelers can improve models, such as algorithms, description of physical processes, and use of data and interfaces. For example, improved algorithms for solving the model equations can lead to faster model execution and/or greater numerical stability of the results (Kavetski and Clark 2011). New regression methodologies may also improve statistical models.

Many operational forecasters use simple “bucket style” rainfall-runoff models (Hartmann et al. 2002), but since the 1990s there has been increased development of fully spatially distributed and physically based models, including many more physical processes than were previously modeled. Modelers can also build better simulations of processes that are the result of human activity, such as floodplain obstructions and diversions. Rather than adding processes and complexity, others have taken the approach to reduce existing models to their simplest yet most effective forms, testing the results on data from thousands of catchments (e.g., Perrin et al. 2003).

Dynamical models typically accept time series data of forcing variables, but also simulate intermediary catchment processes such as soil moisture. Data assimilation (e.g., Liu et al. 2012) is the process of incorporating observations into the model. This may include, for example, comparing recent simulations and observed discharge so as to update simulated soil moisture to bring the simulated discharge closer to the observed. Researchers can develop improved data streams and methods for using available data.

Modelers can also study the interfaces between models. For example, precipitation forecasts from weather models have historically had substantial biases and coarse spatial resolution that should be addressed so as to improve hydrologic forecast accuracy (Cuo et al. 2011). The weather model may provide a single deterministic forecast, and the modelers may develop a process to convert this into a probabilistic or ensemble forecast (see, for example Robertson et al. 2013; Weerts et al. 2011).

While the scientific community can improve models in a general sense, in-house developers (and consultants) play a role in refining and tuning systems in any forecasting agency. For example, an agency may have adopted an approach for calibrating dynamical model parameters, but alternative approaches may be evaluated and compared for specific forecasting challenges (i.e., climate, landscapes, data availability). The in-house developers may also decide that the parameters of a model of a certain catchment could be improved and may want to compare model performance with the old and new parameters. Finally, in-house developers may be responsible for the interfaces for visualizing and interacting with the model guidance and may find that adjustments may lead to more reliable interpretation of the information available to the hydrologists creating the official forecasts (Demargne et al. 2009).

4.3 Context and Constraints

Research scientists have developed and improved rainfall-runoff models for decades. Given that these models are not solely used for operational forecasting (e.g., they may be used to estimate historical water availability when observations are incomplete, or design structures in the floodplain to withstand a given level of risk), there is often only an indirect link between forecaster needs and modelers efforts. Similarly, the weather models that generate rainfall forecasts

are also designed to predict many other variables such as air pressure, temperature, winds, humidity.

The technical sophistication of research scientists is commonly very high, and therefore this audience likely requires little additional training to make effective use of formalized and complex verification approaches. However, each discipline may have its own terminology, and therefore there is the potential for miscommunication and misinterpretation. For example, terms like optimization, parameterization, calibration, tuning, and postprocessing have different meanings in the weather forecasting, hydrology, and water management communities. Additionally, independent researchers may give the different names to the same verification measure. Further, research scientists may ultimately develop complex verification approaches that are inaccessible or incomprehensible to other audiences because of their sophistication, terminology, and specialized purposes.

4.4 Verification Needs

This audience is likely to ask questions such as “How can the system create better forecasts?” and “Does the accuracy of the forecasts match our expectations, and, if not, why not and what can be done to improve them?”

In diagnosing problems or identifying ways to improve the forecasts, these users would like to control as many factors as possible. They may not even study the final official forecasts received by users but may generate retrospective forecasts from the system with some operational realism but without confounding factors, such as poor-quality or latent observational data. For example, the Australian Bureau of Meteorology sponsored research to set up and evaluate six hydrologic models for several hundred Australian catchments (Pagano et al. 2009). Each model was calibrated using the same method and forced by identical historical time series data. Skill was evaluated using a variety of measures (Nash Sutcliffe, Nash Sutcliffe of log-transformed flows, correlation, bias, and four other diagnostic scores). The overall best performing model was selected for implementation as part of the Bureau’s new operational short-term river-forecasting service.

In the model-selection experiment, models did not generate forecasts but rather simulations by using observed rainfall data, not forecast rainfall. Later research used one rainfall-runoff model, but combined it with a variety of forecast rainfall sources, again testing on many catchments, to determine the most suitable forecast rainfall source (Bennett et al. 2013). Further research isolated other factors, such as statistical postprocessing.

Finally, upon implementation of the system, Bureau hydrologists monitored the forecasts and observations daily during a preoperational trial, documenting unexpected or unusual behavior. The historical verification datasets were used to detect the unusualness of the error in any specific forecast. Some unusual behavior was evident by visual inspection of the hydrograph time series (e.g., the duration of the flood event, rates of recession). Hydrologists created algorithms to detect such

behavior and then computed the measures across the past year of forecasts at all catchments to determine how widespread/frequent any particular problem was.

There are many examples of targeted research studies to improve specific model processes, some of which have an operational focus. For example, Bryant and Painter (2009) correlated surface radiative forcing with errors in the calibration dataset of the US National Weather Service's operational model in four mountainous catchments. This identified that a currently unmodeled process (dust on snow changing the surface albedo, leading to earlier snowmelt) was likely responsible for some part of the operational forecast error. The agency then knew it could either improve the model itself or warn operational forecasters of this effect so that they could make adjustments during dust events.

5 Verification for Operational Forecasters

5.1 Role in the Forecast Process

Operational forecasters are responsible for the routine production of streamflow forecasts. Forecasters typically operate in a time and data constrained environment where they must produce a forecast by certain times of day based on the best data available at that time. Forecasters are frequently involved in some or all of the following activities: reviewing data, performing data quality control, running and adjusting models, interpreting model output, assessing forecast confidence, interacting with other forecast producers (e.g., meteorologists), communicating forecasts, coordinating with water managers whose actions both depend on and affect river flow, translating model output into the decision-maker's context, and responding to user requests (Pagano et al. 2014). Depending on the context of the organization and forecast conditions, forecasters can engage in some or all of these activities multiple times per day. Some operational forecasters also serve as system developers and/or administrators during nonoperational periods.

5.2 Actions Stakeholder Can Take to Improve the Forecast Process

Being central to it, forecasters are well positioned to improve the forecast process. Forecasters routinely assess the quality of and seek to improve the forecast model runs. Forecasters also routinely apply expert judgement based on experience to improve forecasts. For example, in the US NWS, hydrologists may run a dynamic simulation model several times making modifications to the model states or input data as part of the process to achieve the best possible forecast.

Experienced forecasters often have very good mental models of how nature behaves and have first-hand experience with the performance of the guidance available to them. They often can readily recall situations in which the models

have failed or the outcome was not as expected. Forecasters may provide system developers with lines of investigation on how to improve the models.

While forecasters often successfully employ intuition based on experience and conduct “sanity checks” using heuristics, they also have their own biases. For example, forecasters may exhibit overconfidence, for example, thinking that there is an 80% chance of a typically rare event happening, when it is only 20% likely (Nicholls 1999). This highlights the value that adequate systems of feedback can bring to assist forecasters identify and limit these biases and test the effectiveness of their heuristics.

5.3 Context and Constraints

Forecasters have a range of statistical sophistication. Most forecasters have an academic background in science and/or engineering and therefore have at least a basic level of training in statistics. Given the range of approaches to improving the forecast, and possible limited exposure of the forecaster to verification practices, an individual forecasters’ experience with applying statistical techniques common to forecast verification can be variable. Some forecasters and their organizations are well versed in verification and have incorporated it into the forecast production cycle. Other forecasters and their organizations have never routinely incorporated forecast verification and are typically much less familiar with the use of statistical techniques.

Forecasters, and the forecast verification systems available to them, frequently emphasize recent forecasts. Particularly during a long duration flood or even drought event when hydrologic conditions persist over time, forecasters (and people more generally) have a tendency to examine and utilize recent forecast performance as a proxy for current forecast accuracy (Muir and Moray 1996). While this practice may add value to the current forecast by identifying and correcting forecast biases and problems in real time, it also may lead to an overweighting of recent forecast skill at the expense of a longer verification analysis.

Forecasters are often constrained by time and may not have access to systems to configure their own verification measures. They may informally compare their forecasts to observations and have subjective impressions of performance, which may be difficult to quantitatively articulate. Forecasters may also have concerns about the consequences of verification on the human aspects of the forecasting system (e.g., if they are shown to be underperforming relative to their peers, will there be professional ramifications for them?).

5.4 Verification Needs

This audience is likely to ask questions such as “Should I trust a particular source of guidance?” “What are the likely errors in today’s forecast?” “How can I best blend together multiple sources of guidance, along with my own understanding of the

situation that may not already be captured by the models?" "Do my operational rules-of-thumb have a scientific basis?" and "Am I adding value to the forecast process?"

Verification can be used to ground forecasters' understanding of the performance of the models. Impressions of overall system reliability that are derived subjectively often do not match the actual system performance (Skitka et al. 1999). People particularly struggle to know how much to trust a model when its quality is not consistent (Parasuraman and Riley 1997), in part because trust is conditioned on the worst behaviors of the system, i.e., the largest errors in recent memory (Muir 1994; Muir and Moray 1996).

One of the best ways to let a forecaster know if particular model guidance should be trusted is to integrate measures of uncertainty in the real-time products themselves. For example, seasonal climate forecasters produce long-lead outlooks of precipitation and temperature, and in many locations, seasons, and lead times, individual tools may have no appreciable skill. On the forecast maps, these cases are displayed with a shading, indicating skill is below a threshold value to warn the forecaster against putting too much trust in that product (Fig. 2).

Naturally, the reliability of such uncertainty quantifications is its own verification issues. For example, many ensemble streamflow predictions only account for uncertainty in future rainfall, whereas there is also uncertainty in the model structure, parameters, and other factors. It is common for forecasters to view time series charts of current and recent forecast hydrographs along with recent observations, so as to visualize the errors of past forecasts and place the current forecast in context. This assumes that there is persistence in errors and biases. Here, the forecaster is using verification information combined with expertise and judgment to correct and blend forecasts, a process that is done formally and objectively with techniques such as Bayesian Model Averaging (Duan et al. 2007).

While valuable, forecaster judgment and intuition require training, to be communicated effectively and feedback so that such judgements and intuition improve. Rapid, relevant, and unambiguous feedback is the key to improving intuitive expertise (Kahneman and Klein 2009). If the final official product contains expert input, it is good to also keep a separate record of unadjusted models and/or objectively blended guidance to compare to. Therefore, the forecaster can see if his/her adjustments had a positive or negative impact on product accuracy. Such feedback should be given rapidly after the event, while the situation is still fresh in the forecaster's mind.

Finally, an avenue for expertise development is formal training, for example, in relation to the testing of judgements. Some elements of the natural system cannot be modeled (because of insufficient models, data or a variety of other factors), but the forecaster may have some awareness of this and attempt to incorporate this into the forecasts manually. The previous section gave the example of dust on snow affecting hydrograph simulations. Prior to the research study, the forecaster may have developed a heuristic rule-of-thumb (e.g., when dust happens, the forecasts should be 30% lower than otherwise). Verification can be used in training to recognize, question, and improve those heuristics. For example, the historical error of the model may

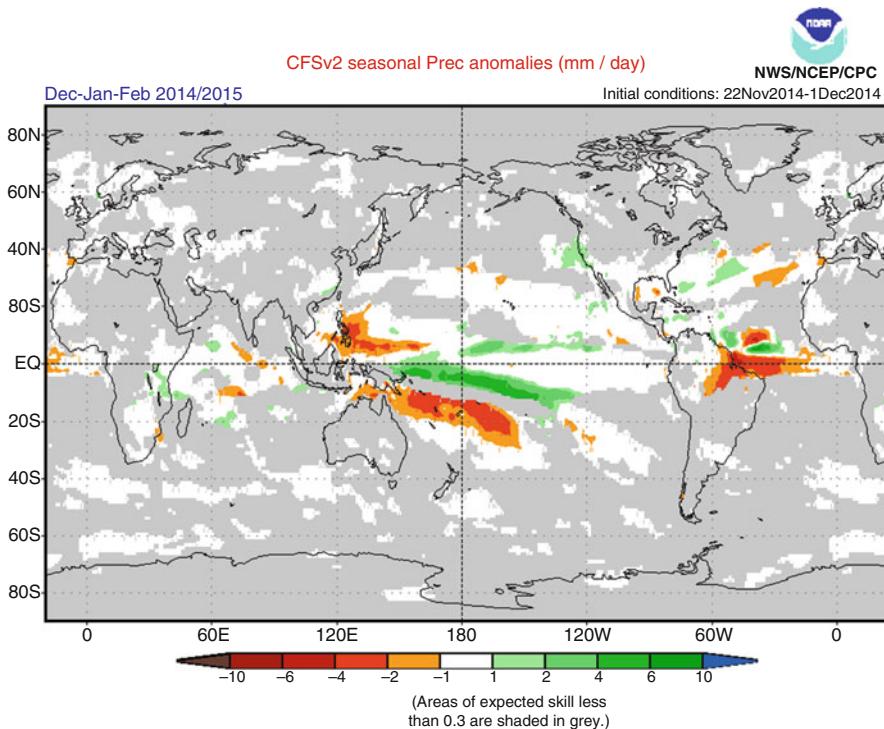


Fig. 2 Coupled Forecast System model version 2 seasonal precipitation forecast with skill mask. Forecasts expected to have low skill, based on the historical performance of the model for that region, season, lead time, and element, are censored by a gray mask (From <http://www.cpc.noaa.gov/products/CFSv2/CFSv2seasonal.shtml>)

have been 20%, and therefore a 30% change would be an overadjustment. Such training should include case studies to introduce the issue to forecasters but should also include broad sets of forecasts to ensure the generalizability of the training. Ultimately, forecasters should drive this training not only to improve their own forecasts but also to identify opportunities for improvement for model developers and administrators as well.

6 Verification for Forecast Users

6.1 Role in the Forecast Process

As the intended recipients of the forecasts, users play a critical role in verification and the forecast process. Just as the hydrologist may consider multiple sources of guidance while determining the final official forecast, the user may consolidate forecasts from multiple sources and place them in context of their situational

awareness of other factors, natural and societal. The users interpret the forecast and assess their confidence in the product. The forecast then informs a decision-making process that is largely specific to a given user, and is affected by the user's objectives, culture, resources, and other factors.

There is a diverse array of user communities, including citizens in danger of being flooded, reservoir operators, irrigators, financial traders (e.g., commodities, hydro-power), insurance agencies, emergency responders, disaster relief agencies, and the media. The forecasts may affect decisions to evacuate a community, store or release water from a reservoir, plant a certain type of drought-resistant crop, and so on.

6.2 Actions Stakeholder Can Take to Improve the Forecast Process

There are frequent cases of interested stakeholders contacting forecast providers for clarifications or more details associated with a particular forecast, or they may generally communicate with forecast providers about their needs (e.g., where they would like forecasts, how often updates are needed). If forecast products are not meeting the user needs, they may decide not to use it, ask to have the product changed, change to another source of forecasts, or even change to another forecast provider who packages the same forecast in a more relevant and accessible form.

Users can also highlight the most important forecasts, such as those for certain locations, or situations (e.g., forecasts crossing above flood-stage are more important to emergency response agencies than forecasts during low flows). Users can specify the minimum level of accuracy required, so as to help prioritize areas for improvement. They can communicate their vulnerability to errors in the forecast, which is tied to their risk tolerance. For example, when making a deterministic forecast, the emergency manager may prefer the forecaster to be conservative, adjusting the forecast towards the worst case scenario, because overpreparing for a disaster is not as dangerous as underpreparing.

Users can ensure that they have a clear understanding of how to interpret forecast products and are familiar with the forecast uncertainties. They can be aware of biases in the forecast and in certain cases develop in-house systems for adding value to the forecasts (e.g., relate a forecast river level to an inundation extent). Verification can help manage expectations around the capabilities of the forecasting systems. Given that users' decisions affect outcomes, they can feed back information to the forecasters about that process, for example, when planned reservoir releases may contribute significantly to downstream flow (and these releases may depend on the forecasts themselves).

6.3 Context and Constraints

Users encompass the entire spectrum of statistical literacy, from highly sophisticated users who may run their own modeling systems in-house and perform their own

verifications, to those with no background at all in statistics. The latter category of users may not understand verification concepts, techniques, or measures and may even have issues interpreting the forecasts themselves. Some users may analyze the forecasts on a daily basis, but others may use the forecasts only very infrequently, such as every few years when floods occur. Just as each scientific field has its own terminology, jargon, and definition of terms, users may have their own language in discussing and evaluating forecasts. Further, because users are often external to forecasting agencies, extra effort may be needed to communicate verification information with them.

Users typically have other inputs to their decision process that are unrelated to forecasts. For example, a reservoir operator may need to adjust reservoir releases to account for dynamic requirements of the ecosystem below the reservoir, the water supply requirements of downstream users, or political considerations to address competing interests. Each of these inputs to users' decision-making processes requires knowledge of diverse areas of input that may not be related to forecasts at all.

6.4 Verification Needs

This audience is likely to have questions such as "Should I trust this forecast?" "How, if at all, should I use this forecast?" "To what extent is forecast quality conditional on attributes of streamflow?" "Can I trust a forecast for an extreme value as much as a routine value?" and "Is the forecast uncertainty compatible with my risk tolerance?" If using probabilistic forecasts, additional questions from the users include "are the forecasts probabilistically reliable? (e.g., when they say 30% probability of flooding, does flooding indeed happen 30% of the time?)". Many members of this audience are also likely not to know how to express the verification questions they have and may likely need verification architects to act as translators.

Users need to know when forecasts are sufficiently reliable for their purposes (Sarewitz et al. 2000). Consistent communication of forecast uncertainty and historical performance can increase forecast credibility (O'Grady and Shabman 1990). Without this credibility, forecasts may not be used (Rayner et al. 2005). Further, the costly consequences of bad outcomes from the use of a particular forecast can devastate user confidence in subsequent forecasts (Glantz 1982).

Given that forecasters face a similar challenge of converting guidance into official warnings as users do converting official forecasts into effective responses, some of the verification information requirements for users are similar to those of forecasters. They may want to give a quantitative basis to their subjective impressions of the forecast skill. They may want a better understanding of the uncertainty in a specific forecast, informed partly by the recent performance of similar forecasts. In that regard, many of the same approaches, such as integrating verification/uncertainty information into the products themselves, are useful. A critical distinction between forecast producers and users is that the users may be interested in

a subset of “high impact” forecasts only. Also, users’ risk tolerance may not be the same as that of forecasters.

For example, forecast producers may look at national maps of forecast skill to get an overall sense of the performance of the forecast system. In contrast, water managers in the Upper Colorado River may only be interested in forecasts for their catchment, issued December to April, forecasting for January to September (Hartmann et al. 2002). Emergency managers on the Mekong River may be most interested in those forecasts indicating a future flood, while the current observation is still below the flood level, a situation that only happens on less than 1% of days (Pagano 2014). Not all users are interested in floods; when the Rio Grande River in the USA is about to dry up, fish biologists must prepare to scoop out endangered species from isolated pools (Paskus 2003). User-oriented verifications often segregate certain forecasts of interest based on their location, lead time, season, and magnitude. To ensure users’ needs are met, the user must be involved in the process of how to subset forecasts.

Additionally, the user has a risk tolerance in the sense that he/she may be more vulnerable to forecasts that are too high than too low. For example, the damage from underpreparing for a flood may be much larger than the cost of overpreparing. Standard forecast evaluation measures such as Nash Sutcliffe or correlation would not reflect this asymmetry. Ideally, the forecast should be probabilistic, and then the user could estimate their own risk tolerance (perhaps with the assistance of forecasters, researchers, extension agents, or other specialists) and decide that, for example, it is most appropriate for them to base their planning around the 80% nonexceedence level of the forecast. This process may involve using a collection of historical forecasts in a formal framework (e.g., input into a reservoir optimization tool) or tabletop exercise (Baldwin et al. 2006). The user may even be able to use verification to quantify financial benefit from using the forecasts (Faber and Stedinger 2001; Hamlet et al. 2002; Verkade and Werner 2011).

Regrettably, probabilistic forecasts are underutilized and deterministic, single-valued forecasts are commonly preferred by users. The forecaster may attempt to compensate for this lack of risk management by users by adjusting the products themselves. For example, the agency that issues deterministic forecasts (e.g., 3.4 m peak river height by Thursday) may informally recommend to the hydrologists that it is “better to forecast too high and too early than too low and too late.” Here, the forecasts are purposefully biased because the costs of overpreparedness are much less than those of a disaster (and the associated damages to the forecasters’ reputation). When combined with an understanding of user decision-making, verification can be used to determine if the forecasters’ estimate of the users’ risk tolerance serves the actual user needs.

A key element of verification for decision-makers is to use language that is meaningful and relevant. For example, knowing that a forecasts’ Root Mean Squared Error is 0.6 m or that the Brier Skill Score is 0.3 may not be useful. In contrast, a user may want to know that in 6 out of 10 cases the forecasts are too high, or half of the time the error is more than 0.4 m or the worst error in the past 5 years was 1.5 m. Analogues are easy to understand and visualize (e.g., in the 1983 flood we predicted 7 m and the observed was

6.7) although care should be taken to emphasize that individual forecasts may not be representative of overall performance.

7 Verification for System Administrators

7.1 Role in the Forecast Process

System administrators – sometimes called program managers – are responsible for developing and evaluating business cases, investment decisions, and financial analyses related to the maintenance and development of forecasting systems. While operational forecasters are responsible for the forecasts, system administrators are accountable for their quality. For example, system administrators may be brought to account for the forecasts of high-profile events, such as major floods that had widespread impacts that were poorly forecast. Such activities are important in maintaining the credibility of, and support for, the forecasting agency. If a particular poor-quality forecast gains the public's attention, it is useful for the agency to provide quantitative evidence that most other forecasts have been quite accurate or that the forecast system overall has a certain level of skill. System administrators may facilitate and/or monitor the dialogues among modelers, forecasters, and users. Verification may also be part of the reporting of key performance indicators to government ministers and key stakeholders.

7.2 Actions Stakeholder Can Take to Improve the Forecast Process

While modelers may investigate improvements into a particular aspect of a model, system administrators direct resources (people, money, and assets) to make such an investigation possible and a priority. System administrators may also determine the policy related to forecasting procedures and may direct investment in creating new or modifying existing forecast products. They may choose to centralize or decentralize forecasting and may recruit and train additional forecasters. They may invest in improvements in the data networks, forecaster workflow management software, and forecast center facilities and computing or display equipment.

7.3 Context and Constraints

Similar to forecasters, system administrators are typically time limited, and verification may only be a minor aspect of their overall activities. Administrators must also concentrate on all aspects of forecast quality, such as timeliness, accessibility, and system reliability. In this bigger picture, investments in, for example, forecast communication or digital delivery mechanisms may be a higher priority than improving core aspects of forecast accuracy. Typically verification will be a

component (along with user-based assessment) of a program evaluating the overall effectiveness of the service. Administrators' familiarity with statistical concepts and verification terminology varies, with some administrators being former researchers and/or forecasters and others coming from a business management background or other fields entirely.

7.4 Verification Needs

This audience is likely to ask questions such as "What is the overall quality of service I am providing?" "How is the agency performing compared to its peers?" "Where should I invest in improvements in the system?" "Has the agency realized the benefits of past investments?" and "How was the skill during particularly high-profile events?" System administrators generally have three types of verification information needs: "headline scores," evaluations of system upgrades, and event-specific analyses.

"Headline scores" provide an overall health check of the entire forecasting enterprise and often involve distilling important aspects of quality across many forecasts into a few measures. For example, Pagano et al. (2004) developed an index of performance (Nash Sutcliffe of each site within a 20 year moving window, averaged across 29 spatially and climatically representative sites across the Western USA) that later became a key performance indicator for the agency. This updated annually along with other measures, such as the number of forecasts issued, and reported publicly and to other government agencies. The agency set short-range performance targets. Regrettably, it was difficult to meet these targets because the score, while relatively easy to calculate and communicate, was often influenced by factors that were out of the agency's control, such as climatic variability and extreme events.

When making decisions about future investments in the forecasting system, system administrators may seek a clear value proposition, possibly supported by evidence that past investments have yielded the expected benefits. For example, some authors quantified the operational river-forecasting benefits of accurate rainfall forecasts, relative to other sources of information (Pappenberger et al. 2011, 2015b; Rossa et al. 2011; Welles 2005; Welles and Sorooshian 2009; Zappa et al. 2010). If evidence can be provided that an equivalent investment (e.g., \$1 million) in a more dense raingage network yields much less operational forecast improvement than a \$1 million investment in better rainfall forecasts, then the administrative choice is clear. Part of the challenge lies in defining "forecast improvement": Which forecasts? Improved how? Does the change make some forecasts better and some worse? Currently, operational hydrologic forecasting agencies often lack the resources and frameworks for conducting such quantitative, structured experiments and instead rely on the informed impressions of subject matter experts.

Finally, system administrators may need to investigate the performance of individual forecasts that have captured the public interest. For example, in January 2011, large floods in Queensland Australia inundated a major metropolitan area. The Commission of Inquiry that followed lasted several years and called hundreds of

witnesses and received hundreds of written submissions. Of critical interest was the quality of the river and rainfall forecasts, in particular how well the magnitude of the event was predicted. Administrators in the Bureau provided detailed information about the forecasts, their accuracy, the precedence of the flood magnitude, and forecast quality. The agency also provided information about how the forecasts were produced so as to show that standard operating procedures were followed. Similar investigations followed the 1997 Red River floods in the USA (Pielke 1999), the 1983 Colorado River floods (Rhodes et al. 1984), the Yakima River drought (Glantz 1982), and also earthquake predictions associated with the disaster in L'Aquila, Italy in 2009 (American Geophysical Union 2010).

8 Recommendations and Conclusions

Several documents provide guidelines on the performance assessment of public forecasting services (Gordon and Shaykewich 2000; World Meteorological Organization 2013). Based on this study and the recommendations of those other studies, several features of a good verification system can be offered:

Archival: Systematically preserve historical operational forecasts, as well as corresponding observations, in a consistent machine readable format to facilitate easy processing. Organize the archive in such a way that historical forecasts can be easily retrieved later on. It is essential to archive official products, but also consider archiving original model inputs, outputs, parameters as well as forecaster interventions. Keeping the original data allows scores to be recomputed over time, if the methods or scores ever change.

Planning: Have a verification plan. Know why you want to verify and understand what questions you are attempting to answer, what new information you want to discover. Recognize the diversity of verification needs but do not try to satisfy every need imaginable. Initially at least, keep the verifications simple and start small and grow over time to suit your needs.

Measures: Choose elements and scores that are relevant to the needs of the verification audience. Recognize the differences between accuracy and skill. Weigh the merits (simplicity, cost, relevancy, reproducibility, and others) of subjective versus objective verification methods, favoring objective measures where possible. Use multiple measures to evaluate various aspects of forecast accuracy – rarely can “one number” paint the entire picture.

Grouping: Group similar forecasts so as to identify systematic errors and accumulate enough examples to calculate results with statistical significance. However, beware of overaggregation, lumping together disparate climates, lead times, events, and so on, because this can conceal useful information. Furthermore, if performing user-oriented verification, attempt to include only forecasts that are relevant to that decision-maker’s context. Forecast verification should be stratified to focus on “high impact” and/or difficult forecasts and be done in a way that informs system improvement.

Use: Do not simply verify the forecasts and file away the results in a report. Be prepared to act on the results of the verification, be it to adjust the forecasting system, investing in system improvements, changing forecaster training, and so on.

Engagement: Share the results of the verification in a timely manner, especially providing rapid feedback to operational forecasters on the quality of their performance. Keep stakeholders updated regularly and do not simply deliver the numerical scores but also include an interpretation of the results. Communicate the results in an easy to understand way and make the results easily accessible. Seek feedback from users that the verification is meaningful and effective and is achieving its intended purpose. Have a verification communication plan.

Investments in forecast verification capacities that incorporate these aspects will pay dividends for forecast agencies and their stakeholders. Unfortunately such verification capacities are the exception rather than the rule today. This has created an environment where the various users described in the chapter are forced into making decisions that are not informed by forecast skill often leading to suboptimal forecast application by stakeholders, forecast development by developers and administrators, or forecast production by forecasters.

Instead, we propose that forecast agencies routinely invest in the development and operation of forecast verification capabilities that support data-driven decisions for all stakeholders in and around the hydrologic prediction enterprise. Given that, we believe continual improvements to forecasts would occur as a matter of course through focusing forecaster, developer, and administrator efforts on areas to reduce forecast error and that greater optimization of forecast application would lead to more resilient decision-making by forecast consumers.

References

- American Geophysical Union, AGU statement: investigation of scientists and officials in L'Aquila, Italy, is unfounded. EOS Trans. Am. Geophys. Union **91**(28), 248 (2010). <https://doi.org/10.1029/2010EO280005>
- C. Baldwin, M. Waage, R. Steger, J. Garbrecht, T. Piechota et al., Acclimatizing water managers to climate forecasts through decision experiments, in *Climate Variations, Climate Change, and Water Resources Engineering* (ASCE, Reston, 2006), pp. 115–131
- N.D. Bennett, B.F.W. Croke, G. Guariso, J.H.A. Guillaume, S.H. Hamilton, A.J. Jakeman, S. Marsili-Libelli, L.T.H. Newham, J.P. Norton, C. Perrin, S.A. Pierce, B. Robson, R. Seppelt, A.A. Voinov, B.D. Fath, V. Andreassian, Characterising performance of environmental models. Environ. Model. Softw. **40**, 1–20 (2013). <https://doi.org/10.1016/j.envsoft.2012.09.011>
- A.A. Bradley, S.S. Schwartz, Summary verification measures and their interpretation for ensemble forecasts. Mon. Weather Rev. **139**(9), 3075–3089 (2011). <https://doi.org/10.1175/2010MWR3305.1>
- A.A. Bradley, S.S. Schwartz, T. Hashino, Distributions-oriented verification of ensemble streamflow predictions. J. Hydrometeorol. **5**(3), 532–545 (2004)
- G.W. Brier, R.A. Allen, Verification of weather forecasts. Compend, in *Compendium of Meteorology* (1951), pp. 841–848
- A.C. Bryant, T.H. Painter, Radiative forcing by desert dust in the Colorado River Basin from 2000 to 2009 inferred from MODIS data, in *AGU Fall Meeting Abstracts*, vol. 1 (2009), p. 0501.

- Online Available from <http://adsabs.harvard.edu/abs/2009AGUFM.C33B0501B>. Accessed 27 Jan 2015
- Bureau of Meteorology, *Verification in the Bureau. Framework Report* (Bureau of Meteorology, Melbourne, 2015)
- L. Cuo, T.C. Pagano, Q.J. Wang, A review of quantitative precipitation forecasts and their use in short-to-medium-range streamflow forecasting. *J. Hydrometeorol.* **12**(5), 713–728 (2011)
- J. Demargne, M. Mullusky, K. Werner, T. Adams, S. Lindsey, N. Schweiin, W. Marosi, E. Welles, Application of forecast verification science to operational river forecasting in the US National Weather Service. *Bull. Am. Meteorol. Soc.* **90**(6), 779–784 (2009)
- Q. Duan, N.K. Ajami, X. Gao, S. Sorooshian, Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**(5), 1371–1386 (2007)
- B.A. Faber, J.R. Stedinger, Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts. *J. Hydrol.* **249**(1–4), 113–133 (2001). [https://doi.org/10.1016/S0022-1694\(01\)00419-X](https://doi.org/10.1016/S0022-1694(01)00419-X)
- M.H. Glantz, Consequences and responsibilities in drought forecasting: the case of Yakima, 1977. *Water Resour. Res.* **18**(1), 3–13 (1982)
- N. Gordon, J. Shaykewich, *Guidelines on Performance Assessment of Public Weather Services* (World Meteorological Organization, Geneva, 2000). Online Available from www.wmo.int/pages/prog/hwrp/documents/FFI/expert/Guidelines_on_Performance_Assessment_of_Public_Weather_Services.pdf
- A. Hamlet, D. Huppert, D. Lettenmaier, Economic value of long-lead streamflow forecasts for Columbia River Hydropower. *J. Water Resour. Plan. Manag.* **128**(2), 91–101 (2002). [https://doi.org/10.1061/\(ASCE\)0733-9496\(2002\)128:2\(91\)](https://doi.org/10.1061/(ASCE)0733-9496(2002)128:2(91))
- H. Hartmann, R. Bales, S. Sorooshian, *Weather, Climate, and Hydrologic Forecasting for the Southwest U.S.* (The University of Arizona, Tucson, 1999). Online Available from <http://www.climas.arizona.edu/publication/report/weather-climate-and-hydrologic-forecasting-southwest-us>
- H.C. Hartmann, R. Bales, S. Sorooshian, Weather, climate, and hydrologic forecasting for the US Southwest: a survey. *Clim. Res.* **21**(3), 239–258 (2002)
- I.T. Jolliffe, D.B. Stephenson, *Forecast Verification: A Practitioner's Guide in Atmospheric Science* (Wiley, 2012). Online Available from http://books.google.nl/books?hl=en&lr=&id=DCxsKQeaBH8C&oi=fnd&pg=PT8&dq=jolliffe+stephenson&ots=3Ojk_X1AOy&sig=8hGKrwljwaUgKxxYwxzmafUB8k. Accessed 27 Jan 2015
- D. Kahneman, G. Klein, Conditions for intuitive expertise: a failure to disagree. *Am. Psychol.* **64**(6), 515 (2009)
- D. Kavetski, M.P. Clark, Numerical troubles in conceptual hydrology: approximations, absurdities and impact on hypothesis testing. *Hydrol. Process.* **25**(4), 661–670 (2011)
- Y. Liu, A.H. Weerts, M. Clark, H.-J. Hendriks Franssen, S. Kumar, H. Moradkhani, D.-J. Seo, D. Schwanenberg, P. Smith, A. Van Dijk et al., Advancing data assimilation in operational hydrologic forecasting: progresses, challenges, and emerging opportunities. *Hydrol. Earth Syst. Sci.* **16**(10), 3863–3887 (2012)
- B.M. Muir, Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics* **37**(11), 1905–1922 (1994)
- B.M. Muir, N. Moray, Trust in automation. Part II. Experimental studies of trust and human intervention in a process control simulation. *Ergonomics* **39**(3), 429–460 (1996)
- A.H. Murphy, What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**(2), 281–293 (1993)
- A.H. Murphy, Forecast verification, in *Economic Value of Weather and Climate Forecasts* (Cambridge University Press, Cambridge, UK/New York/Melbourne, 1997)
- N. Nicholls, Cognitive illusions, heuristics, and climate prediction. *Bull. Am. Meteorol. Soc.* **80**(7), 1385–1397 (1999). [https://doi.org/10.1175/1520-0477\(1999\)080<1385:CIHACP>2.0.CO;2](https://doi.org/10.1175/1520-0477(1999)080<1385:CIHACP>2.0.CO;2)
- K. O'Grady, L. Shabman, Communicating the probability of Great Lakes water levels and storms, in *Proceedings of Great Lakes Water Level Forecast and Statistics Symposium*, Windsor (1990) pp. 197–204
- T.C. Pagano, Evaluation of Mekong River commission operational flood forecasts, 2000–2012. *Hydrol. Earth Syst. Sci.* **18**(7), 2645–2656 (2014)

- T.C. Pagano, D. Garen, S. Sorooshian, Evaluation of official western US seasonal water supply outlooks, 1922–2002. *J. Hydrometeorol.* **5**(5), 896–909 (2004)
- T.C. Pagano, H. Hapuarachchi, Q.J. Wang, *Continuous Soil Moisture Accounting and Routing Modelling to Support Short Lead-Time Streamflow Forecasting* (CSIRO Water for a Healthy Country National Research Flagship, Melbourne, 2009)
- T.C. Pagano, A.W. Wood, M.-H. Ramos, H.L. Cloke, F. Pappenberger, M.P. Clark, M. Cranston, D. Kavetski, T. Mathevet, S. Sorooshian, J.S. Verkade, Challenges of operational river forecasting. *J. Hydrometeorol.* (2014). Online Available from <http://journals.ametsoc.org/doi/abs/10.1175/JHM-D-13-0188.1>. Accessed 27 Jan 2015
- F. Pappenberger, J. Thielen, M. Del Medico, The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrol. Process.* **25**(7), 1091–1113 (2011)
- F. Pappenberger, M.H. Ramos, H.L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, P. Salamon, How do I know if my forecasts are better? Using benchmarks in hydrological ensemble prediction. *J. Hydrol.* **522**, 697–713 (2015a). <https://doi.org/10.1016/j.jhydrol.2015.01.024>
- F. Pappenberger, H.L. Cloke, D.J. Parker, F. Wetterhall, D.S. Richardson, J. Thielen, The monetary benefit of early flood warnings in Europe. *Environ. Sci. Pol.* **51**, 278–291 (2015b). <https://doi.org/10.1016/j.envsci.2015.04.016>
- R. Parasuraman, V. Riley, Humans and automation: use, misuse, disuse, abuse. *Hum. Factors J. Hum. Factors Ergon. Soc.* **39**(2), 230–253 (1997)
- L. Paskus, *Why the Silvery Minnow Matters*, AlterNet (2003). Online Available from http://www.alternet.org/story/17152/why_the_silvery_minnow_matters. Accessed 27 Jan 2015
- C. Perrin, C. Michel, V. Andréassian, Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **279**(1), 275–289 (2003)
- R.A. Pielke Jr., Who decides? Forecasts and responsibilities in the 1997 Red River flood. *Appl. Behav. Sci. Rev.* **7**(2), 83–101 (1999)
- M. Pocernich, Appendix: verification software, in *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd edn. (Wiley, Chichester, 2012), pp. 231–240. Online Available from <http://onlinelibrary.wiley.com/doi/10.1002/9781119960003.app1/summary>. Accessed 27 Jan 2015
- S. Rayner, D. Lach, H. Ingram, Weather forecasts are for wimps: why water resource managers do not use climate forecasts. *Clim. Chang.* **69**(2–3), 197–227 (2005)
- S.L. Rhodes, D. Ely, J.A. Dracup, Climate and the Colorado River: the limits of management. *Bull. Am. Meteorol. Soc.* **65**(7), 682–691 (1984)
- D.E. Robertson, D.L. Shrestha, Q.J. Wang, Post processing rainfall forecasts from numerical weather prediction models for short term streamflow forecasting. *Hydrol. Earth Syst. Sci. Discuss.* **10**(5), 6765–6806 (2013). <https://doi.org/10.5194/hessd-10-6765-2013>
- A. Rossa, K. Liechti, M. Zappa, M. Bruen, U. Germann, G. Haase, C. Keil, P. Krahe, The COST 731 action: a review on uncertainty propagation in advanced hydro-meteorological forecast systems. *Atmos. Res.* **100**(2–3), 150–167 (2011). <https://doi.org/10.1016/j.atmosres.2010.11.016>
- D. Sarewitz, R.A. Pielke, R. Byerly, *Prediction: Science, Decision Making, and the Future of Nature* (Island Press, 2000). Online Available from http://books.google.nl/books?hl=en&lr=&id=O0nxEU-deAUC&oi=fnd&pg=PR11&dq=sarewitz+pielke+prediction&ots=F3r_mNYv9p&sig=78GiOAFyglce8xbodoqOVanjRZA. Accessed 27 Jan 2015
- L.J. Skitka, K.L. Mosier, M. Burdick, Does automation bias decision-making? *Int. J. Hum. Comput. Stud.* **51**(5), 991–1006 (1999)
- H.R. Stanski, L.J. Wilson, W.R. Burrows, *Survey of Common Verification Methods in Meteorology* (World Meteorological Organization, Geneva, 1989). Online Available from http://www.eumetcal.org/resources/ukmeteocal/verificationSAV/www/english/msg/library/SWB_Chapter1.pdf. Accessed 27 Jan 2015
- J.S. Verkade, M.G.F. Werner, Estimating the benefits of single value and probability forecasting for flood warning. *Hydrol. Earth Syst. Sci.* **15**(12), 3751–3765 (2011). <https://doi.org/10.5194/hess-15-3751-2011>

- R.Y. Wang, D.M. Strong, Beyond accuracy: what data quality means to data consumers. *J. Manag. Inf. Syst.* **12**, 5–33 (1996)
- A.H. Weerts, H.C. Winsemius, J.S. Verkade, Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrol. Earth Syst. Sci.* **15**(1), 255–265 (2011). <https://doi.org/10.5194/hess-15-255-2011>
- E. Welles, *Verification of River Stage Forecasts* (2005). Online Available from <http://arizona.openrepository.com/arizona/handle/10150/195133>. Accessed 27 Jan 2015
- E. Welles, S. Sorooshian, Scientific verification of deterministic river stage forecasts. *J. Hydrometeorol.* **10**(2), 507–520 (2009)
- E. Welles, S. Sorooshian, G. Carter, B. Olsen, Hydrologic verification: a call for action and collaboration. *Bull. Am. Meteorol. Soc.* **88**(4), 503–511 (2007)
- F. Wetterhall, F. Pappenberger, H.L. Cloke, J. Thielen-del Pozo, S. Balabanova, J. Daňhelka, A. Vogelbacher, P. Salamon, I. Carrasco, A.J. Cabrera-Tordera et al., Forecasters priorities for improving probabilistic flood forecasts. *Hydrol. Earth Syst. Sci. Discuss.* **10**(2), 2215–2242 (2013)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences* (Academic, 2011). Online Available from <http://books.google.nl/books?hl=en&lr=&id=IJuCVtQ0ySIC&oi=fnd&pg=PP2&dq=wilks+statistical&ots=anHlqQBLNU&sig=w7ZsTvkiX5BaOjYRzngpdtC91M>. Accessed 27 Jan 2015
- World Meteorological Organization, *Guide to the Implementation of a Quality Management System for National Meteorological and Hydrological Services* (World Meteorological Organization, Geneva, 2013). Online Available from http://www.wmo.int/pages/prog/hwrp/qmf-h/documents/ext/wmo_1100_en.pdf
- WWRP/WGNE Joint Working Group on Forecast Verification Research, *Forecast Verification: Issues, Method and FAQ* (2015). Online Available from <http://www.cawcr.gov.au/projects/verification/>. Accessed 27 Jan 2015
- M. Zappa, K.J. Beven, M. Bruen, A.S. Cofino, K. Kok, E. Martin, P. Nurmi, B. Orfila, E. Roulin, K. Schröter et al., Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmos. Sci. Lett.* **11**(2), 83–91 (2010)

Part IX

Communication and Use of Ensemble Forecasts for Decision Making



Overview of Forecast Communication and Use of Ensemble Hydrometeorological Forecasts

Jutta Thielen-del Pozo and Michael Bruen

Abstract

Over the last few decades, hydrometeorological forecasting, warning and decision making has benefited greatly from advances in the natural, physical, computing and social sciences. A fast developing computing capability has enabled meteorologists to produce ensemble prediction systems (EPS) that quantify the uncertainty in forecasting and simulating floods, droughts, and in water management decision making. At the same time, the social sciences have helped to understand the human perceptions of risk information and how different actors communicate hazard, risk and uncertainty information. Ultimately hydrometeorological forecasts are used in making decisions. However, to be effective, such decisions must be communicated to the hazard response organisations and to the general public. For this, the communication must be simple and clear, it must be relevant and should come from a trusted source. This overview summarises how such communication is organised for a variety of applications in different countries. It is the effectiveness of the entire system which must be considered and assessed. As ensembles are increasingly used in increasingly longer term management and policy decisions, the range of end-users and their differing requirements can only expand and flexibility and adaptability to individual circumstances will be required from both the natural and social scientists involved.

J. Thielen-del Pozo (✉)
European Commission, Joint Research Centre, Ispra, Italy
e-mail: jutta.thielen@jrc.ec.europa.eu

M. Bruen (✉)
UCD Dooge Centre for Water Resources Research, UCD School of Civil Engineering, Dublin,
Ireland
e-mail: michael.bruen@ucd.ie

Keywords

Ensemble prediction systems (EPS) · Flood · Forecast communication · Hydrometeorological forecasts

Over the last few decades, hydrometeorological forecasting, warning, and decision making have benefited greatly from advances in the natural, physical, computing, and social sciences. A fast developing computing capability has enabled meteorologists to produce ensemble prediction systems (EPS) that are now applied in various sectors including hydrological applications for the forecasting and simulation of floods, flash floods, droughts, and water management decision making. Ensemble prediction systems allow us to quantify the uncertainty in our forecasts in a meaningful way and to better understand the limits of predictability (Thielen et al. 2009; Rossa et al. 2010; Zappa et al. 2010) and so provide decision makers with longer useful lead times to prepare for the event. The amount and breath of material in this handbook is evidence of the importance that hydrological ensemble prediction systems are playing in various applications in hydrology.

At the same time as the natural and technical sciences improved forecasting technology, the social sciences have helped to understand the human perceptions of risk information and how different actors communicate and respond to hazard, risk, and uncertainty information between peers (horizontally) as well as within a hierarchy (vertically) (e.g., Drobot and Parker 2007; Demeritt et al. 2007). The natural sciences and social sciences communities are united in the realization that uncertainties must be managed rather than ignored or eradicated at all cost (Drobot and Parker 2007), a view that is also shared within the response community (IFRCRC 2012).

Ultimately, hydrometeorological forecasts are used in making decisions. Decision making or planning under uncertain conditions can be a complex, yet everyday, experience that people encounter. *“Uncertainties exist when details of situations are ambiguous, complex, unpredictable, or probabilistic; when information is unavailable or inconsistent”* (Brasher 2011, p. 478). In addition to the uncertainty of whether a predicted future event, e.g., a flood or drought, will occur, forecasters are also faced with the uncertainty an individual person feels about taking the right or wrong decision on a course of action.

Decisions could be urgent ones, i.e., related to immediate emergency response, where life is at risk, or less time-sensitive ones possibly relating to the longer term management of resources or the formulation of policy. The most important requirement is that the decisions are effective. This means they save lives, reduce damage to health, property or quality of life, or improve the management of infrastructure and resources. Hydrometeorological forecasts are just one source of information available to the decision maker who must take all other types of information from many other sources into account in deciding what should be done. Examples of other sources of information could be (i) the demographics of populations at risk for instance their age profile – the very young and very old may be particularly vulnerable or their health/disability status, (ii) hydropower energy demand, (iii) temporal variations in

requirements for irrigation water or public water supply, and (iv) transport traffic on water ways, just to name a few. Some indication of the reliability or uncertainty in each data source is essential to guide their influence on the final decisions. This is straightforward if the information is a direct measurement, e.g., of flows, water levels, or precipitation, for which the measurement uncertainties are well understood. However, it is much more problematic when the information is a forecast and even more so when it involves such complex system and models as for the atmosphere or a river basin. Nevertheless, such information is essential if the decision maker is expected to consistently make good decisions. While forecasters tend to want this information, they find it difficult to quantify any improvement in skill due to having this information, Frick and Hegg (2011).

For hydrometeorological forecasts, based on ensembles of simulations, possibly with multiple models, some uncertainty information is already implicit in the ensembles. The critical issue is how to extract the uncertainty information and how to present it to the decision maker. It is important that the decision maker is not overburdened with unnecessary information nor is expected to do complex analyses in a hurry, particularly in emergency response situations. Many approaches have been suggested, including some based on a cost-benefit analysis (Dale et al. 2014) and others based on a decision scaling method (Turner et al. 2014). Simplicity, clarity, relevance, and trust are thus the ideal characteristics of the way ensemble information is presented to a decision maker. Their resulting actions may be instrumental in saving lives and property, and they may, after an event, be held accountable for the decisions. This serious responsibility has a large influence on their approach to making decisions, often characterized by a risk-averse conservatism (Block 2011) and a tendency towards institutional inertia (Rayner et al. 2005). The different attitudes of the decision maker to costs and losses can lead to suboptimal decisions (Millner 2009) so there is a complex challenge to demonstrate the effectiveness of model-based decision support systems, Moser (2009). Millner (2008) develops a decision model for use with ensembles which incorporates the user's cost-loss profile. Weaver et al. (2013) identifies one barrier to the increased use of climate models as a failure to incorporate information from the decision science and social science disciplines and argue for a paradigm shift towards a multidisciplinary decision support framework. Marshall et al. (2011) maintain that social factors are important in the take-up of climate model information and Renn (2011) identifies the perception of risk information and its social amplification as an important issue. End users' requirements span many factors other than technical model or resolution improvements and include education and training (Wetterhall et al. 2013).

The characteristics of good forecast communication, mentioned above, are:

- (i) **Simplicity and Clarity:** Spaghetti plots are too complex and potentially confusing (Zappa et al. 2013). This has been recognized by forecasters for some time, e.g., WMO (2003). Demeritt et al. (2010) explain the particular issues that decision makers have with spaghetti plots. Stephens et al. (2012) reviews methods of visualization and Bruen et al. (2010) describe the visualization strategies of some operational systems.

-
- (ii) **Relevance:** This depends on the use of the forecast, e.g., emergency response, hazard warning, or water resources management. The timescale may be immediate, medium or longer term, typically ranging from single event scales to climate change impacts on water resources. Actively seeking end user feedback and responding accordingly is a key element of maintaining relevance (Demeritt et al. 2013). Contextually sensitive approaches to presenting information are important (PytlakZillig et al. 2010).
 - (iii) **Trust:** Sector-specific and local sources of information are most likely to be trusted. However, in order for persons to respond correctly to a warning, it is also important that they are familiar with the event, e.g., occupants of a house that has recently been flooded will most likely act differently to a flood warning compared with occupants who have never experienced this threat. The social aspects of warning are very important and must be correctly understood (Drabek 1999). Individual relationships can be important (Lackstrom et al. 2014), and also the use of established networks (Kirchhoff 2013).

Methods of communicating probabilistic forecasts and their use are as many as their applications on the short, medium, and long ranges as well as for local, national, continental, and global scales. A comprehensive review is provided in this handbook by Pappenberger et al. in Part 10 “Ensemble Forecast Application and Showcases.” In addition, Alfieri et al. “*Flash flood early warning based on precipitation-indices: three examples at the European and regional scale*” highlight the growing trend of the hydrological community to increasingly use ensemble prediction systems based on radar, nowcasting, and high-resolution short-term forecasting data or combinations thereof to improve the prediction of flash floods. Both information from regional monitoring networks and the expertise of local flood forecasters are being integrated for better forecasting of location and intensity of the events (cf. Alfieri), which is also important information to translate the hazard forecasts into impact-based forecasting and risk information as described by Wittwer et al. “*Challenges of decision making in the context of uncertain forecasts in France*”(Section 10). Such information is crucial for the translation of hydrological forecasts into actionable warnings which serve as guidance for decision makers, the response community and the public. Bates et al. illustrate in “*Probabilistic Inundation Modelling*” (Section 10) how the recent advances in computing power and high-resolution mapping have made probabilistic forecasting of flood inundation possible, allowing now to quantify the probability of flood occurrences with sufficient advanced warning for taking useful preventive actions. Taking for example the June 2013 floods in Central Europe, Bates et al. demonstrate how such applications can work across scales, e.g., how input from a medium-range, continental system such as the European Flood Awareness System (Thielen et al., Section 10) can be successfully combined with a high-resolution hydraulic model (LISFLOOD-FP) to derive probabilistic inundation maps. Such maps could provide important guidance to decision makers and provide information on the temporal as well as spatial uncertainty on expected flood extent. Thus, the range of examples provided in section 10 shows that ensemble prediction systems have become a trusted and

well-established feature for flood forecasting on the short ranges as well as medium range, for local systems as well as global systems. However, the application of ensemble prediction systems is by far not limited to flood forecasting only but extends for example also to shipping (“*Probabilistic shipping forecast*” by Meissner et al., Section 10), prediction of droughts (“*Seasonal drought forecasting*” by Wood et al., Section 10), hydropower (“*Hydro-power forecasting in Brazil*” by Tucci et al., and “*Ensemble forecasting for hydropower in Canada*” by Boucher and Ramos, Section 10), or generally water resource management.

The chapters in this section of the handbook aim at summarizing at a few key issues which are illustrated with concrete examples: These complement the examples in Section 10 and present a snapshot of methods of communication of EPS in forecasting at a number of locations in three very different continents, Australia, Asia, and the USA. A broad range of challenges, both in climate and computational resources, is represented. A number of different timescales are involved, varying between shorter term flood forecasting for emergency response to longer term streamflow forecasting for water resources planning.

Tuteja, Zhou, Lerat, Wang, Shin, and Robertson explain that because of the very high variability in stream flows in Australia, seasonal forecasts, which are used in practical water resources applications, have a high degree of uncertainty which must be communicated with the forecasts. The focus is on minimizing the risks of misunderstanding as well as on communicating the forecast skill. A Bayesian Joint Probability model is used (Wang et al. 2009) with a 5000 member ensemble (some results are sensitive to ensemble size) generated from an empirical multivariate model of stream flows. A number of different forecast skill metrics were evaluated. There is some variation of forecast skill with season, but the best forecasts are for the latter half of the year when storages are filling.

Hartman describes different approaches to the specific forecast information communicated in the USA. There, the resources to generate ensemble flood forecasts are widely available, and the issue is whether to generate an uncertainty product from the ensembles and deliver this to the end user or to deliver all of the raw, unprocessed, ensemble forecasts to the end user, allowing them to analyze them in whatever way they deem appropriate. Usually the provider of the forecasts is best placed to produce the decision support “tool” for the end user. Use of ensembles in longer range water resources forecasting in the USA is complicated by the attenuating effects of its many reservoirs. The management of (and releases from) each must be included in the simulations. Interestingly, end users often want to see the related input data, such as precipitation and snowpack forecasts, as these convey significant information for the longer term predictions. Considerable emphasis is placed on validating EPS forecasts using hindcasts, e.g., simulating what forecasts the system would have made for past weather. This helps to identify bias in an EPS system, but is computationally demanding.

The following chapter, consisting of a number of forecast communicating examples from around the world, addresses the requirements for communicating ensembles and uncertainty for short-term, medium-term, and long-term applications and decision making. They address the issues of (i) how the information is produced and

used at different scales? (ii) how the specific needs of decision makers, forecasters, and water managers are reflected in the set-up of the early warning systems, and (iii) identifying where ensemble prediction systems can play a useful role and where their application is limited.

Pegram highlights the importance of robust and reliable information for disaster managers when emergency actions such as evacuations are required. Repeated false alarms result in lack of trust between the public and the decision makers which is particularly critical when the time to react is limited, e.g., in urban and flash flood prone areas. He gives the example of the city of Durham in South Africa and explains how ensembles are generated from radar data to quantify the uncertainty in the rainfall and hydrological response to produce reliable forecasts for the decision makers.

While early warning based on radar ensembles is typically limited in lead time from 0 to 6 h, high-resolution EPS from numerical weather prediction can be useful to provide an early warning indication that potentially critical conditions are possible within a time frame of 1–3 days as illustrated in the second section by Raynaud. Although not sufficiently precise in terms of location and timing for local civil protection to take specific actions, such indicators can be useful for local authorities to be prewarned, take precautionary actions, and put staff and equipment on standby.

In particular for trans-national river basins, where typically different agencies and multiple authorities are involved in monitoring, forecasting and decision making, the additional lead times gained through HEPS proves to be important. For instance, *Sprokkereef, Ebel & Rademacher* illustrate how HEPS have been used for the Rhine, a river basin shared by nine countries, for more than 10 years to calculate probabilistic water levels and discharges and they describe how they are communicated and distributed on a daily basis to expert users. The added value of the probabilistic forecasts in comparison to deterministic ones has been demonstrated especially for navigation-related water-level forecast.

Finally, longer term EPS such as monthly and seasonal forecasts find application more in water resources and hydropower management than in emergency response. Being the most common form of renewable energy worldwide, it is not surprising that the optimization of hydropower output is of particular interest for different sectors. Olsson, Alionte-Eklund, Johansson, Lindström, and Spångmyr describe how HEPS are used in Sweden for hydropower optimization. The added value of hydrological probability forecasts as support in decision making to hydropower plant operators is demonstrated, but also the limitations with regard to spatio-temporal resolution are also highlighted.

Hirpa, Fagbemi, Afiesimam, Shuaib, and Salamon highlight the special challenges of translating scientific information into practical operations in developing countries where hydrometeorological disasters have profound impacts on human lives. While, for example, the previous chapters illustrate advanced early warning systems benefiting from high-density observational networks, skillful weather forecasts and ensemble systems as well as state-of-the-art web technologies for communication and information sharing, there are still many places in the world where such levels of sophistication and technology are not yet utilized.

The use and communication of ensemble predictions and uncertainty require a minimum data sharing and IT infrastructure, while for communities without Internet there is a need for a different means of information dissemination such as by phone or through a face-to-face conversation. Trans-disciplinary and multistakeholder partnerships offering global solutions for basic coverage and provision for medium-range forecasts and near-real-time remote sensing detection of events can play a major role in accelerating the process of incorporating science into more effective disaster planning to save human lives and protect the economy.

Demeritt, Stephens, Créton-Cazanave, Lutoff, Ruin, and Nobert conclude this section by investigating the practical challenges of communicating and using ensemble forecasts at the example of operational flood incident management. Based on recent social science research on the variety and effectiveness of visualizing hydrological ensemble prediction, the chapter highlights cognitive and other difficulties experienced by the users of probabilistic forecasts to understand the information correctly. Recognizing that the way uncertainty information is presented has influence on its perception, this chapter illustrates the most common way of communicating EPS to different users with varying level of expertise in probabilistic forecasting. It also highlights that probabilistic information can be misunderstood, for example, by the general public but equally by expert users, if the underlying concept of probability is not fully understood. Dialogue between the producers of HEPS and the users, for example, through repeated discussions or training is identified key action for the correct uptake of the information by the decision makers.

However, even if the information is well communicated and received, there remains the issue that anticipatory actions and decisions based on uncertain forecasts will always be associated with a risk and different mechanisms for an improved management still need to be further explored.

Accepting that there is not a single solution to fit all applications, it emerges that the visualization and the communication of HEPS need to be tailored to the respective end users. This is particularly challenging the less defined the end user community is, e.g., in continental and global systems.

Scientists and engineers involved in hydrometeorological forecasting have realized for some time now that their role is not confined to the generation of forecasts and associated uncertainty information. To be effective and make a real difference in the lives of people, this information must be communicated in a convincing way to the right decision makers whether these are civil protection, water resource managers, or the public. It is the effectiveness of the entire system which must be considered and assessed. Different approaches to what is communicated and how such communication takes place, described above, illustrate how diverse the area can be, how much progress has already been made, and how much more there is to be done. As ensembles are increasingly used in longer term management and policy decisions, the range of end users and their differing requirements can only expand and flexibility and adaptability to individual circumstances will be required.

References

- P. Block, Tailoring seasonal climate forecasts for hydropower operations. *HESS* **15**(4), 1355–1368 (2011)
- D.E. Brashers, Communication and uncertainty management, *Journal of Communication*, **51**(3), 477–497 (2001)
- M. Bruen, P. Krahe, M. Zappa, J. Olsson, B. Vehvilainen, K. Kok, K. Daamen, Visualising flood forecasting uncertainty: some current European EPS platforms – COST731 Working Group 3. *Atmos. Sci. Lett.* **11**(2), 92–99 (2010)
- M. Dale, J. Wicks, K. Mylne, Probabilistic flood forecasting and decision-making: An innovative risk-based approach. *Nat. Hazards* **70**(1), 159–172 (2014)
- D. Demeritt, H. Cloke, F. Pappenberger et al., Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environ. Hazard.* **7**, 115–127 (2007). <https://doi.org/10.1016/j.envhaz.2007.05.001>
- D. Demeritt, S. Nobert, H. Cloke, F. Pappenberger, Challenges in communicating and using ensembles in operational flood forecasting. *Meteorol. Appl.* **17**(2), 209–222 (2010)
- D. Demeritt, S. Nobert, H.L. Cloke, The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol. Process.* **27**(1), 147–157 (2013)
- T.E. Drabek, Understanding disaster warning responses. *Soc. Sci. J.* **36**(3), 515–523 (1999)
- S. Drobot, D.J. Parker, Advances and challenges in flash flood warnings. *Environ. Hazard.* **7**, 173–178 (2007)
- J. Frick, C. Hegg, Can end-users' flood management decision making be improved by information about forecast uncertainty? *Atmos. Res.* **100**(2–3), 296–303 (2011)
- International Federation of Red Cross and Red Crescent Societies, Community and Early Warning Systems: guidelines, (2012), 81pp. <https://www.ifrc.org/PageFiles/103323/1227800-IFRC-CEWS-Guiding-Principles-EN.pdf>
- C.J. Kirchhoff, Understanding and enhancing climate information use in water management. *Clim. Change* **119**(2), 495–509 (2013)
- K. Lackstrom, N.P. Kettle, B. Haywood, Climate-sensitive decisions and time frames: a cross-sectoral analysis of information pathways in the Carolinas. *Weather Clim. Soc.* **6**(2), 238–252 (2014)
- N.A. Marshall, I.J. Gordon, A.J. Ash, The reluctance of resource-users to adopt seasonal climate forecasts to enhance resilience to climate variability on the rangelands. *Clim. Change* **107**(3–4), 511–529 (2011)
- A. Millner, Getting the most out of ensemble forecasts: a valuation model based on user-forecast interactions. *J. Appl. Meteorol. Climatol.* **47**(10), 2561–2571 (2008)
- A. Millner, What is the true value of forecasts? *Weather Clim. Soc.* **1**(1), 22–37 (2009)
- S. Moser, Making a difference on the ground: The challenge of demonstrating the effectiveness of decision support. *Clim. Change* **95**(1–2), 11–21 (2009)
- L.M. PythikZillig, Q. Hu, K.G. Hubbard, Improving Farmers' perception and use of climate predictions in farming decisions: a transition model. *J. Appl. Meteorol. Climatol.* **49**(6), 1333–1340 (2010)
- S. Rayner, D. Lach, H. Ingram, Weather forecasts are for wimps: why water resources managers do not use climate forecasts. *Clim. Change* **69**(2–3), 197–227 (2005)
- O. Renn, The social amplification/attenuation of risk framework: application to climate change. *Wiley Interdiscip. Rev. Clim. Chang.* **2**(2), 154–169 (2011)
- A. Rossa, K. Liechti, M. Zappa, M. Bruen, U. Germann, G. Haase, C. Keil, P. Krahe, The COST 731 action: a review on uncertainty propagation in advanced hydro-meteorological forecast systems. *Atmos. Res.* **100**(2/3), 150–167 (2010)
- E.M. Stephens, T.L. Edwards, D. Demeritt, Communicating probabilistic information from climate model ensembles—lessons from numerical weather prediction. *WIREs Clim. Change*, **3**, 409–426 (2012). <https://doi.org/10.1002/wcc.187>

- J. Thielen, K. Bogner, F. Pappenberger, M. Kalas, M. del Medico, A. de Roo, Monthly-, medium-, and short-range flood warning: testing the limits of predictability. *Meteorol. Appl.* **16**, 77–90 (2009). <https://doi.org/10.1002/met.140>
- S.W.D. Turner, D. Marlow, M. Ekstrom, B.G. Rhodes, U. Kularathna, P.J. Jeffrey, Linking climate projections to performance: a yield- based decision scaling assessment of a large urban water resources system. *Water Resour. Res.* **50**(4), 3553–3567 (2014)
- Q.J. Wang, D.E. Robertson, F.H.S. Chiew, A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resources Research*, **45**, W05407 (2009)
- C.P. Weaver, R.J. Lempert, C. Brown, Improving the contribution of climate model information to decision making: the value and demands of robust decision frameworks. *Wiley Interdiscip. Rev. Clim. Chang.* **4**(1), 39–60 (2013)
- F. Wetterhall, F. Pappenberger, L. Alfieri, H.L. Cloke, J. Thielen-del Pozo, S. Balabanova, J. Danhelka, A. Vogelbacher, P. Salamon, I. Carrasco, *HESS* **17**(11), 4389–4399 (2013)
- WMO, Present and planned configurations of ensemble prediction systems at the National Centers for Environmental Protection (NCEP). Report CBS ET/EPS/Doc.3(6) of WMO Expert Team on Ensemble Prediction Systems (2003)
- M. Zappa, K.J. Beven, M. Bruen, A. Cofino, K. Kok, E. Martin, P. Nurmi, B. Orfila, E. Roulin, K. Schröter, A. Seed, J. Stzurc, B. Vehviläinen, U. Germann, A. Rossa, Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmos. Sci. Lett.* **11**(2), 83–91 (2010)
- M. Zappa, F. Fundel, S. Jaun, A ‘Peak-Box’ approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. *Hydrol. Process.* **27**(1), 117–131 (2013)



Present and Future Requirements for Using and Communicating Hydrometeorological Ensemble Prediction Systems for Short-, Medium-, and Long-Term Applications

Geoff Pegram, Damien Raynaud, Eric Sprokkereef, Martin Ebel,
Silke Rademacher, Jonas Olsson, Cristina Alionte-Eklund,
Barbro Johansson, Göran Lindström, and Henrik Spångmyr

Contents

1	Introduction	1049
2	Requirements for a Flood Nowcasting System Based on Radar Ensembles	1050
2.1	Catchment Model	1051
2.2	Forecast Updates Using the Kalman Filter	1052
2.3	Short-Term Rainfall Forecasting	1053
2.4	Forecasting with the “String of Beads” Model	1054
2.5	The S_PROG Model	1055
2.6	Forecast Model Comparisons	1056

G. Pegram (✉)

Satellite Applications and Hydrology Group, School of Civil Engineering, Surveying and Construction Management, University of KwaZulu-Natal, Durban, South Africa

e-mail: Pegram@ukzn.ac.za

D. Raynaud

Université Joseph Fourier, Grenoble, France

e-mail: damien.raynaud@ujf-grenoble.fr

E. Sprokkereef

Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, The Netherlands

e-mail: eric.sprokkereef@rws.nl; eric.sprokkereef@kpnmail.nl

M. Ebel

Bundesamt für Umwelt, Ittigen, Switzerland

e-mail: martin.ebel@bafu.admin.ch

S. Rademacher

German Federal Institute of Hydrology (BfG), Koblenz, Germany

e-mail: rademacher@bafg.de

J. Olsson · C. Alionte-Eklund · B. Johansson · G. Lindström · H. Spångmyr

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

e-mail: jonas.olsson@smhi.se; cristina.alionte.eklund@smhi.se; barbro.johansson@smhi.se; barbro.johansson@cfk.gu.se; goran.lindstrom@smhi.se; henrik.spangmyr@midvatten.se

2.7	Areal Rainfall Estimates	1059
2.8	Dissemination of Outputs to Disaster Managers	1060
2.9	Communicating Uncertainty to the End Users	1064
2.10	Conclusion on Nowcasting Systems Based on Radar Ensembles	1064
3	Increasing Warning Time for Flash Floods with High-Resolution Ensemble Prediction Systems	1066
3.1	Current Flash flood Warning Systems in Europe	1067
3.2	Using Limited Area Ensemble Prediction System for Extending the Predictability of Flash Floods	1069
3.3	Conclusions for LEPS-Based Early Warning Systems for Flash Floods	1070
4	Medium-Range Ensemble Prediction Requirements for a Multinational and Well-Controlled River	1071
4.1	The International Rhine Basin	1072
4.2	Operational Forecasting in the Rhine Basin	1074
4.3	Experiences in Probabilistic Forecasting	1075
4.4	Challenges and Developments in Probabilistic Forecasting. Towards Actual Predictive Uncertainty	1076
4.5	Conclusions for Medium-Range Flood Forecasting for Controlled Rivers	1079
5	Requirements for Using and Communicating Hydrometeorological EPS Medium-Seasonal Forecasting	1080
5.1	Hydropower, Forecasting and Communication Requirements	1080
5.2	Seasonal (Spring-Flood) Ensemble Forecasting	1081
5.3	Medium-Range (10-Day) Forecasting	1084
5.4	Conclusions and Future Outlook for Ensemble-Based Hydropower Forecasting	1085
6	Summary	1088
	References	1089

Abstract

This chapter provides representative examples for using and communicating hydrometeorological ensemble prediction systems (HEPS) for short-, medium-, and long-term applications. The needs of the specific end users in disaster management, flood forecasting centers, and water management are highlighted.

In the first section, the requirements are presented for a nowcasting system based on radar data designed to provide sufficient lead time for decision makers responsible for urban areas. The generation of rainfall ensembles from radar measurements is described using the so-called string of beads methodology. The aspects required by decision makers for flood management are followed by the technical set-up and constraints.

The second section illustrates how short-term HEPS can contribute to increasing the predictability of flash flood events on a regional scale with complementary indicators to higher-resolution local information systems based on short-term forecasts and observations.

In the third section, the benefits of hydrological ensembles are elaborated for flood forecasting and shipping for a well-controlled, trans-national river basin such as the river Rhine. For several years, medium-range ensemble prediction systems have been explored for flood forecasting in medium to large river basins such as the Rhine.

The last section describes the requirements of HEPS for water management and how they are used at the example of hydropower in Sweden. In snow-dominated hydrological regimes such as in Scandinavia, reservoirs need to be carefully managed in order to have enough capacity for storing spring flood volume, while keeping enough water for securing the required power generation. The chapter concludes with the strengths and the limitations of HEPS for various applications.

Keywords

Ensemble prediction systems · Radar · Trans-national river basins · Decision support

1 Introduction

Ensemble Prediction Systems (EPS) were launched in early 1990s by the US National Meteorological Centre (NMC) and the European Centre for Medium-Range Weather Forecast (ECMWF) (Tracton and Kalnay 1993; Palmer et al. 1993). It was only after they became an integral part of operational weather forecasts in the medium-range that the hydrological community started exploring their benefit also for the prediction of hydrological processes (de Roo et al. 2003; Schaake et al. 2006). Although research studies clearly indicated the potential for improving predictions in the medium range, the translation from research into operational hydrological applications, e.g., for flood forecasting, was slow initially (Clove and Pappenberger 2009). With the development of EPS with different resolutions and lead times, this has changed in the recent years and the use of EPS now spans across a wide range of hydrometeorological applications including short-term flash floods, medium-range riverine floods, and long-terms water and energy resource management. Furthermore, where numerical weather predictions still do not provide sufficient spatial and temporal resolution and skill, alternative ways of constructing ensembles, e.g., from radar estimates, have been developed.

In contrast to using single, deterministic forecasts, EPS allow quantifying the magnitude of the uncertainty in the predictions. This has opened the door to entire new fields of research ranging from the correct representation of probabilities, evaluation of probabilistic forecasts, and the communication of probabilistic results to decision makers. Depending on the situational context, decision makers have different requirements: a mayor responsible for evacuating campsites or urban areas in case of flash flood danger has different time constraints and critical thresholds to consider for decision making than a manager of a hydropower plant responsible for meeting production targets.

It is beyond the scope of a single chapter to describe all requirements for using and communicating hydrometeorological EPS for short-, medium-, and long-term applications. Instead, representative examples for nowcasting, short-term, medium-term, and long-term applications are described illustrating the specific end user requirements, how the requirements are met by using EPS and how the results are communicated to end users and decision makers.

2 Requirements for a Flood Nowcasting System Based on Radar Ensembles

In the year 2000, the Municipality Disaster Management Centre of the city of Durban in South Africa did not have any facility for anticipating floods except from emergency weather reports and forecasts. They typically found themselves reacting to information phoned in by people who had either experienced damage or noticed that flooding was occurring. By the year 2003 they had, in the Disaster Management Centre, a GIS display overlain by real-time images of rainfall measured by radar at 5-min intervals, showing them where the rain was and had fallen. This quantum improvement meant that the people living near rivers had now got the potential for some warning about impending floods. In addition, most knew that the Disaster Management Group was working towards mitigating floods in their area in a proactive rather than reactive way. Major industrial developments had been established in Mgeni Park and in the Mlazi Basin in the Durban Metro. Some of these were strategic industries. With the flood forecasting capability in the Durban Metro Disaster Management Centre, a 6–12 h warning of an impending flood would enable industry to evacuate staff and perform controlled shut downs or take steps to reduce the damage to the sensitive plants.

The requirements of an effective flood forecasting system region, which carry over to other problem sites, were set out based on a definition of what constitutes such a system. The Provincial Umgeni Water expected a model or suite of models that

- Use radar data as precipitation input
- Forecast storm movement and precipitation
- Determine the expected runoff from catchments
- Are easily updateable
- Have a user friendly front end

whereas the Durban municipality expected a system which:

- Is quick and easy to run
- Is easy to update
- Is spatially presented (preferably with a GIS)
- Is able to provide predictions with a good degree of confidence
- Makes it easy to obtain information
- Provides reasonable advance warning of potentially serious events (2–3 h)

Of key importance are reliable real-time measurements of rainfall and streamflow, which can be difficult to obtain in bad weather. It is usually difficult to get people who are in danger to move away from their homes and possessions; therefore, it is crucial that the warnings are reliable, in order to facilitate the development of a relationship of trust between the affected communities and the Disaster Managers – there is a very high cost associated with “crying wolf.”

To obtain a measure of the reliability of forecasts, so that they can be meaningfully transferred to the public, it is important that they are accompanied by confidence bands to indicate their precision. These confidence bands can realistically be obtained from examining the forecasts of historical events and determining how quickly they decay in terms of realistic information. The best way to do this is to determine confidence intervals, generated from ensembles of forecasts, seeded with small initial perturbations. The next step, before issuing warnings to the public, is to translate the technical measures of uncertainty (as measured by the confidence bands) into easy to understand messages, e.g., visually through “traffic light” colors *Green*, *Yellow*, *Orange*, and *Red* bands of alert associated with future periods.

This aspect of measuring uncertainty is typical of other systems, but the alerting measures were left to the disaster managers to interpret, with assistance from the designers of the system. Some of the requirements of the Alerting Procedure have to be balanced with practical requirements due to the limited resources available.

In the following development, first the catchment model which was selected for the flood forecasting system is briefly described, then the methodology of short-term rainfall forecasting using the “String of Beads” rainfall simulation model is outlined. The incorporation of the real-time flood forecasting rainfall-runoff model and the operation of the String of Beads model in parallel forecast streams is described. Finally, the manner in which the system developers responded to the decision makers needs for user friendly interfaces and clear information is reported.

2.1 Catchment Model

The catchment model selected was composed of an arrangement of three interconnected linear reservoirs, parameterized by the reservoir response constants and including a loss term for each reservoir; a complete description of the model is given by Pegram and Sinclair (2002), based on the ARMA-based linear tank model introduced by Pegram (1980). The model is used in a semidistributed form where each of 12 quaternary subcatchments (each approximately 300 km^2) of the Umgeni catchment is treated in a lumped manner. These are forced with a mean areal rainfall input in aggregations of radar estimates of rainfall measured in 5-min intervals. For the purposes of making forecasts, the model is formulated according to its governing state-space equations rather than in the more efficient pseudo-ARMA form. This was to take advantage of the Kalman (1960) and MISP (Todini 1978) filters for online forecast updates. In our application, the Kalman filter updates the model states while the MISP algorithm is a modified version the filter, which provides both state and parameter updates at each time step.

The advantage of recursive estimation techniques (such as the Kalman filter) is that they only require the measurements from the immediate past in order to provide the optimal (least squares) estimates; this makes them ideally suited to online computational situations. A further important benefit is the routine computation of the error covariance matrix, which can provide an estimate of the forecast confidence at each time step.

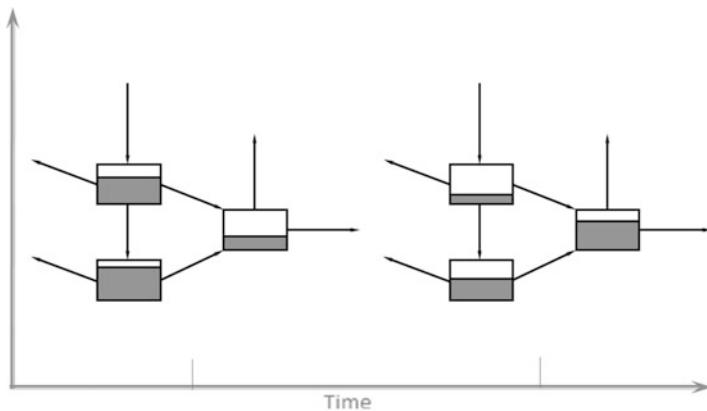


Fig. 1 Model state updates (note the different storage levels at the different epochs)

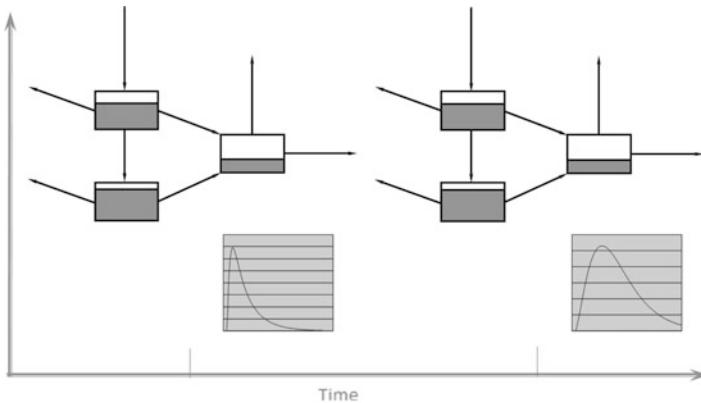


Fig. 2 Model parameter updates (note the same storages, but differing responses)

Figures 1 and 2 indicate the difference between state updates and parameter updates, in the context of the catchment model used in this study. State updates refer to a change of storages within the model's reservoirs, while parameter updates change the way the model will respond to a given set of storages to reflect the changes in the catchments expected response with time.

2.2 Forecast Updates Using the Kalman Filter

The Kalman filter is a recursive state estimation scheme. It produces optimal estimates (in a least squares sense) of a system's state vector x_t , where the system may be defined by the following linear stochastic difference equation (Eq. 1)

$$x_t = Ax_{t-1} + Bu_t + w_{t-1} \quad (1)$$

where A is the state transition matrix, B is the system input conversion matrix, u_t is the exogenous input vector, and w_{t-1} is the system noise, which is assumed to be normally distributed white noise with mean \bar{w} and covariance matrix Q . The system measurement equation is given by Eq. 2

$$y_t = Hx_t + v_t \quad (2)$$

where y_t is the system measurement vector, H is the state to measurement transfer matrix, and v_t is the measurement noise which is assumed to be normally distributed white noise with mean \bar{v} and covariance matrix R .

The derivation of the filter equations can be found in various texts (e.g., Gelb 1974; Young 1984) and is omitted here. The filter equations fall into two categories: prediction and correction equations. The prediction equations provide a priori (forecast) estimates of the system states and state error covariance matrix; the correction equations compute the a posteriori estimates of the system states. The matrices A , B , Q , R & H , the measurements y_t , and the input vector u_t are all assumed to be known until time t ; the unknowns are the states \hat{x}_t and the noise terms w_t and v_t .

The filter is applied recursively with initial estimates of the system states and associated errors being made to start the filter. The prediction equations are used to forecast to the next time step and the correction equations applied once a measurement becomes available. This is illustrated in Fig. 3, along with the filter equations. In the figure, $\hat{x}_{t+1/t}$ is the a priori estimate of the system states, $\hat{x}_{t/t}$ the a posteriori estimate of the system states, $P_{t+1/t}$ the a priori estimate of the system error covariance matrix $P_{t/t}$ the a posteriori estimate of the error covariance matrix K_t the gain of the filter.

2.3 Short-Term Rainfall Forecasting

Rainfall forecasts applicable to flood forecasting typically span a whole range of timescales from seasonal outlooks to several days ahead as well as right down to forecasts for the next few minutes or hours. In order for the longer range forecasts to be significant the uncertainties associated with them need to be reduced by decreasing the spatial resolution at which the forecasts can be made. This unfortunately has the effect of reducing their usefulness for flood forecasting, even for relatively large catchments, for this reason two statistically based spatial rainfall models were investigated as possible candidates to provide short-term forecasts of rainfall fields in the study. These were the “String of Beads” Model (SBM) of Pegram and Clothier (2001) and the “Spectral prognosis” (S_PROG) model (Seed 2001). Numerical weather prediction is not able to provide sufficient quantitative detail at the appropriate spatial and temporal scales for the purposes of this urban flood forecasting system.

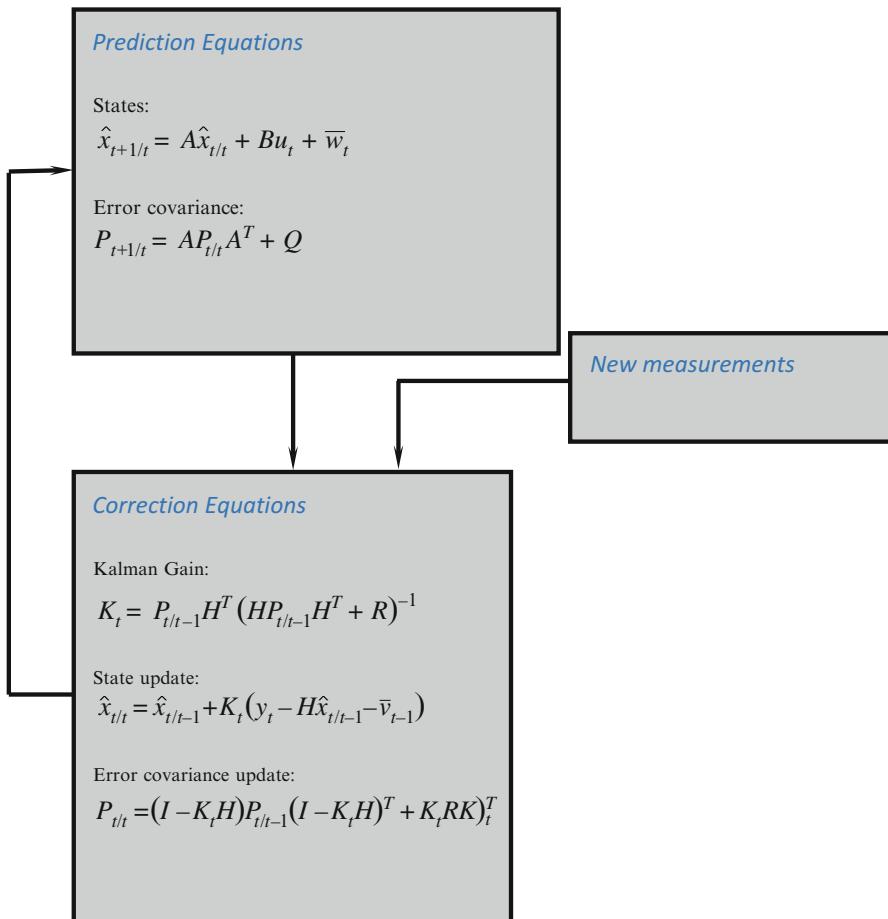


Fig. 3 The Kalman filter cycle

2.4 Forecasting with the “String of Beads” Model

A full description of the String of Beads Model in forecast mode has recently been published by Berenguer et al. (2011). A complete description of the SBM in simulation mode is given by Clothier and Pegram (2002), with a basic outline given here.

A typical simulation begins with the generation of a stack of random noise fields. The values of the noise at each pixel in the field are independent and distributed according to the standard normal distribution with a mean of zero and unit variance. A new field is constructed (pixel by pixel) using an Autoregressive lag 5, AR(5), model linking the new frame to the previous 5 and placed at the first position on the stack, with the existing fields moving backwards one position and the final field falling away (Fig. 4).

A mean field advection vector is used to establish the correct temporal alignment of the pixel values in the field. A warm-up period is required to ensure that the

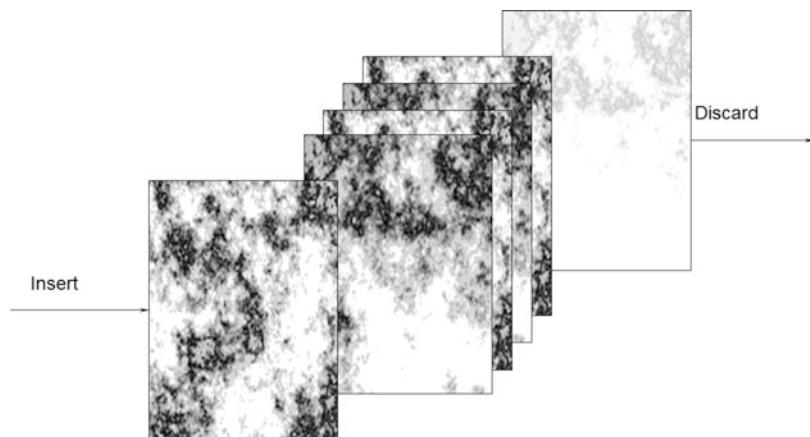


Fig. 4 The SBM noise stack

sequence of fields used in the generation is properly conditioned. Once the stack has been given a sufficient number of recursions to be correctly conditioned, the newly generated field is power-law filtered using the fast Fourier transform to ensure the spatial correlation has the desired shape, then the correcting image scale statistics “Wet Area Ratio” (WAR) and “Image Mean Flux” (IMF) are imposed on the field by a thresholding and scaling process. The resulting field is then exponentiated to produce a field of simulated rainfall rates.

In forecast mode the process is more streamlined. The spatial correlation structure from the observed fields is retained, and the pixel scale development is computed directly from the observed fields. The image scale statistics are forecast using the same bivariate AR(5) process of the simulation mode, but this time conditioned on the mean values of WAR and IMF from the previous five observations. Instead of using Lagrange advection, a sophisticated motion-tracking algorithm (Bab-Hadiashar et al. 1996; Seed 2001) was used to estimate the field advection. A dense grid of advection vectors is computed at each time step and used to advect the forecasts and maintain the appropriate temporal alignment between forecasts. The smoothed advection grid is updated, at each time step, as new information becomes available. The computation of the advection vectors is efficient enough to allow it to be used for real-time applications. Optical flow algorithms that provide comparable (or better) accuracy are generally prohibitively inefficient to compute without specialized hardware components (Camus and Bulthoff 1995).

2.5 The S_PROG Model

The S_PROG model exploits the idea that rainfall fields exist as structures encompassing a range of spatial scales with the persistence of structures being proportional to their spatial scale i.e., larger scale structures have a longer persistence

time than smaller scale ones (Seed et al. 1999). Generation of new forecast rainfall fields is achieved by first disaggregating the observed field into a multilayered hierarchy of fields each of which represents features at different scales. This disaggregation is achieved by applying a notch filter in the Fourier domain to separate the field into a number of spectral components. Each of the fields in the hierarchy is modeled by a dynamically fitted AR(2) process at the relevant space scale and combined to produce the forecasts. The forecast field is conditioned to have the same mean rainfall rate and wetted area as the latest observed image; this is in contrast to the SBM where the evolution of these quantities is predicted using a bivariate AR(5) model.

2.6 Forecast Model Comparisons

Both qualitative and quantitative comparisons were made between the forecasts of SBM and S_PROG. Both models were configured to produce forecasts conditioned on an observed initialization period of a rainfall event measured by the Durban weather radar and the resulting forecasts compared to the observations.

Figure 5 shows one of a set of 4 comparisons of forecasts made using the SBM and S_PROG models as well as the observed sequences from the equivalent times.

The forecasts in Fig. 5 have been made at time steps of approximately 5 min; thus, the forecasts are up to half an hour ahead. The images shown in the figure represent observed and forecast reflectivity fields measured in dBZ. All fields are thresholded below a value of 10 dBZ because the rainfall rates below this level are negligible (approximately 0.15 mm/h).

To illustrate an ensemble of radar forecasts, Fig. 6 gives an example of fine scale (1 km) radar image forecasting showing four members of an ensemble of forecasts using simulated sequences, comparing the expected, a possible growth and a possible decay scenario with the actual sequence. The images are shown at 10 min intervals. The missing quadrant in each scan at lower right was to remove irritating ground-clutter. This image is taken from Clothier and Pegram (2002).

The forecast fields in Fig. 5 were compared to the observed fields in terms of two measures. The first was the square root of the mean of the sum of squared errors over all pixels in the field (RMSSE), while the second was a comparison of the mean spatial difference error, which is closely related to the IMF (average spatial rain rate) in the rainfall rate domain. Figure 7 compares the magnitude of these two measures for the advancing forecast sequences produced by the models.

S_PROG performs better than SBM if RMSSE at the primary pixel scale is the criterion used for the comparison. The RMSSE for S_PROG increases rapidly with forecast time, while the RMSSE for SBM remains fairly constant with time, despite starting at a higher initial value. It makes sense that S_PROG performs better (in this case) since the model is formulated to give an optimum forecast in the least squares sense, with the forecasts being degraded to a mean field as the forecast time increases, resulting in errors with a smaller magnitude. The SBM does not behave in this way since the image scale parameters (WAR and IMF) evolve according to the

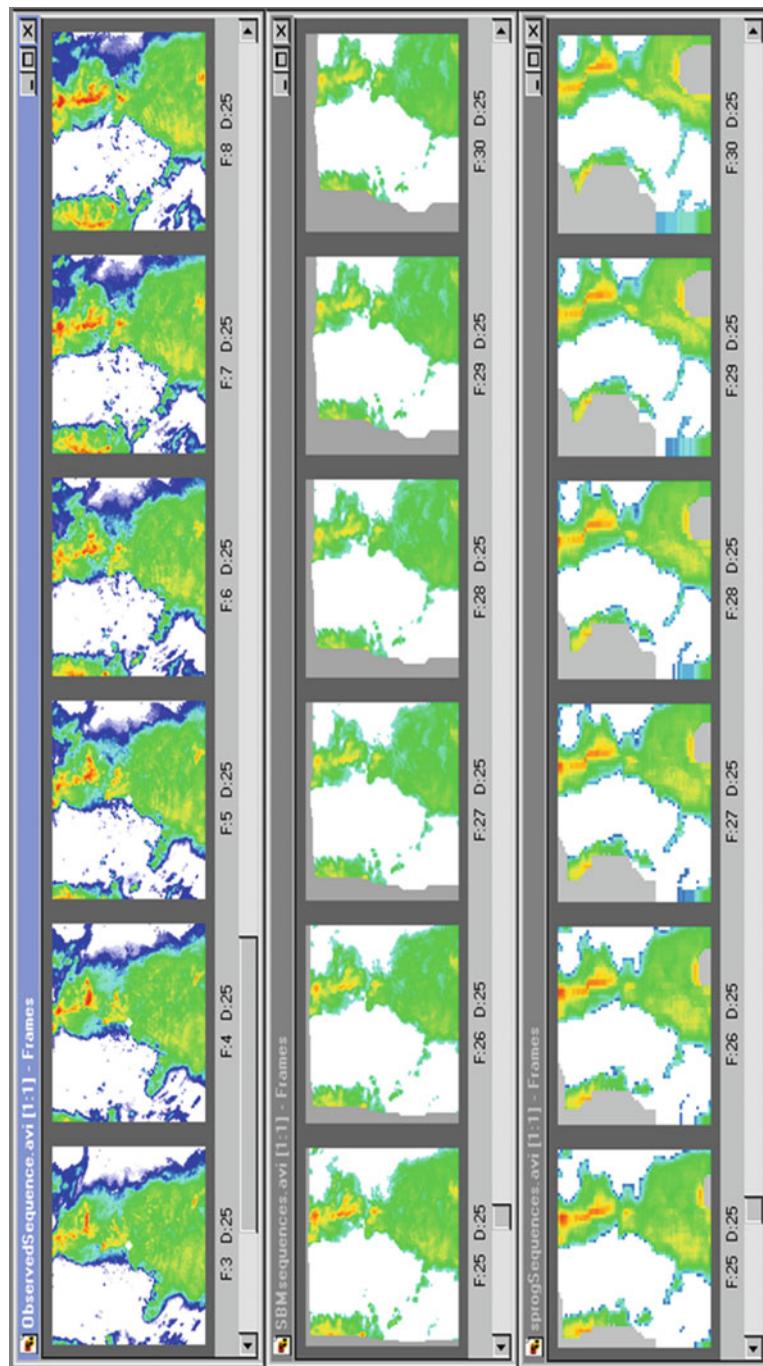


Fig. 5 Comparative forecasts and observed reflectivity fields – *top row* Observed, *second row* SBM forecasts, *third row* S_PROG forecasts – 5 min intervals starting simultaneously

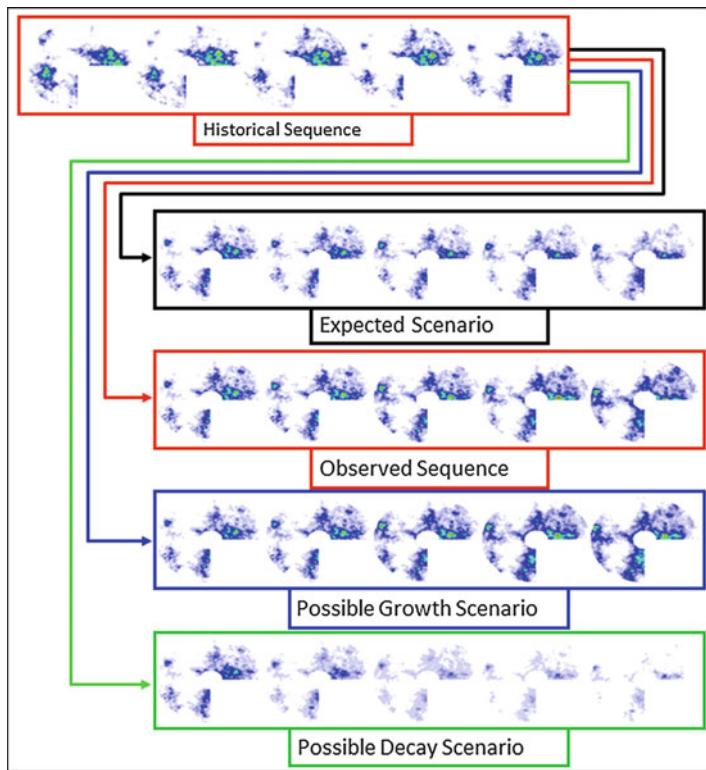


Fig. 6 An example of fine scale (1 km) radar image forecasting showing four members of an ensemble of forecasts using simulated sequences, comparing the expected, a possible growth and a possible decay scenario, together with the observed sequence

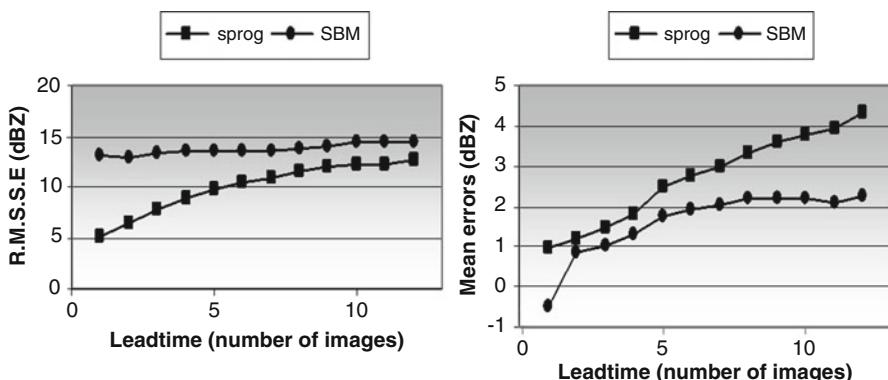


Fig. 7 Root mean sum of squared forecast errors and mean errors at 5-min intervals

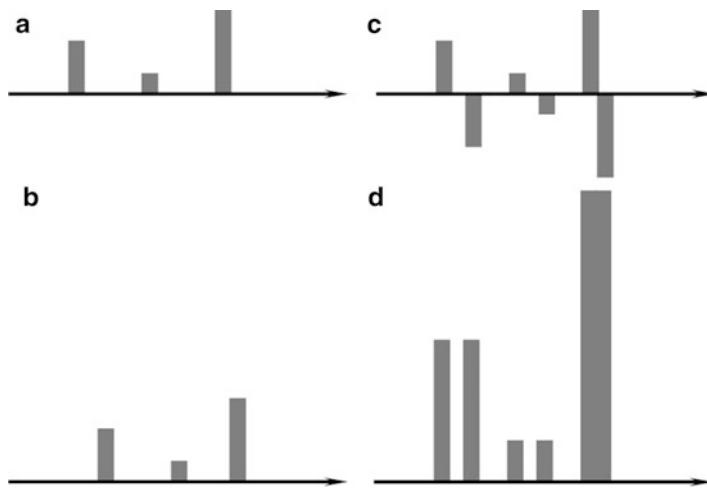


Fig. 8 Sensitivity of RMSSE to the position of peaks; (a) and (b) are time series which are nearly the same in that the same amount of rain fell in both cases, but are staggered by one time step; (c) shows the differences, while (d) shows the squared differences

bivariate AR(5) model, meaning that changes in intensity and wetted area are evident in the forecasts. The RMSSE is very sensitive to the relative positioning of peaks in the field, having a large effect on the squaring of the errors. Squaring the error values removes the effects of sign from the comparison. Figure 8 shows this effect graphically in one dimension. The RMSSE can be misleading in this context; because the right amount of rainfall integrated over subcatchments was our goal (not matching at the pixel scale), the mean errors were considered more appropriate in the model comparison.

The final decision was to use SBM as the forecasting model in our application, instead of S_PROG, as it is better suited to event scale simulation.

2.7 Areal Rainfall Estimates

The first step is to accumulate the spatial rainfall fields over the required time period as described in Pegram and Sinclair (2002). The radar rainfall fields are sampled, instantaneously, at discrete time steps, while the physical rainfall field is evolving continuously with time. This evolution encompasses the growth and decay of the field's structures as well as complex field advection processes. If the sampling interval is not fine enough to capture the dynamic changes in the field, then simple linear accumulation techniques are inadequate and more sophisticated accumulations schemes are required.

Figure 9 shows a 6 h accumulation done using a linear accumulation technique (where the time between radar scans is 15 min), and Fig. 10 the same accumulation period using a morphing approach described by Hannesen

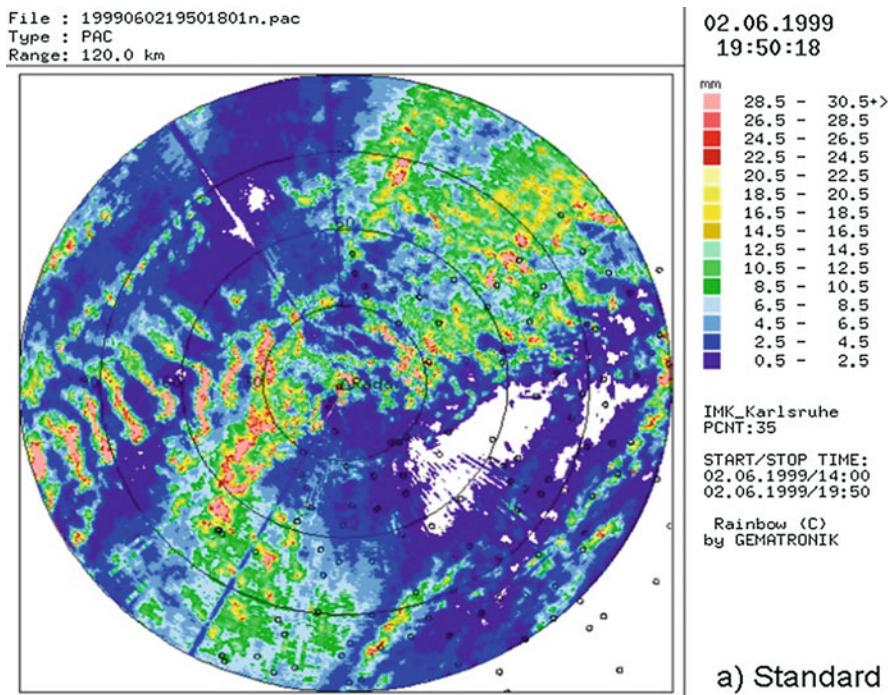


Fig. 9 A crude superimposed accumulation technique (After Hannesen 2002)

(2002). It is obvious that the second approach produces an accumulated field which is what we would expect due to a rainstorm moving through the radar's field of view.

2.8 Dissemination of Outputs to Disaster Managers

The dissemination of outputs refers to the kinds of information presented to the end user of the flood forecasting system, in this case the disaster managers (DMs). The products were agreed upon after consultation with the members of the disaster management team at Durban Metro and Umgeni Water. The consultation process yielded two major requirements from the DM's point of view; some warning should be provided ahead of a likely flood and the affected areas should be indicated (preferably in a graphical format). To meet these requirements, ArcView GIS was chosen as a display tool since the DMs already made use of this system. Images of current and historical rainfall were available for display in the disaster management control room, and dynamically selected flood lines could be called up in response to current observed and forecast streamflows.

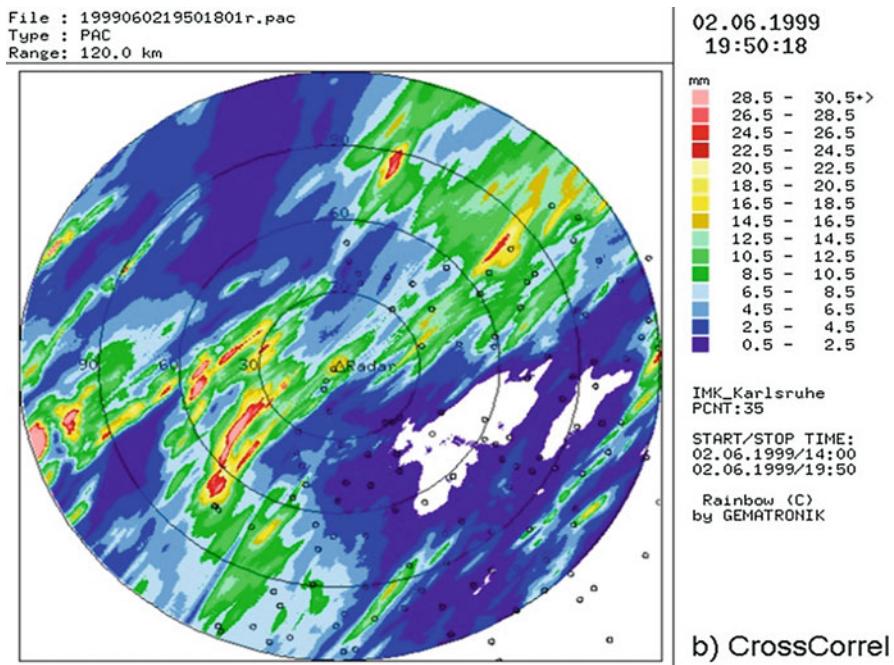


Fig. 10 Cross-correlation tracking and morphing accumulation (After Hannesen 2002)

2.8.1 Integration of the Radar Rainfall Images into a GIS

When radar rainfall images are presented in a spatial context, they can provide a visual advanced warning of heavy rainfall approaching sensitive (flood-prone) catchments. Figure 11 shows an example of an instantaneous radar image of a major storm event that moved over the Mlazi catchment and the lower reaches of the Mgeni catchment.

The images were available in an ArcView compatible format at the Durban Municipality's Disaster Management Centre in near real-time.

2.8.2 Indicating the Flood Affected Areas in a GIS

There was a need expressed by the disaster managers (DMs) to determine which areas they might expect to be affected by floodwaters. It is all very well having a sequence of observed streamflows and possible forecasts of the future flows described in m^3/s , but these numbers usually have no meaning to the DMs. They wanted images interpreting flood levels and depths of inundation showing which areas are to be affected.

The chosen process for producing floodlines was as follows: a Hydraulic model of the river channel and adjacent flood plains was produced, using a Digital Elevation Model (DEM) in combination with field surveys. This is a once off procedure since the same terrain model may be used to hydraulically model the effects of many different

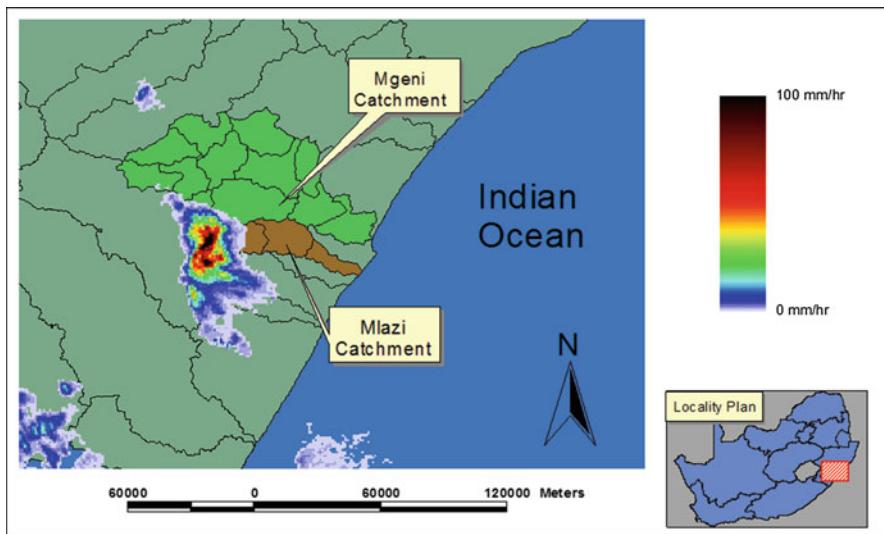


Fig. 11 A major storm over the upper Mlazi catchment (17 December 2002 at 14:58)

flow rates. The hydraulic model used in this study was the well-known HEC-RAS model. HEC-RAS routes the flood-wave down the modeled channel and computes flood levels at each model cross-section. A typical output from HEC-RAS (steady state) is shown in Fig. 12, modeling the river near its mouth to the sea. HEC-RAS can also output the flood levels as a set of points in a three-dimensional coordinate system as a text file. In order to interpret these points as flood lines and inundation levels, an interpolation between channel cross-sections is required, preferably in conjunction with information from a DEM. A separate floodline must be produced for each flow rate considered. A typical one for the 50-year flood in the Umgeni river, 21 km from the river mouth, is shown in plan in Fig. 13.

For the purposes of this pilot study, a simple solution was proposed. Floodlines at various recurrence intervals are available in the ArcView shapefile format for the Umgeni River through previous studies commissioned by the Durban municipality. These served as the starting point and an ArcView script was written which selects the appropriate floodlines to display based on the current observed streamflow and the forecast flow. Then images such as Fig. 11, tracking storms, and Fig. 13, showing computed levels, were mounted in the city's disaster management center and updated automatically as recent rainfall-runoff data were collected and processed.

Figure 14 shows an ensemble of spatially averaged rainfall forecasts over the Mlazi catchment indicated in Fig. 11. These were modeled using the String of Beads Model. For this application, ensembles of future possible riverflows were not computed; however, the methodology is an extension of the above and the reader is referred to Sect. 4 following for a description of the process.

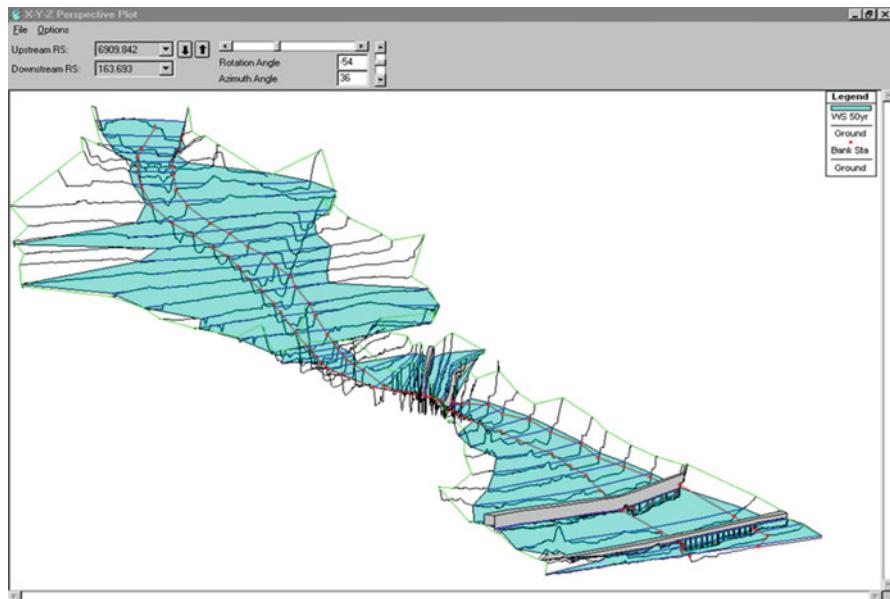


Fig. 12 HEC-RAS output for the lower reaches of the Umgeni River near its mouth, Northeast of the City of Durban, the turquoise surface being the modeled 50-year flood

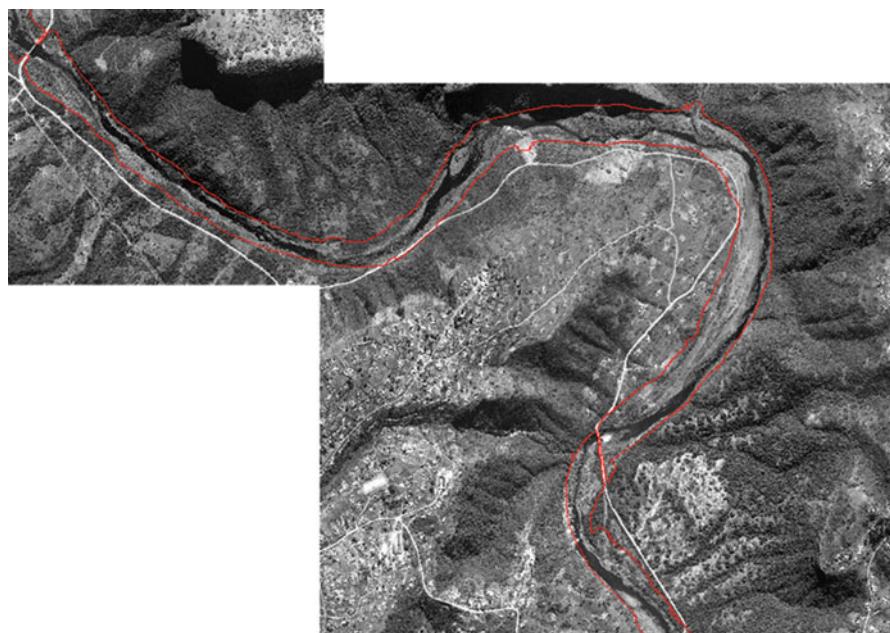


Fig. 13 A 2% Annual exceedance probability (50 year) flood line (Umgeni catchment downstream of Inanda Dam) also using HEC-RAS, upstream of the river in Fig. 12

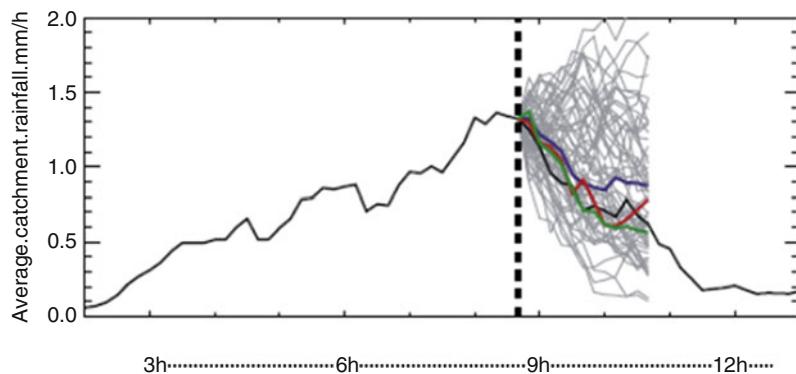


Fig. 14 An ensemble of forecast rainfall over the Umlazi catchment created using the String of Beads Model. Three are colored up to distinguish the individual paths from the others

2.9 Communicating Uncertainty to the End Users

The major aims in communicating flood forecast information to the end users are clarity and simplicity. People in charge of decisions during flood events are not necessarily technical people and are working under stress; therefore, efforts have to be made to convert probabilities into meaningful decision tables or support matrices.

An example of one is offered in Table 1.

The proposed methodology requires the preliminary definition, during a planning phase, of a number of alert levels (blue, yellow, and red) corresponding to bankfull (blue), inundation of property (yellow), and potential loss of life (red). The decision to take (or not to take) action is derived as a function of (1) the alert level, (2) the likelihood of the event, and (3) the reliability of the forecast. The likelihood of the event is defined as a function of the probability of exceedance (Very Low, Low, Medium, High, Very High) of each specific alert level, while the reliability (Very Low, Low, Medium, High, Very High) of the forecast is defined in terms of its coefficient of variation (CV). The probability of exceedance is computed as a function of the probability of a future value conditional to all the prior knowledge one can establish, including the actual forecast of the flood forecasting model.

The rationale for this approach is that on the one hand it becomes critical to take the right decision when the Red level is reached and on the other hand one has to be more careful when the uncertainty of the forecast is low.

2.10 Conclusion on Nowcasting Systems Based on Radar Ensembles

This section set out to describe the main factors to keep in mind when designing a flood forecasting system for small to medium-sized catchments. For this purpose, weather radar is an indispensable tool because the images are available in near real

Table 1 An example of a flood decision support matrix

		Likelihood				
Reliability	Blue	Very low	Low	Medium	High	Very high
	Very low	NA	NA	NA	NA	NA
	Low	NA	NA	NA	NA	NA
	Medium	NA	NA	NA	NA	A
	High	NA	NA	A	A	A
	Very high	NA	NA	A	A	A
		Likelihood				
Reliability	Yellow	Very low	Low	Medium	High	Very high
	Very low	NA	NA	NA	NA	NA
	Low	NA	NA	NA	NA	NA
	Medium	NA	NA	A	A	A
	H	NA	A	A	A	A
	Very high	NA	A	A	A	A
		Likelihood				
Reliability	Red	Very low	Low	Medium	High	Very high
	Very low	NA	NA	NA	NA	NA
	Low	NA	NA	NA	A	A
	Medium	NA	A	A	A	A
	High	A	A	A	A	A
	Very high	A	A	A	A	A

time. In addition, models like the String of Beads can be used to provide ensemble forecast of rainfall in near real time starting within 5–10 min ahead of their assimilation into the system. This is in contrast to ensembles available from Numerical Weather Prediction models which have delays of the order of hours. With all the sophistication of the process, this needs to be seamless and invisible to the disaster manager who is interested in the forecast outcome, not the math. The key words are *speed* and *simplicity*, for the amelioration of the fearful damage and loss of life caused by heavy rainfall.

3 Increasing Warning Time for Flash Floods with High-Resolution Ensemble Prediction Systems

Flash floods can be generated by a variety of different processes or combination of processes that are often influenced by very small scale and local features. For example, flash floods can be the result of short-term but intense precipitation but can equally be triggered by moderate rainfall combined with snow melt. Rainfalls also can suddenly be enhanced by local features related to orography, e.g., convection through sun facing slopes (e.g., Thielen and Gadian 1996) or urban heat islands (e.g., Bornstein and Lin 2000; Kusaka et al. 2014; Thielen et al. 2000; Thielen and Gadian 1997), which can lead to extreme rainfall rates or stationary regeneration of storm cells. In principle, the small scales of most of the meteorological processes involved in flash flood triggering can be resolved to some extent by high-resolution numerical models (Younis et al. 2008; Alfieri et al. 2011). With steadily increasing computing power, there is also an increasing number of operational numerical weather prediction models with grid spacings well below 10 km and thus approach flash flood relevant scales (e.g., Stephan et al. 2008; Rotach et al. 2009).

However, even with such high-resolution operational numerical weather prediction models, uncertainty about the exact intensity, timing and location of precipitation remains high. More importantly, even if meteorologists were able to forecast rainfalls accurately, predicting flash floods would remain difficult because the efficiency of the runoff processes in the catchments play an important role as well – moderate precipitation on totally saturated soils can lead to flash flood generation as well as heavy precipitation on unsaturated soils. Therefore, in order to forecast flash floods, both the meteorological rainfalls and the runoff processes in the catchments need to be well described (UCAR 2010).

Due to the difficulty in predicting rainfalls sufficiently skillful at the local scales, flash flood predictions are often based on observations through in situ networks, radars, lidars, or other remote sensing techniques which have the drawback that the lead times are very short as illustrated in the previous section. Efforts in extending the lead time are made by blending remote sensing fields with very short-term forecasts allowing to extend the predictability from 1–3 up to 6 h (e.g., Germann et al. 2009; Liguori et al. 2012; Leijnse et al. 2007; Kober et al. 2012). Such systems provide high-resolution and high-quality rainfall fields that can be applied in hydrological models specifically developed for certain catchments. The estimation of uncertainty in the estimation and prediction of the rainfall fields is increasingly taking into account (Villarini et al. 2010).

However, while flash floods are local phenomena, they can occur almost anywhere and a shift in rainfall of only a few hundred meters can lead to flash floods in a neighboring catchment where a comprehensive data collection system, rainfall-runoff model, or flash flood guidance system may have not been set up. In case the other catchment belongs to another authority or even country, possible prewarning information may be entirely lost because information is not shared and bi-lateral communication protocols may not have been established. This situation could be improved with a regional flash flood early warning platform which integrates products based on

numerical weather prediction which are available over large spatial areas with specific, local information, e.g., from radar and in situ measurements. In this section, the requirements for such a seamless regional flash flood forecasting system are being laid down at the example of the Mediterranean, a region particularly prone to flash floods.

3.1 Current Flash flood Warning Systems in Europe

Few operational early warning systems for flash flood are documented in Europe. Table 2 summarizes locations, geographical extents together with some general properties such as their spatial resolution and forecasting methodologies. Most likely other systems exist, but the lack of publications or information available online or from other sources make it difficult to establish an exhaustive state of the art.

The used methodologies vary a lot from one early warning system (EWS) to another. Some of them are only based on rainfall analysis and do not perform any hydrological simulation. They can be used for greater lead times, but their accuracy is often limited by the lack on hydrological processes like for the RISKMED warning system (Savvidou et al. 2009). The Extreme Rainfall Alert (ERA) developed in UK is only based on rainfall threshold exceedance of some given return periods events and for storm durations from 1 to 6 h. Thus, it is mostly design for urban and flash floods. It gives a probabilistic assessment of the risk by taking into account the positional uncertainty of the quantitative precipitation estimate from the high-resolution UK Met Office 4 km model. The first evaluation of the warning systems proved that ERA is better at timing the possible rainfall threshold exceedance rather than for effectively forecast the flood occurrence (Hurford et al. 2012). The second type of systems uses hydrological models to assess the flash flood threat sometimes with a very high resolution (Verzasca EWS: Liechti et al. 2012), but the heavy computation time required do not make them suitable for forecast more than a few hours in advance or for spatial domain larger a few hundreds kilometers. A third approach is adopted by the FLASH system which is completely different from the others as it assumes a direct link between rainfall intensity and the electric activity of the associated thunderstorms (Llasat et al. 2010). For this warning system, the lead time of prediction is again very short as it relies on the triggering and observation of thunderstorms.

The rainfall data chosen as input also change from one system to another. Most of the nowcasting systems take the radar rainfall images as meteorological information (Verzasca EWS, AIGA), while the others use the quantitative precipitation estimates (QPE) from the numerical weather prediction (RISKMED, EPIC). The PFFGS by using blending methods between the two types of inputs develop an innovative technique (Atencia et al. 2010).

Ensemble prediction techniques are being explored for hydrological forecasting since about a decade in projects such as the European Flood Forecasting System (FP5 Research project), through initiatives such as the Hydrologic Ensemble Prediction EXperiment (HEPEX), or the COST 731 project (Rossa et al. 2010; Zappa et al. 2010) and many others (Croke and Pappenberger 2009). Among the systems

Table 2 General characteristics of the current flash flood early warning systems in Europe

Name	Domain	Resolution	Lead-time	Methodology
Verzasca EWS	Verzasca watershed Switzerland	500 m	Nowcasting	Hydrological simulations
GFWS	Guadalhorce watershed, Spain	1 km	Nowcasting	Meteorological and hydrological warnings
AIGA	South East of France	1 km	Nowcasting	Meteorological and hydrological warnings
EHIMI	Catalonia	1 km	Nowcasting	Meteorological and hydrological warnings
ERA	UK	4 km	120 h	Meteorological warnings
Flood-PROOFS	Vall d'Aosta (Italy)	—	60 h	Hydrological warning
FLASH	Whole Mediterranean	≈6 km (accuracy of the lightning detection system)	3 h	Lightning based
PFFGS	Catalonia	1 km	6 h	Meteorological and hydrological warnings
PIEDMONT EWS	Piedmonte region, Italy	Depends on lead time From 1 km with radar images to 7 km with QPF	48 h	Hydrological simulations
RISKMED	South of Italy (except Sicily), Western Greece, Malta, Cyprus	7 km	72 h	Meteorological warnings
EPIC	European part of the Mediterranean	1 km (7 km pour the input rainfall data)	132 h	Meteorological warnings

listed above, four already explore the technique for flash floods. The Verzasca EWS uses a new technique called NORA, an acronym for Nowcasting of Orographic Rainfall by mean of Analogues, which built 25 possible rainfall forecasts up to 1 h in advance thanks to the radar data archive and some similar meteorological situations (Panziera et al. 2011). The GFWS creates 16 possible rainfall inputs thanks to the a radar rainfall ensemble (developed during the IMPRINTS European project) up to 2 h with 1 and 10 min spatial and temporal resolution. The Flood-PROOFS

forecasting system transforms a low-resolution deterministic forecast into a probabilistic high-resolution one thanks to downscaling methods.

3.2 Using Limited Area Ensemble Prediction System for Extending the Predictability of Flash Floods

Currently, a quasi-European flash flood indicator named EPIC, an acronym for European Precipitation Index based on Climatology, is included in the European Flood Awareness System (EFAS, see chapters ► “[Flash Flood Forecasting Based on Rainfall Thresholds](#)” and ► “[Medium Range Flood Forecasting Example EFAS](#)” in this handbook). EPIC is designed to detect catchments with a high probability for heavy precipitation leading to flash flooding within the upcoming 5 days (Alfieri et al. 2011, 2012). One drawback of EPIC is that it is a purely rainfall-driven indicator and thus not capable of capturing flash flooding generated by other processes such as snow melt or the combination of saturated soils with moderate rainfalls. However, many studies have shown that certain geomorphologic features of the basin as well as initial soil moisture should not be neglected when assessing flash flood potential (Penna 2011). EPIC has thus been modified to take into account catchment properties and soil moisture in order to be more accurate in the identification of areas where flash floods can occur in the next 5 days. The new indicator, named ERIC for European Runoff Index based on Climatology, takes advantages of EPIC’s simple methods and introduces a dynamic runoff coefficient to account for the missing hydrological contribution to flash flood triggering in EPIC.

Equation 3 illustrates how the upstream runoff is computed in ERIC. It sums up at each cell of the 1 km river network the rainfall data falling on the whole upstream area are for durations of 6, 12, and 24 h after weighting each of these contributions with a dynamic runoff coefficient.

$$UR_{d_k}(t) = \frac{1}{N} \sum_{i=1}^N C_{f_i}(t) \times [P'_{d_k, i}(t)] \quad (3)$$

where $d_k \in \{6, 12, 24\}$, duration chosen for the rainfall accumulation

N number of grid cells in the upstream area

$P'_{d_k, i}$ Rainfall accumulation within d_k in over the grid cell i

C_f is the runoff coefficient

This coefficient is derived from rainfall–runoff relationships for different initial soil moisture. They have been determined thanks to the distributed hydrological model LISFLOOD which runs operationally over Europe as part of the EFAS. This model also provides twice a day the initial soil moisture information necessary to compute the simulated upstream runoff. Then the novel indicator, ERIC, is computed similarly to

EPIC by comparing the current upstream runoff to the mean annual maxima determined thanks to the COSMO-LEPS and LISFLOOD climatologies (Eq. 4).

$$\text{ERIC} = \max_{\forall d_k} \left(\frac{UR_{d_k}}{\frac{1}{M} \sum_{j=1}^M \max(UR_{d_k})_j} \right) \quad (4)$$

where $M = 19$ number of years of the climatology

$\max(UR_{d_k})_j$ Maximum value of during year j

Results from a 1-year analysis showed that ERIC correlates well with reported flash flood events after selecting a suitable threshold for return periods of upstream runoff, proving that merging meteorological EPS with hydrological features with simple methods helps providing more accurate flash flood forecast over large domains (Raynaud et al. 2014). The threat score defined by the number of hits divided the sum of hits, false alarms and missed events, has been improved from 0.34 for EPIC to 0.5 for ERIC indicator. Moreover, many flash flood events were detected by EPIC from a few hours to 1 day in advance, whereas ERIC was able to locate the flood threat over several consecutive forecasts increasing the confidence the warning that would have been issued.

One example is the torrential rain that hit Halkidiki in Greece on February 10, 2010, and triggered rapid flooding and mudslides. EPIC detected the event a day in advance, but ERIC already showed significant return periods of forecast upstream runoff over the region for lead time exceeding 3 days. Figure 15 presents the outputs of ERIC's forecast 2 days before the event. The reporting points correspond to river sections where the 20-year return period of upstream runoff should be exceeded. For this event, COSMO-LEPS had difficulties catching the intensity and the location of the events with sufficient lead time, but the wet conditions of soils at this time explain why ERIC successfully detected a potential flood.

3.3 Conclusions for LEPS-Based Early Warning Systems for Flash Floods

Short-term EPS have been demonstrated to be useful for the calculation of flash flood indicators over a wide range of spatial scales, ranging from local to quasi-continental scales. Although the temporal and spatial uncertainties are large, useful indication on the possibility of flash floods can be achieved with lead times of 24–72 h. Although being uncertain, awareness can be raised within the responsible authorities for potentially threatening situations in a certain area with a lead time of 24–72 h. Ideally, the early warning information is complemented with higher-resolution systems based on nowcasting, blending, remote sensing, or in situ observations, like the one described in the first part of this chapter. As the

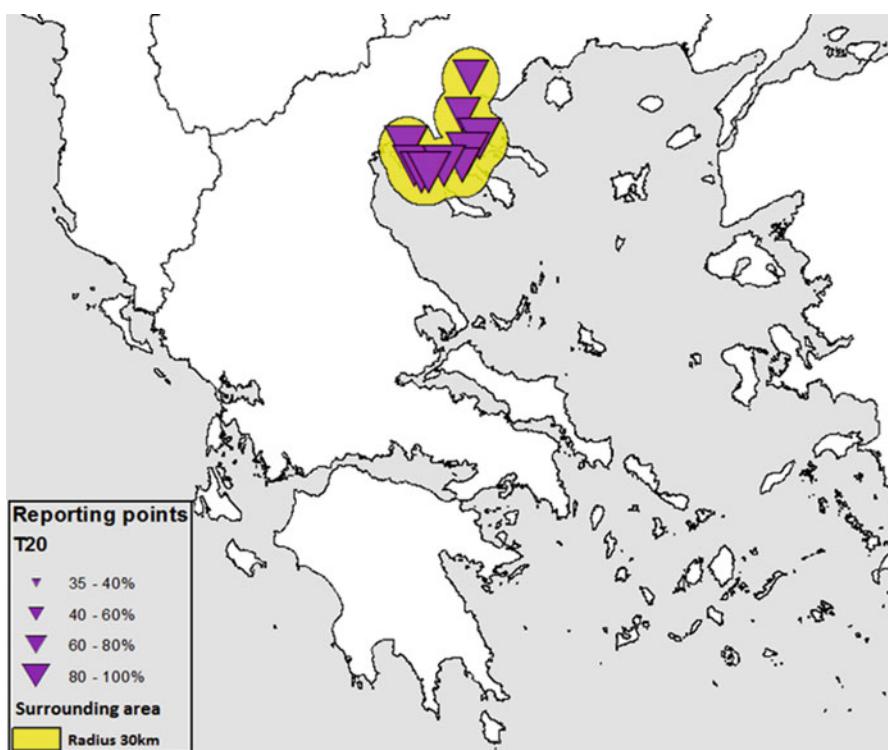


Fig. 15 Two days lead time flash flood forecast of ERIC indicator for the flash flood event of February 10th 2010. The reporting points represent river sections where the 20-year return period of upstream runoff is exceeding by at least 35% of COSMO-LEPS members

forecasted events draw nearer such systems then allow quantitative estimates of the severity, duration and location of the event and provide guidance for the decision makers.

The building of such “staggered warning information flow” with systems providing information with different accuracies for different levels of preparedness actions is very well accepted in the response community (IFRCRCS 2009).

4 Medium-Range Ensemble Prediction Requirements for a Multinational and Well-Controlled River

Flood damage can generally be limited by taking timely precautionary measures. However, this depends on the availability of reliable information about the course of the flood with sufficient lead time. While the mathematical models currently available enable a good understanding of the runoff in rivers, operational hydrological forecasts of future events remain per definition uncertain. Although knowledge

about hydrological behavior of river basins is growing, and model input data are abundantly available, there remain several significant sources of uncertainty. The initial conditions of the catchment at the start of the flood period are not known precisely, nor is it precisely known how much rain or snow will fall when and at which location. Furthermore, even sophisticated operational models used today are still a simplification of reality and are unable to precisely forecast all relevant hydrological processes in the catchment.

In this section, the methods of including ensemble prediction system data for real-time estimation of hydrological uncertainty in the river Rhine forecasting system are described. First the Rhine basin, a catchment which constitutes six different country borders, is characterized. In the following part, several forecasting systems for the Rhine and the experiences with ensemble prediction flood forecasting are described. Finally the challenges and developments in probabilistic forecasting towards actual predictive uncertainty in a connected multinational forecasting environment are discussed.

4.1 The International Rhine Basin

With a length of 1,239 km and a catchment area of 185,000 km² (CHR 2010), the Rhine belongs to the larger rivers in Europe. On a global scale, however, the Rhine does not even belong to the 100 largest rivers of the world and is no more than a medium-sized river basin. Nevertheless, the Rhine is well known all over the world. This is partly due to the variety of attractive scenery in the river basin, for instance, the famous waterfall of Schaffhausen in Switzerland and the Loreley in Germany, are famous tourist attractions. More importantly, the Rhine is one of the busiest shipping routes of the world, with a navigable length of more than 800 km.

Approximately 60 million people live in the Rhine basin, which is shared by nine European states (Kalweit et al. 1993). Larger areas belong to Germany, France, The Netherlands, and Switzerland. Medium-sized and smaller areas belong to Austria, Luxemburg, Italy, Liechtenstein, and Belgium (Fig. 16).

The Rhine rises in the Swiss Alps at a height of 2,345 m. The river flows through the Alpine foreland, the German uplands, as well as the German and Dutch lowlands. The Rhine is the only river connecting the Alps to the North Sea and the Atlantic Ocean.

The Alpine Rhine and the major tributaries rising from the high mountains show a rather uneven discharge regime with large variations between low and high discharges. The discharge regime of the Rhine downstream from the Alpine region is more steady, partly due to the hydrological conditions, but also as a result of human interference. In the Alpine region, the ratio between the highest and the lowest discharge is about 1:70, whereas in the downstream part of the basin this ratio drops to 1:20 or even less.

The Alpine part of the basin contributes a varying amount to the discharge at the outlet depending on the time of year. In average, almost half of the discharge originates from the Alps. In summer, this contribution may rise to over 70%; in



Fig. 16 The Rhine basin

winter it will drop to less than 30% due to the fact that winter precipitation in the Alps is mainly stored as snow.

In comparison to the average discharge, the fluctuation in the discharge downstream from Lake Constance is moderate. In the Alpine region low flows occur in winter and floods in summer. From Lake Constance to the Rhine Delta, the temporal appearance of high and low discharges reverses. Traveling downstream, the

influence of the German upland tributaries becomes more important, leading to higher winter discharges and lower summer discharges.

4.2 Operational Forecasting in the Rhine Basin

Operational flood forecasting is an essential part of flood protection. Information on expected water levels is an important basis for flood management.

In Switzerland the Federal Office for the Environment is responsible for hydrological forecasting for the major rivers and lakes in the Swiss Rhine catchment. Every day at least one set of forecasts for all important gauges in the catchment are published. These forecasts are based on different numerical weather prediction models (NWP) in combination with different hydrological models. Currently COSMO-2, COSMO-7, and ECMWF medium-range forecasts provided by Meteo Swiss are implemented into the Swiss Forecasting System. New model results are available between two and eight times per day, depending on the lead time (from 33 h to 10 days) and spatial grid resolution (between 10 km and 2 km). In critical periods, hydrological forecasts are calculated and published as often as necessary and supplemented by written warning reports. Challenges include not only the correct measurement, interpolation and distinction of regional rainfall and snow accumulation in the alpine topography, but also the correct calculation of snow and glacier melt, as well as the consideration of anthropogenic influences such as large dams and hydropower stations or the runoff regulation of lakes and reservoirs.

Following two major floods in 2005 and 2007, causing major damage to larger parts of Switzerland, the need became clear for more physically based models with a high spatial resolution to improve the quality of forecasts especially in small and medium-sized alpine catchments. Several new models have been set up, calibrated, and implemented into the Swiss forecasting system for all Swiss catchments. The use of different hydrological models enables the forecaster to evaluate model uncertainty in operational real-time forecasts.

In Germany, the 16 Federal States are responsible for the flood forecasting service within their own territory. In the German Rhine basin, competence is shared between six Federal States. Some of them run their own Flood forecasting center, other share the responsibility. Hence, there are several forecasting centers in the German Rhine basin with a variability of used systems, models, and publishing times. Most of them use forecasting systems which are a combination of hydrological models and hydraulic models driven by different NWPs provided by the German Weather Service. Nationally and internationally, the centers are well connected – the more upstream centers supply forecasts to the downstream ones. Although until now only a few use EPS-based systems, the discussion about how the upstream-downstream chain will work for EPS already started.

In the Netherlands, flood forecasting for the river Rhine is within the jurisdiction of the Dutch Water Management Centre, which is part of the Ministry of Infrastructure and Environment. Under normal conditions, the Centre publishes daily water-level forecasts for the gaging station Lobith on the German-Dutch border, mainly

focusing on the navigation on the Rhine. In flood periods, the frequency of forecasts increases to a maximum of four times a day. Flood forecasts for the Rhine are published when the water level at the Lobith gauge rises to 14 m above mean sea level and when a further rise to at least 15 m is expected, a situation that occurs twice a year on average.

Until January 1999, a statistical model based on multiple linear regression was used for daily forecasts as well as for flood forecasting. After the large floods in 1993 and 1995, it was clear that a new more physically based modeling approach was necessary, which led to the development of the current forecasting system RWSoS Rivers. The system is a combination of hydrological (HBV) and hydraulic (Sobek) models, developed for short- as well as for medium-range forecasts (4–14 days). It uses multiple weather forecasts, both deterministic and probabilistic.

4.3 Experiences in Probabilistic Forecasting

In Switzerland, the combination of hydrological forecasts and probabilistic meteorological ensemble forecasts was extensively tested and evaluated within the research project MAP D-Phase (Arpagaus et al. 2009). Promising results and positive end user feedback led to the decision to include this approach also in the daily operational runoff forecast. The probabilistic forecast is currently based on the COSMO-LEPS ensemble (Marsigli et al. 2005), which is a local area model ensemble, developed and run by the COSMO consortium. This ensemble is a 16 member high-resolution ensemble model with a spatial resolution of 7 km and a lead time of up to 132 h.

The operating hydrologist usually discusses the spread and reliability of the current probabilistic NWP with a meteorological forecaster, as is the case with purely deterministic forecasts. In many cases, and especially in critical flood situations, there is a large variation in the forecast results of the different deterministic NWP. The spread of the probabilistic forecasts gives a good impression of the current predictability of the weather development in the next 3–4 days and the variability of the meteorological input to be expected (Fig. 17). Nevertheless, there are many situations when even a wide spread of the resulting runoff calculated by the different ensemble members cannot guarantee that the actual runoff lies between the limits of the calculated forecasts. Another problem for flood forecasting is the long calculation time needed by the ensembles and the rather low updating cycle. The initialization date of the COSMO-LEPS data is more than 11 h old when it is sent to the forecasting center. As the current update cycle is 12 h, only two probabilistic forecasts per day are available.

Nonetheless, including the probabilistic NWP in the hydrological forecast has considerably facilitated the process of understanding and evaluating the uncertainty of a forecast situation. All forecasts are published on the Internet, including those based on probabilistic NWP. Even though the interpretation of the results require some training and knowledge, the reactions of the end users are generally very positive.

In Germany and the Netherlands, probabilistic water level and discharge forecasts for the River Rhine have been calculated for more than 10 years (Renner et al. 2009).

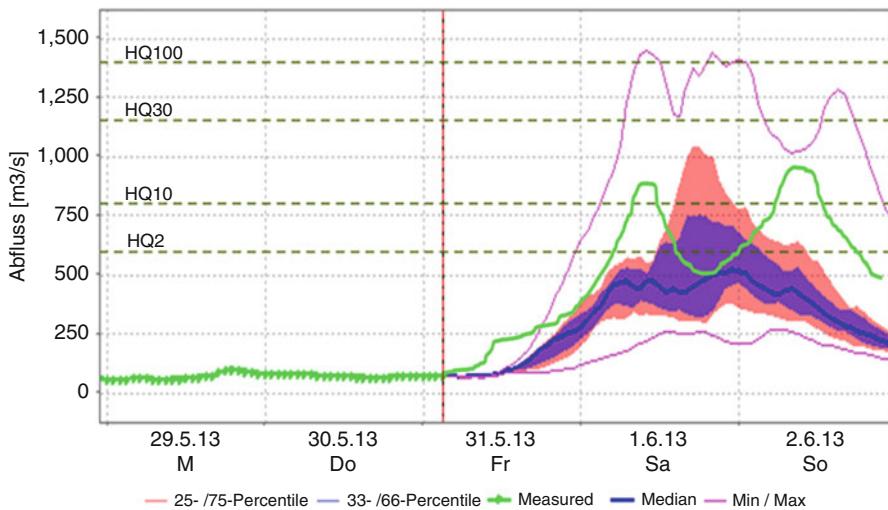


Fig. 17 Runoff forecast for the gauge Andelfingen, Thur, based on COSMO-LEPS forecast

In The Netherlands, the probabilistic results are distributed on a daily basis to professional users only. Currently different meteorological forecasting products are being used to estimate the uncertainty in water-level prediction caused by the numerical weather prediction. These include both deterministic forecasts and ensemble forecasts at different temporal and spatial resolutions. To determine the bandwidth in meteorological uncertainty, the ECMWF-EPS ensemble and the COSMO-LEPS ensemble are used.

The ECMWF Ensemble Prediction produces 15-day probabilistic forecasts daily at 00 and 12UTC (Buizza et al. 1999, 2006, 2007). Since 2010, the EPS probabilistic forecast has been based on 51 integrations with approximately 32-km resolution up to forecast day 10 and 65-km resolution thereafter. An impression of the resolution of the ensemble can be seen in Fig. 18.

Both meteorological ensemble products are run through the combination of hydrological HBV and hydraulic Sobek models. The model output is error corrected by a statistical postprocessing, which then leads to an ensemble water-level prediction.

As a result, all 51 ensemble member forecasts are shown. This Figure displays the 25% and 75% percentile forecast and the median.

4.4 Challenges and Developments in Probabilistic Forecasting. Towards Actual Predictive Uncertainty

Probabilistic forecasts do provide added value in comparison to deterministic ones, especially for navigation-related water-level forecast (see chapter ► “[Probabilistic Shipping Forecast](#)” by Meißner and Klein within this handbook). But they must be

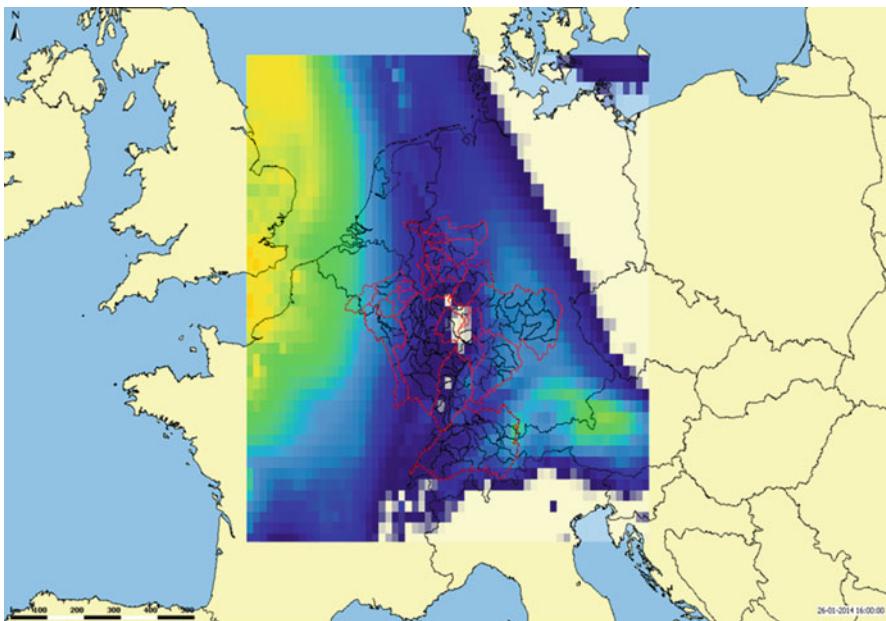


Fig. 18 Precipitation forecast from ECMWF ensemble prediction system

improved to guarantee reliable flood forecasts. A more rapid updating cycle and faster availability are essential to achieve this. A reliable probabilistic forecast will become even more important in the future, especially due to the higher spatial resolution of future NWP generations. New generations of NWP ensemble models should improve the quality of the results significantly.

Experiences with ensemble water-level prediction based only on meteorological uncertainty in the Netherlands showed the need to implement other sources of uncertainty. Figure 19 shows an ensemble water-level forecast during a medium size flood in January 2011. The orange lines are the forecasted water levels based on ECMWF weather prediction, the blue line is the observed water level. A hindcast of the event has shown that in this specific case the main sources of uncertainty were the initial conditions and the snow melt.

Rainfall input is the largest obstacle to reliable flood forecasts and is responsible for most of the uncertainty in flood forecast. The inclusion of other sources of uncertainty such as modeled snow melt, initial conditions (e.g., soil properties), and various sources of uncertainty in the hydrological models can help to create a more realistic spread. Stable, robust and easily interpreted approaches are essential to evaluate ensemble forecasts accurately and to improve the actual usability for end users.

Different approaches have been used to integrate these sources of uncertainty. In recent years, Bayesian Mean Averaging (BMA) (Liang et al. 2013) and Quantile Regression (QR) (Weerts et al. 2011; Lopez et al. 2014) have been tested in the Netherlands. In BMA, a correction for the bias and the uncertainty of

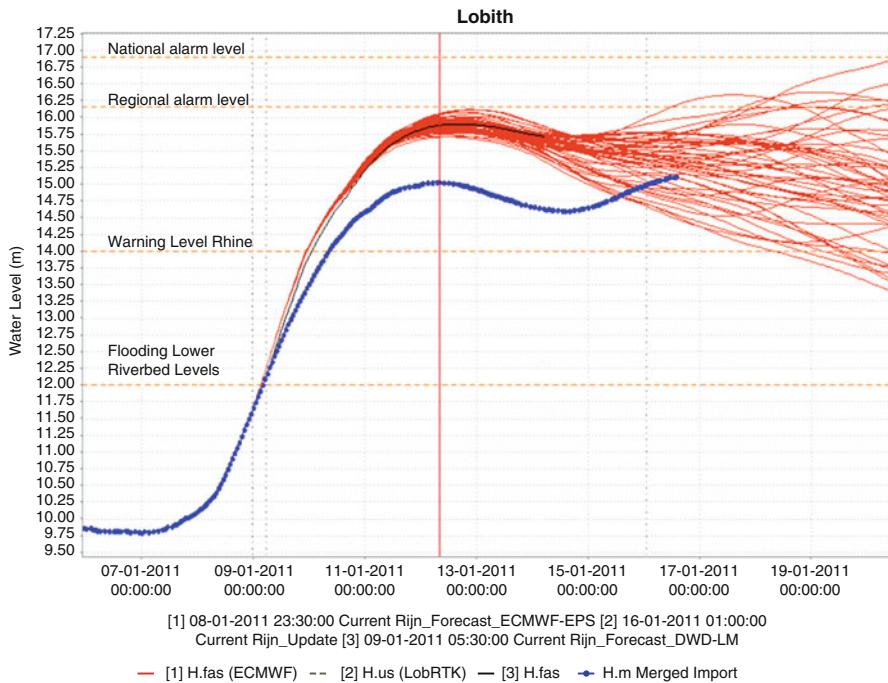


Fig. 19 Water-level forecast for the Rhine at Lobith based on ECMWF ensemble weather prediction during a medium size flood in January 2011. The *orange lines* are the 50 ensemble forecasts, the *blue line* the observed water level

an ensemble forecast in a training period prior to the present forecast is used. In the training period, historical model forecasts are compared with observations. The spread within and between the model realizations is used to quantify the uncertainty of the overall forecast. From the performance of each model in this training period, it is possible to calculate the likelihood of accuracy of the current forecast. This is used as a weight in the overall forecast (Ebel and Beckers 2009). The QR technique conditions forecast uncertainty of the forecasted value itself, based on a retrospective quantile regression of hindcasted water-level forecasts and forecast errors (Weerts et al. 2011). Both methods provide a relatively simple, efficient, and robust method for estimating predictive uncertainty and will be further tested in the coming years.

A relatively new way to produce operational forecasts in the Netherlands is the “Ensemble Dressing” method (Verkade et al. 2013). This method assumes that the meteorological forecast ensemble is unbiased. The “naked” ensemble prediction as it is published by the Dutch forecasting center does not contain information on the hydrological uncertainty (model, parameters, initial conditions). The hydrological uncertainty is estimated through a hindcast based on observed precipitation and temperature (“perfect forcing”) and characterized with Quantile Regression. The

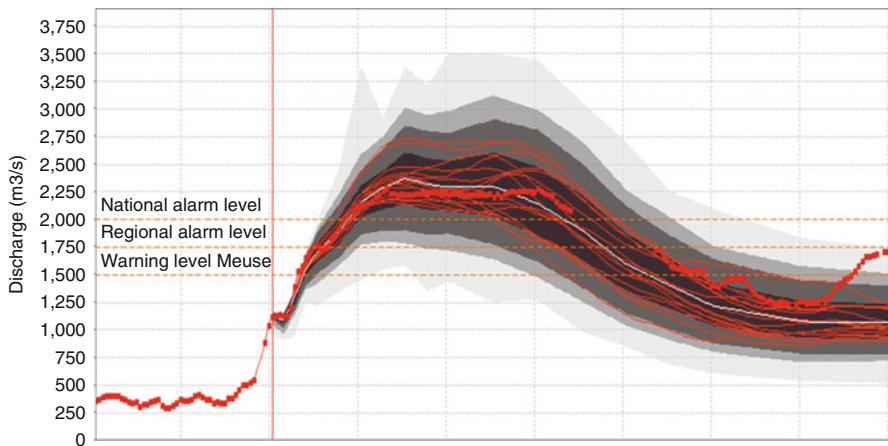


Fig. 20 COSMO-Leps ensemble discharge forecast, dressed with hydrological uncertainty post processor. An example for the River Meuse in the Netherlands

estimation of the hydrological uncertainty is determined for each of the ensemble members of the Cosmo-LEPS forecast depending on the lead time and on the height of the forecast at that lead time. The total uncertainty is estimated as the mean of the individual density functions. An example (here for the river Meuse) of the COSMO-Leps ensembles, “dressed” with hydrological uncertainty is shown in Fig. 20.

4.5 Conclusions for Medium-Range Flood Forecasting for Controlled Rivers

The introduction of information about the uncertainty of operational forecasts can be considered a prerequisite for state-of-the-art forecasts in the future, not only for operational flood forecasts but also for seasonal and for drought forecasts. Currently many end users are still used to only receiving deterministic information on forecasted runoff and water levels, but there is growing appreciation for the added value of probability information in forecasts. Good communication, easy to handle information, and trainings are essential for decision makers to accept these new forms of forecasting.

Currently not all forecast centers along the Rhine are using EPS-based systems, but the discussion about how the upstream-downstream chain will work for EPS has already started. In a connected trans-boundary catchment with different models and systems as the one presented, it is challenging but crucial to pass on information about uncertainty from one forecasting center to the next. EPS will become an integral part of the flood forecasting system of the Rhine, allowing an extension of reliable lead times in many cases and helping to identify uncertain forecasting situations. To increase the level of acceptance, the calculations need to be constantly reliable and be complemented with bias corrections. Stable approaches, which are not only easy to configure and calibrate, but also easy to use and interpret, will evolve with time and

experience. The introduction of improved ensemble NWP with higher spatial resolution and higher updating cycles will significantly help to reduce the level of uncertainty and to enhance the reliability of these new forecasting products.

5 Requirements for Using and Communicating Hydrometeorological EPS Medium-Seasonal Forecasting

While the previous sections on the this chapter have addressed flash flood and flood forecasting for which nowcasting, short-term and medium-range ensembles are useful, there are many other hydrological applications which act on longer time scales, e.g., for drought applications, water resource management or hydropower. In the following the requirements for using and communicating hydrometeorological EPS in the medium-seasonal scale is being illustrated at the example of hydropower.

5.1 Hydropower, Forecasting and Communication Requirements

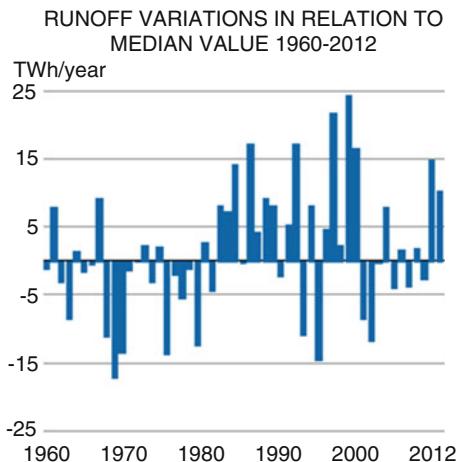
Worldwide, hydropower is the most common form of renewable energy. In Sweden, some 45%, or 65 TWh, of the total electricity production comes from hydropower. Most of the major rivers are affected by hydropower production. Dams and reservoirs have been constructed to store runoff, generated during snow melt and autumn rains, throughout the winter season when runoff is low and electricity demands are high. In hydropower plants, the production can quickly be increased in response to increased demands. In combination with the availability of stored water, it makes hydropower especially valuable to match supply both from less flexible sources, like nuclear power, and more variable renewable sources like wind.

In Sweden the annual runoff, expressed in energy terms, may vary by as much as 30 TWh (Fig. 21). Whether it is a wet year or a dry year has a large effect on the electricity prices in Scandinavia. The increased transfer capacity between Scandinavia and the European continent means that the water availability in Norway and Sweden also starts to affect prices outside Scandinavia.

Inflow forecasts are of interest for different lead times as well as time horizons:

- Seasonal forecasts are used to predict the inflow during the next 3–6 months. Of particular interest is the estimation of the snow melt runoff volume. The runoff forecasts are used for long-term production and reservoir planning as well as for price estimates.
- Medium-range forecasts are of importance for run-of-river plants with low storage capacity. Predictions of high inflow will most likely decrease electricity prices. During flood periods when reservoirs are full, medium-range forecasts becomes essential in reservoir management and the prevention of damages along the rivers. At the end of the winter when many reservoirs are nearly empty, the interest is in predicting the date when the snow melt runoff starts.

Fig. 21 Variation in annual inflow to Swedish hydropower plants expressed in TWh. Deviation from median value. The conversion to TWh is based on the current production capacity (Source: Swedenergy)



The HBV rainfall/runoff model was originally developed in Sweden in the early 1970s (Bergström 1972; Bergström and Forsman 1973; Lindström et al. 1997). Since then it has been widely used by the hydropower industry in Scandinavia to forecast the inflow to reservoirs and power plants. Seasonal forecasting with a climatological ensemble was an early application. Around 1990, the first Windows user interface called Integrated Hydrological Modelling System (IHMS) was created for the HBV model. From being a fairly complicated tool handled by a few, the HBV became an everyday tool for the hydrologists at the hydropower companies. At the beginning, seasonal forecasts were made once a month during winter and medium-range forecasts on special occasions. Today, seasonal forecasts are made several times a week and medium-range forecasts every day. The uncertainty of the forecasts has largely been handled subjectively, taken into consideration in the decisions made on production plans. Only recently have the hydropower companies started to include the full forecast ensembles in their planning systems.

5.2 Seasonal (Spring-Flood) Ensemble Forecasting

In Sweden seasonal runoff forecasts for hydropower applications are mainly carried out by hydrologists employed by the hydropower companies. They use a common tool, the Integrated Hydrological Modelling System (IHMS) which includes the HBV model. The HBV was originally developed to assist hydropower operations. The aim is to create a conceptual hydrological model with reasonable demands on computer facilities and calibration data. The HBV approach has proved flexible and robust in solving water resource problems and applications now span over a broad range.

As was mentioned in the previous section, hydrological forecasts of future events are per definition uncertain and the longer the lead time the larger the uncertainties. In order to capture this uncertainty, ensembles are applied.

The seasonal forecasts are traditionally based on a climatological ensemble including all years from 1961 and onwards. For the catchment in question, a set-up and calibrated HBV model is initialized by running it until the day of the forecast using observed meteorological inputs (temperature and precipitation). Then the catchment's meteorological time series in the forecast period from each historical year is used to feed the initialized HBV model that way generating one possible realization of the future runoff (Fig. 22). The variation in forecast results is thus due to the initial conditions at the time of the forecast, mainly the amount of water stored in the snow pack. The main aim of the forecast is to predict the inflow volume to reservoirs. The IHMS includes routines to accumulate the runoff over the forecast period and to compute runoff levels with different exceedance probability.

The IHMS/HBV also includes routines to evaluate the seasonal forecast with respect to the ensemble mean. The forecast volume of each forecast is saved, and at the end of the season, the program is run to compute the forecast error. Normally, the forecast made just before the start of the melt season is discussed by end users and modelers, and the forecast error is compared to the model error. So far, there is no routine to evaluate the probability levels.

Previously, many of the efforts to improve the forecasts have been focused on improving the hydrological model components, which have been estimated to

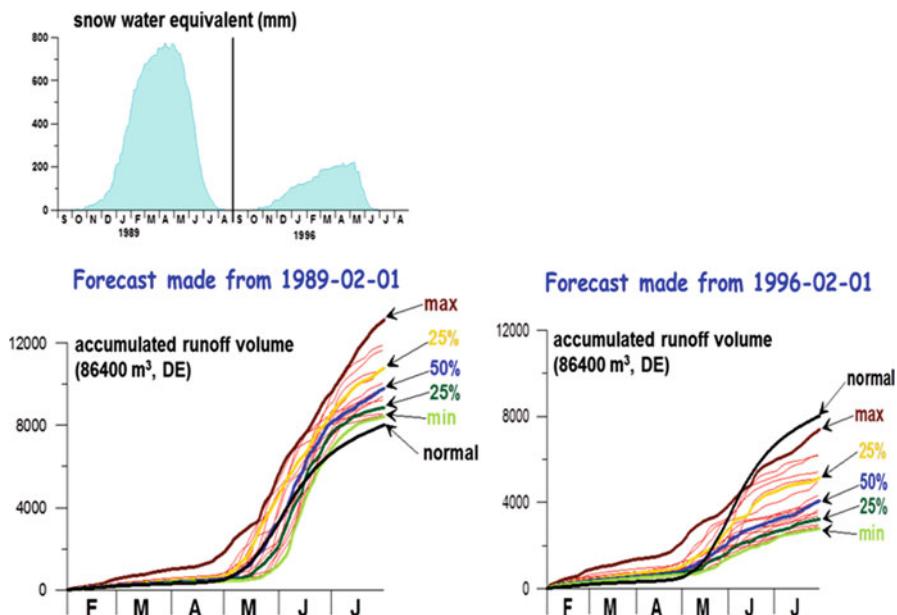


Fig. 22 Examples of seasonal forecasts based on climatological ensembles. The examples from two different years illustrate the importance of the initial state for the outcome of the forecast. In 1989 the snow pack on the first of February was around three times higher than in February 1996

contribute to 60% of the total forecast error (Arheimer et al. 2011). The remaining error is thus related to the meteorological forecast, and recently more focus has been put on improving this part of the forecasting chain. In the climatological ensemble approach, the mean forecast indicates the hydrological response to a future evolution of the weather which corresponds to the normal conditions. The spread between ensemble members indicates the expected range if the weather will deviate from the normal conditions. But the forecast does not contain any information whether the weather in the specific forecast period ahead is expected to become normal or not. Thus, a conceivable way to improve the hydrological forecasts is to incorporate information about the expected future development of the weather and its deviation from the climatological mean.

One obvious possibility is to replace the historical meteorological time series by seasonal (ensemble) meteorological forecasts over the period of interest. Such forecasts are available from, e.g., the European Centre for Medium-range Weather Forecasting (ECMWF). Generally, the accuracy of seasonal meteorological forecasts is limited in Sweden as well as the rest of Northern Europe. This is especially the case for precipitation, whereas temperature forecasts are normally more reliable. As temperature is a key driver for the spring flood generation, this implies a potential added value.

Tests of using ECMWF seasonal forecasts for spring flood forecasting in Sweden have been performed in the last 5–10 years but with limited success so far. No distinct improvement over the climatological ensemble approach has been attained. This is likely a combination of scale issues and model bias. The size of the grid cells in the ECMWF forecasting model (e.g., $1^\circ \times 1^\circ$) produces smoothed fields of temperature and precipitation that do not fully reflect the impact of altitudinal gradients and catchment-scale variability. This fact as well as general model uncertainty can make the forecasts differ systematically from catchment-scale observations. Better performance is likely attainable by statistical downscaling and bias correction of the ECMWF forecasts prior to the hydrological forecast simulations.

An alternative approach, which is not affected by scale issues, is to try to reduce the historical, climatological ensemble by estimating which of the historical years that are most likely representative of the future evolution of the weather during the current year. The general features of the weather during a certain year or period may be characterized using, e.g., teleconnection indices (TCI) or circulation patterns (CP). In these methods, the large-scale atmospheric circulation is described in terms of pressure anomalies (i.e., deviations from the climatological mean) or frequency of different weather situations. By characterizing the current year's weather in terms of TCIs and/or CPs, using either observations up until the forecast time or forecasts over the period of interest, and comparing with the historical years' values, the most similar historical years may be identified and used in the forecast simulations. Preliminary testing of the approach has failed to show any distinct overall improvement over the climatological ensemble approach although there are signs of increased accuracy in the early forecasts, e.g., spring flood forecasts issued already in the beginning of the year.

A further approach that does not involve hydrological modeling is statistical atmospheric downscaling of spring flood volume directly. As (1) the spring flood is a direct reflection of the winter climate and (2) the winter climate in Scandinavia is known to be governed by large-scale climate teleconnections, as manifested in e.g., the North-Atlantic Oscillation (NAO) index, such a direct link is conceivable. By advanced statistical tools, it has proved possible to link the spring flood volume to large-scale atmospheric variables such as pressure, wind speed, and humidity in winter. Also this approach has demonstrated a potential to improve the accuracy of early spring flood forecasts.

The Swedish hydropower industry is closely following the development of the new approaches and is partly funding the research (Olsson et al. 2016). An attempt to combine the current climatological procedure with the new approaches in a multimethod system has indicated that there is scope for increasing the overall forecast accuracy by 5–10%. This would translate into a quite substantially more effective energy production and, in turn, increased economical revenues. Preoperational testing of the system is ongoing.

5.3 Medium-Range (10-Day) Forecasting

The hydrological forecasting division at SMHI has been running a hydrological ensemble prediction system (EPS) to generate probability forecasts operationally since 2004. The inputs to the HBV model are the meteorological 10-day EPS forecasts from ECMWF, consisting of 50 ensemble members and one undisturbed control forecast. An auto-regressive updating of the simulated flow is applied in each single forecast run, making the simulated discharge at the time of the forecast agree with the concurrent observation. From the resulting 51 hydrological ensemble forecasts, “raw” nonexceedance probabilities are estimated. Evaluations of these raw forecasts have however clearly showed that the resulting nonexceedance probabilities are heavily biased (Olsson and Lindström 2008). Mainly, the spread in the hydrological ensemble is far too narrow, especially in the first days of the forecast. Thus, the forecasts often falsely indicate a high degree of certainty, whereas in reality there is a high probability of the actual flow being outside the ensemble range. The reason for the underestimated spread is not yet fully clarified, but it is likely a combination of limitations in both the meteorological forecasts (such as an underestimated spread also in them as well as scale issues making the forecasts less variable than catchment-scale weather fluctuations) and in the hydrological process descriptions. A simple correction is applied to the raw probabilities in order to provide more realistic estimates, including the combined effect of uncertainties in both the meteorological forecast and the hydrological model.

Ensemble stream flow predictions based on the HBV model are done both for selected catchments and on the national scale. The selected catchments comprise 80 so-called indicator basins in Sweden, which are designed to represent the hydrological conditions in different parts of the country (and catchments of different size) and where there is a real-time discharge gauge. The ensemble stream flow

prediction at SMHI is integrated with the national forecasting system and tailored products are available to specialized end users via Internet. An example of how the single-catchment ensemble forecasts are visualized for the end users is given in Fig. 23a.

On the national scale, flood probability maps for exceeding a certain threshold, i.e., a certain warning level, are produced automatically once a day (Fig. 23b). The flood probabilistic forecasts are based on a HBV-model set-up that covers the whole country (“HBV-Sweden”) divided into 1001 subbasins of sizes between 200 and 700 km². Probabilities for exceeding a certain warning level (corresponding to return periods of e.g., 2 or 10 years) are calculated for each one of these 1001 subbasins, where the levels have been determined by frequency analysis of historical simulations.

Hydrological probability forecasts should be seen as an early warning product that can give improved support in decision making to end users communities, for instance, Civil Protections Offices and County Administrative Boards as well as hydropower plant operators. An early indication that “something is going to happen,” that may not be evident from a deterministic forecast, may be highly valuable especially in conditions with already high flows. Besides the actual estimated levels, the ensemble spread itself gives an indication of the forecast uncertainty, which may be qualitatively useful. There are however also some conceived limitations with the medium-range probability forecasts when used in an operational context. One is the previously discussed limitations related to the spatial resolution of the meteorological ensemble forecasts and the resulting difficulties in capturing local variations. This particularly affects un-gauged catchments without possibility of updating the hydrological forecasts.

5.4 Conclusions and Future Outlook for Ensemble-Based Hydropower Forecasting

Ensemble forecasts on both medium-range and seasonal timescales have been available for the Swedish hydropower industry since more than a decade. The importance is supported by the fact that the industry continuously funds research aiming at improved forecasts by probabilistic approaches. In actual practice, the use of ensembles and probabilities for concrete planning and management still appears limited although the hydropower companies very recently have asked for full forecast ensembles to be delivered. There is clearly a general understanding of the potential benefits of probability forecasts, but procedures for using the added information in the actual operation remain to be developed. Further, the problems encountered with, e.g., scale issues and biased probabilities may have made hydropower end users hesitate and await the development of more robust and reliable products.

A good overview of the experiences and expectations from the international community of hydropower practitioners was provided by the 2011 Workshop on Operational River Flow and Water Supply Forecasting arranged by the Canadian Society for Hydrological Forecasting (Weber et al. 2012). Concerning ensemble

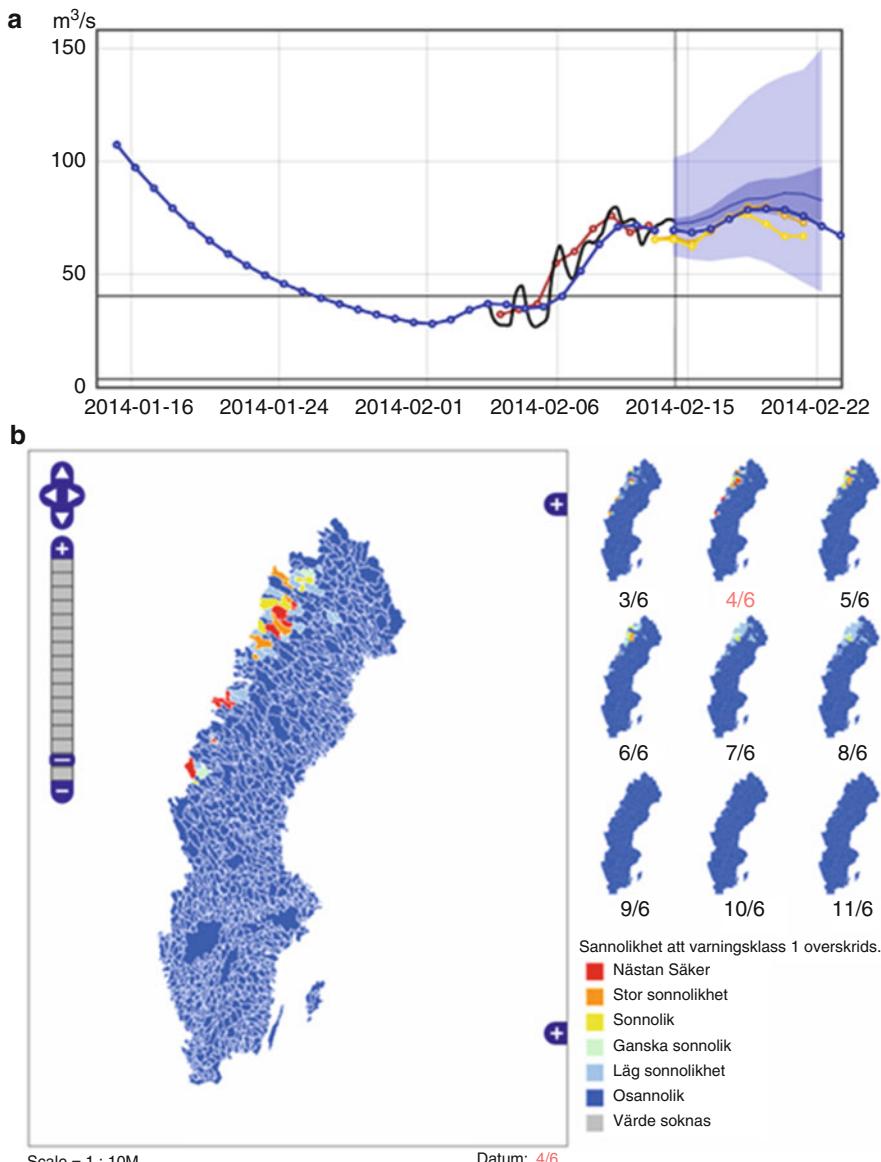


Fig. 23 (a) Presentation of hydrological model runs based on different sets of meteorological forecasts. Before the forecast time (2014-02-15), black and red lines represent observations of different time resolution and the blue circles model HBV simulations. After the forecast, blue circles represent deterministic HBV forecast, solid blue line and blue fields different probabilities of exceedance (top to bottom: 2%, 25%, 50%, 75% and 98%) from the HBV EPS forecast, and the yellow lines deterministic forecast and EPS median from the S-HYPE model. (b) Flood probability maps for exceeding a given threshold

forecasting, it was recommended to sample the full range of uncertainties including not only the future weather but also uncertainties in the initial state as well as hydrological model structure and parameter values. The latter issues are being explored also at SMHI, e.g., by studying new ways to estimate the snow cover and by using new, alternative model concepts.

In addition to hydrological forecasting using the HBV model, the S-HYPE model (Strömqvist et al. 2012) is gradually being introduced at the SMHI. S-HYPE is a high-resolution model for all of Sweden, based on the HYPE model code (Lindström et al. 2010). The average basin area in the current S-HYPE version is about 13 km². Some advantages of using the S-HYPE model for hydrological forecasts are that it covers the whole country and updates predictions of discharge and water levels at gaging stations. Hydrological statistics are available for the whole country, based on a method described by Bergstrand et al. (2014). Today, S-HYPE is run with both alternative deterministic forecasts and ensemble forecasts from the ECMWF (Fig. 23a), and the plan is to also introduce seasonal forecasts. The operational forecasts are stored in a database, which makes it possible to follow up and evaluate forecast quality (see e.g., Pechlivanidis et al. 2014). From late 2013, hydrological forecasts based on S-HYPE are available at the operational flood warning service at SMHI. From February 2014, S-HYPE forecasts are also made available to the public at <http://vattenwebb.smhi.se/hydronu/>. This service puts the present hydrological situation into perspective by comparing with historic data and also includes a 10 day forecast.

Another recommendation from the above mentioned workshop was to use a thorough suite of statistical and graphical measures for forecast evaluation. Knowledge of the historical forecasting performance is clearly beneficial, both for the forecaster and for the user of the forecast, but unfortunately only limited attention is generally paid to this issue. The workshop further concluded that postprocessing of forecast ensembles will continue to be necessary and also emphasized that the hydrological forecaster needs to have a good understanding of the physical processes involved, in order to properly interpret and convey the results of model forecasts.

Concerning the future role of hydropower production, it is likely to become more and more important as a regulator, to meet both the variation in electricity use and the variation in production from other renewable sources like wind and solar energy. Climate change is however expected to substantially change the future hydropower production as well as the future role of the hydrological forecasts in Sweden. The global warming will make the winter climate substantially milder which means that not all precipitation will accumulate as snow but also runoff episodes will occur. As the power demand is high during winter, this increased inflow is beneficial for the hydropower production. The reduced snow pack will lead to lower spring flood peaks, and overall the annual cycle will even out. With a reduced need for reservoir storage capacity before the spring flood, electricity can be produced with a higher hydraulic head. The added hydropower potential in Sweden by the end of the century owing to climate change has been estimated to equal the capacity of 1–3 nuclear reactors (Andréasson et al. 2007).

Concerning the forecasts, spring flood forecasts will remain important for the foreseeable future, but on the long term their importance is expected to decrease

along with the gradual weakening of the spring flood. Long-term forecasts for other seasons, winter in particular, will become more important. A general expected consequence of the global warming is a change towards higher variability and more pronounced extremes. This will likely make forecasting even more challenging than today. A particular difficulty concerns the discharge and water-level thresholds used for issuing warnings and alerts, e.g., 2- or 10-year levels. In a nonstationary, changing climate, these levels will gradually change, and in snow melt dominated hydrological regimes such as northern Sweden, these changes may happen relatively fast. Developing robust routines for frequency analysis and optimal estimation of warning levels under nonstationary conditions is an important and urgent task.

6 Summary

This chapter has illustrated with four different examples the use and communication of ensemble prediction systems for very short-term applications of flooding in urban areas and flash flood indicators, for medium-range applications for flood forecasting in medium-large water sheds and for medium-seasonal applications in reservoir management. The limitations of using Numerical Weather Prediction-based EPS in very short-term applications on the one hand and long-term applications on the other hand are being clearly demonstrated, while the benefits of EPS for short-medium-range flood forecasting is clearly illustrated. Examples of hydrological ensemble prediction systems produced with different inputs are being shown, examples being remote sensing rainfall estimates, numerical weather prediction systems or climatological data.

The examples have shown that, not surprisingly, the requirements depend largely on the application. Nowcasting systems have different requirements on spatial and temporal scales than seasonal forecasting systems, systems for flood forecasting experts have different requirements for visualization than systems for disaster managers. However, the following common points can be highlighted: First of all, uncertainty is inherent in all forecasting and prediction systems and needs to be dealt with and not ignored. Second, the reliability of the forecasts is considered key and hydrological ensemble prediction system requires postprocessing to ensure this reliability. Biases and unreliability particularly in the meteorological NWP products limit the usability of the data for hydrological applications. Only reliable systems generate the required trust with the experts and the end users that the system is valuable. Third, the uncertainty needs to be quantified and visualized and there is a multitude of different ways to do so. Fourth, information is typically presented either in the form of time series, e.g., hydrographs or rainfall volumes or maps, e.g., flood inundation maps. The complexity of the visualization depends largely on the end user and differs for expert users, e.g., forecasters or operators, or downstream end users such as disaster managers. Fifth, in the examples presented here, the ensemble information is plotted or visualized against “deterministic” thresholds, i.e., warning levels.

References

- L. Alfieri, D. Velasco, J. Thielen, Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. *Adv. Geosci.* **29**, 69–75 (2011)
- L. Alfieri, J. Thielen, F. Pappenberger, Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland. *J. Hydrol.* **424–425**, 143–153 (2012)
- J. Andréasson, S.-S. Hellström, J. Rosberg, S. Bergström, Summary of climate change maps of the Swedish water resources – background material for the Swedish Commission on Climate and Vulnerability, SMHI Hydrology No 106, SMHI, 601 76 Norrköping, 15 pp (in Swedish) (2007)
- B. Arheimer, G. Lindström, J. Olsson, A systematic review of sensitivities in the Swedish flood-forecasting system. *Atmos. Res.* **100**, 275–284 (2011). <https://doi.org/10.1016/j.atmosres.2010.09.013>
- M. Arpagaus et al., MAP D-PHASE: demonstrating forecast capabilities for flood events in the Alpine region. *Veröff. MeteoSchweiz* **78**, 75 pp (2009)
- A. Atencia, T. Rigo, A. Sairouni, J. Moré, J. Bech, E. Vilaclara, J. Cunillera, M.C. Llasat, L. Garrote, Improving QPF by blending techniques at the Meteorological Service of Catalonia. *Nat. Hazards Earth Syst. Sci.* **10**, 1443–1455 (2010)
- A. Bab-Hadiashar, D. Suter, R. Jarvis, 2-D motion extraction using an image interpolation technique. *SPIE* **2564**, 271–281 (1996)
- M. Berenguer, D. Sempere-Torres, G. Pegram, SBMcast – an ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *J. Hydrol.* (2011). <https://doi.org/10.1016/j.jhydrol.2011.04.033>
- M. Bergstrand, S.-S. Asp, G. Lindström, Nation-wide hydrological statistics for Sweden with high resolution using the hydrological model S-HYPE. *Hydrol. Res.* **45**, 349 (2014)
- S. Bergström, The application of a simple rainfall-runoff model to a catchment with incomplete data coverage. SMHI, Notiser och preliminära rapporter, ser. Hydrologi No. 26. IHD-Report (1972)
- S. Bergström, A. Forsman, Development of a conceptual deterministic rainfall-runoff model. *Nord. Hydrol.* **4**, 147–170 (1973)
- R. Bornstein, Q. Lin, Urban heat islands and summertime convective thunderstorms in Atlanta: three case studies. *Atmos. Environ.* **34**(3), 507–516 (2000)
- R. Buizza, A. Hollingsworth, F. Lalaurie, A. Ghelli, Probabilistic predictions of precipitation using the ECMWF Ensemble Prediction System. *Weather Forecast.* **14**, 168–189 (1999)
- R. Buizza et al., The new ECMWF variable resolution ensemble prediction system (VAREPS): methodology and validation. *ECMWF Tech. Memo.* No. 499 (2006)
- R. Buizza et al., The new ECMWF VAREPS (variable resolution ensemble prediction system). *Q. J. Roy. Meteorol. Soc.* **133**, 681–695 (2007)
- T. Camus, H.H. Bulthoff, Real-time optical flow extended in time. Max-Planck-Institute. Technical report No. 13 (1995)
- CHR, Website International Commission for the Hydrology of the Rhine basin. The length of the Rhine (2010)
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009)
- A.N. Clothier, G.G.S. Pegram, Space-time modeling of rainfall using the String of Beads model: integration of radar and raingauge data. Water Research Commission Report, No. 1010/1/02 (2002). ISBN 1 86845 835 0
- M. Ebel, J.L.V. Beckers, Operational river and coastal water level forecast using Bayesian Model averaging. DELTARES (Delft) (2010). Nr. 1200379-003
- A. Gelb, *Applied Optimal Estimation* (MIT Press, Cambridge, MA, 1974)
- U. Germann, M. Berenguer, D. Sempere-Torres, M. Zappa, REAL – ensemble radar precipitation estimation for hydrology in a mountainous region. *Q. J. Roy. Meteorol. Soc.* **135**, 445–456 (2009). <https://doi.org/10.1002/qj.375>
- R. Hannesen, An enhanced surface rainfall algorithm for radar data. *Progress report for MUSIC*. European Commission contract No. EVK1-CT-2000-00058 (2002)

- A.P. Hurford, S.J. Priest, D.J. Parker, D.M. Lumbroso, The effectiveness of extreme rainfall alerts in predicting surface water flooding in England and Wales. *Int. J. Climatol.* **32**, 1768–1774 (2012). <https://doi.org/10.1002/joc.2391>
- International Federation of Red Cross and Red Crescent Societies, *World Disaster Report 2009; Focus on Early Warning, Early Action* (ATAR Roto Presse, Satigny/Vernier, 2009)
- R.E. Kalman, A new approach to linear filtering and prediction problems. *Trans. ASME J. Basic Eng.* **82**(Series D), 35–45 (1960)
- H. Kalweit et al., Der Rhein unter den Einfluss des Menschen – Ausbau, Schifffahrt, Wasserwirtschaft. CHR report No I-11 (in German) (1993)
- K. Kober, G.C. Craig, C. Keil, A. Dörnbrack, Blending a probabilistic nowcasting method with a high-resolution numerical weather prediction ensemble for convective precipitation forecasts. *Q. J. Roy. Meteorol. Soc.* **138**, 755–768 (2012). <https://doi.org/10.1002/qj.939>
- H. Kusaka, K. Nawata, A. Suzuki-Parker, Y. Takane, N. Furuhashi, Mechanism of precipitation increase with urbanization in Tokyo as revealed by ensemble climate simulations. *J. Appl. Meteorol. Climatol.* **53**(4), 824–839 (2014)
- H. Leijnse, R. Uijlenhoet, J.N.M. Stricker, Rainfall measurement using radio links from cellular communication networks. *Water Resour. Res.* **43**, W03201 (2007). <https://doi.org/10.1029/2006WR005631>
- Z. Liang et al., Application of Bayesian model averaging approach to multimodel ensemble hydrologic forecasting. *J. Hydrol. Eng.* **18**(11), 1426–1436 (2013)
- K. Liechti, M. Zappa, F. Fundel, U. Germann, Probabilistic evaluation of ensemble discharge nowcasts in two nested Alpine basins prone to flash floods. *Wiley Online Libr.* (2012). <https://doi.org/10.1002/hyp.9458>
- S. Ligouri, M.A. Rico-Ramirez, A.N.A. Schellart, A.J. Saul, Using probabilistic radar rainfall nowcasts and NWP forecasts for flow prediction in urban catchments. *Atmos. Res.* **103**, 80–95 (2012)
- G. Lindström, B. Johansson, M. Persson, M. Gardelin, S. Bergström, Development and test of the distributed HBV-96 model. *J. Hydrol.* **201**, 272–288 (1997)
- G. Lindström, C.P. Pers, R. Rosberg, J. Strömqvist, B. Arheimer, Development and test of the HYPE (Hydrological Predictions for the Environment) model – a water quality model for different spatial scales. *Hydrol. Res.* **41**(3–4), 295–319 (2010)
- M.C. Llasat, M. Llasat-Botija, M.A. Prat, F. Porcú, C. Price, A. Mugnai, K. Lagouvardos, V. Kotroni, D. Katsanos, S. Michaelides, Y. Yair, K. Savvidou, K. Nicolaides, High-impact floods and flash floods in Mediterranean countries: the FLASH preliminary database. *Adv. Geosci.* **23**, 1–9 (2010)
- P. Lopez et al., Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the Upper Severn River: a comparison. *Geophys. Res. Abstr.* **16**, EGU2014-14591 (2014). EGU General Assembly 2014
- C. Marsigli, F. Boccanferra, A. Montani, T. Paccagnella, The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlinear Proc. Geoph.* **12**, 527–536 (2005)
- J. Olsson, G. Lindström, Evaluation and calibration of operational hydrological ensemble forecasts in Sweden. *J. Hydrol.* **350**, 14–24 (2008). <https://doi.org/10.1016/j.jhydrol.2007.11.010>
- J. Olsson, C.B. Uvo, K. Foster, W. Yang, Technical note: initial assessment of a multi-method approach to spring flood forecasting in Sweden, *Hydrol. Earth System Sci.* **20**, 1–9 (2016). <https://doi.org/10.5194/hess-20-1-2016>
- T.N. Palmer, F. Molteni, R. Mureau, R. Buizza, P. Chapelet, J. Tribbia, Ensemble prediction, in *Proceedings of 1992 ECMWF Seminar: Validation of Models Over Europe* (ECMWF, Reading, 1993), pp. 21–66
- L. Panziera, U. German, M. Babella, P.V. Mandapaka, NORA – nowcasting of orographic rainfall by means of analogues. *Q. J. Roy. Meteorol. Soc.* **137**, 2106 (2011). ISSN 0035-9009
- I.G. Pechlivanidis, T. Bosshard, H. Spångmyr, G. Lindström, G. Gustafsson, B. Arheimer, Uncertainty in the Swedish operational hydrological forecasting systems, in *Proceedings of the Second International Conference on Vulnerability and Risk Analysis and Management (ICVRAM2014)*, University of Liverpool, Liverpool, 13–16 July 2014 (2014)

- G.G.S. Pegram, A continuous streamflow model. *J. Hydrol.* **47**, 65–89 (1980)
- G.G.S. Pegram, A.N. Clothier, High resolution space-time modelling of rainfall: the “String of Beads” model. *J. Hydrol.* **241**, 26–41 (2001)
- G.G.S. Pegram, D.S. Sinclair, A linear catchment model for real time flood forecasting. Water Research Commission Report, No. 1005/1/02 (2002)
- D. Penna, H.J. Tromp-van Meerveld, A. Gobbi, M. Borga, G. Dalla Fontana, The influence of soil moisture on threshold runoff generation processes in an alpine headwater catchment. *Hydrol. Earth Syst. Sci.* **15**, 689–702 (2011)
- D. Raynaud, J. Thielen del Pozo, P. Salamon, P.A. Burek, S. Anquetin, L. Alfieri, A dynamic runoff co-efficient to improve flash flood early warning in Europe: evaluation on the 2013 Central European floods in Germany. *Meteorol. Appl.* **22**(3), 410–418 (2014)
- M. Renner, M.G.F. Werner, S. Rademacher, E. Sprokkereef, Verification of ensemble flow forecasts for the River Rhine. *J. Hydrol.* **376**, 463 (2009)
- A. De Roo, B. Gouweleeuw, J. Thielen, P. Bates, A. Hollingsworth et al., Development of a European Flood Forecasting System. *Int. J. River Basin Manag.* **1**(1), 49–59 (2003)
- A. Rossa, G. Haase, C. Keil, P. Alberoni, S. Ballard, J. Bech, U. Germann, M. Pfeifer, K. Salonen, Propagation of uncertainty from observing systems into NWP: COST-731 Working Group 1. *Atmos. Sci. Lett.* **2**, 145–152 (2010)
- M.W. Rotach et al., MAP D-PHASE: real-time demonstration of weather forecast quality in the Alpine region. *Bull. Am. Meteorol. Soc.* **90**, 1321–1336 (2009). <https://doi.org/10.1175/2009BAMS2776.1>
- K. Savvidou, S. Michaelides, K.A. Nicolaides, P. Constantinides, Presentation and preliminary evaluation of the operational Early Warning System in Cyprus. *Nat. Hazards Earth Syst. Sci.* **9**, 1213–1219 (2009). <https://doi.org/10.5194/nhess-9-1213-2009>
- J. Schaake, K. Franz, A. Bradley, R. Buizza, The Hydrological Ensemble Prediction EXperiment (HEPEX). *Hydrol. Earth Syst. Sci. Discuss.* **3**, 3321–3332 (2006)
- A.W. Seed, A dynamic and spatial scaling approach to advection forecasting, in *Proceedings of the Fifth International Symposium on Hydrological Applications of Weather Radar – Radar Hydrology*, Kyoto, 2001
- A.W. Seed, R. Srikanthan, M. Menabde, A space and time model for design storm rainfall. *J. Geophys. Res.* **100**(D21), 31623–31630 (1999)
- K. Stephan, S. Klink, C. Schraff, Assimilation of radar-derived rain rates into the convective-scale model COSMO-DE at DWD. *Q. J. Roy. Meteorol. Soc.* **134**, 1315–1326 (2008). <https://doi.org/10.1002/qj.269>
- J. Strömqvist, B. Arheimer, J. Dahné, C. Donnelly, G. Lindström, Water and nutrient predictions in ungauged basins: set-up and evaluation of a model at the national scale. *Hydrol. Sci. J.* **57**(2), 229–247 (2012)
- J. Thielen, A. Gadian, Influence of different wind directions in relation to topography on the outbreak of convection in Northern England. *Ann. Geophys.* **14**(10), 1078–1087 (1996)
- J. Thielen, A. Gadian, Influence of topography and urban heat island effects on the outbreak of convective storms under unstable meteorological conditions: a numerical study. *Meteorol. Appl.* **4**(2), 139–149 (1997)
- J. Thielen, W. Wobrock, P. Mestayer, J.-D. Creutin, A. Gadian, The influence of the lower boundary on convective rainfall development; a sensitivity study. *Atmos. Res.* **54**, 15–39 (2000)
- E. Todini, Mutually interactive state/parameter estimation (MISP), in *Application of Kalman Filter to Hydrology, Hydraulics and Water Resources*, ed. by C.-L. Chiu (University of Pittsburgh, Pittsburgh, 1978)
- M.S. Tracton, E. Kalnay, Operational ensemble prediction at the National Meteorological Center: practical aspects. *Weather Forecast.* **8**, 379–398 (1993)
- University Corporation for Atmospheric Research (UCAR), *Flash Flood Early Warning System Reference Guide 2010* (2010), http://www.meted.ucar.edu/communities/hazwarnsys/ffewsg/FF_EWS.frontmatter.pdf. ISBN 978-0-615-37421-5

- J. Verkade et al., Real-time hydrologic probability forecasting using ensemble dressing, with application to river Rhine. *Geophys. Res. Abstr.* **15**, EGU2013-10763 (2013). EGU General Assembly 2013
- G. Villarini, W.F. Krajewski, A.A. Ntelekos, K.P. Georgakakos, J.A. Smith, Towards probabilistic forecasting of flash floods: the combined effects of uncertainty in radar-rainfall and flash flood guidance. *J. Hydrol.* **394**(1–2), 275–284 (2010)
- F. Weber, D. Garen, A. Gobena, Invited commentary: themes and issues from the workshop “Operational River Flow and Water Supply Forecasting”. *Can. Water Resour. J.* **37**, 151–161 (2012). <https://doi.org/10.4296/cwrj2012-953>
- A.H. Weerts et al., Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrol. Earth Syst. Sci.* **15**, 255–265 (2011)
- P. Young, *Recursive Estimation and Time Series Analysis* (Springer, Berlin, 1984)
- J. Younis, S. Anquetin, J. Thielen, The benefit of high-resolution operational weather forecasts for flash flood warning. *Hydrol. Earth Syst. Sci.* **12**, 1039–1051 (2008)
- M. Zappa, K.J. Beven, M. Bruen, A.S. Cofino, K. Kok, E. Martin, P. Nurmi, B. Orfila, E. Roulin, K. Schroter, A. Seed, J. Sztruc, B. Vehvilainen, U. Germann, A. Rossa, Propagation of uncertainty from observing systems and NWP into hydrological models: COST-731 Working Group 2. *Atmos. Sci. Lett.* **11**, 83–91 (2010)



Best Practice in Communicating Uncertainties in Flood Management in the USA

Robert K. Hartman

Contents

1	Introduction	1094
2	Conveying Probabilistic Streamflow Information	1094
2.1	Static Product Generation	1095
2.2	Interactive “User” Product Generation	1097
2.3	Provision of “Raw” and “Postprocessed” Ensembles	1098
2.4	Managing Expectations	1099
3	Applications to Water Resources Forecasting	1100
4	Hindcasting and Validation	1103
4.1	Validation and Associated Services	1106
5	Conclusion	1107
	References	1108

Abstract

Ensemble forecasting has gained a great deal of popularity for addressing and estimating uncertainty associated with both meteorologic and hydrologic forecasts over the past decade. While ensemble-based hydrologic forecasts have been in routine operations for longer-term forecasts for many years, the notion of short- and medium-term probabilistic forecasts in support of water and flood management efforts is relatively new and is a developing science and service. Approaches to effectively conveying and communicating hydrologic forecast uncertainty are being actively developed and vetted with potential user communities. Important experience and insight will be gained over the next few years as the community of developers, forecasters, and end users work to leverage probabilistic forecasts in a

R. K. Hartman (✉)

California-Nevada River Forecast Center, NOAA, National Weather Service, Sacramento, CA,
USA

e-mail: Robert.Hartman@noaa.gov

risk-based decision environment. With proper focus and support, these efforts have the potential to significantly improve flood, ecosystems, and water management with benefits to multiple sectors of our society.

Keywords

Ensemble · Communication · Probability · Risk · Uncertainty · Hydrology · Water resources · Hindcasting

1 Introduction

Hydrologic ensemble forecasting procedures have made great strides over the past decade. Progress has been attributable to a growing acceptance that uncertainty is something that can be leveraged to make more informed decisions (National Research Council of the National Academies 2006) and substantial community support as evidenced through the success of the Hydrological Ensemble Prediction Experiment (HEPEX; www.hepex.org).

Among the most vexing challenges of the hydrologic ensemble prediction process is the appropriate conveyance of uncertainty information to decision makers. These decision makers represent many sectors (e.g., emergency services, power generation, recreation, agriculture, navigation, municipal water supply, industry, ecological management). Each has different and specialized needs and each has a different risk tolerance. In addition, their statistical background varies from nearly nothing (i.e., plays the lotto) to very sophisticated.

This chapter describes approaches and examples of how ensemble-based hydrologic forecast information is conveyed to users by the US National Weather Service today. Positive and negative attributes of approaches along with challenges are presented.

2 Conveying Probabilistic Streamflow Information

There are at least four fundamental approaches that can be used to provide uncertainty information associated with forecast streamflow. The oldest approach is to simply accommodate the expected “error” as a function of the user’s substantial experience with forecasts over time and, in particular, events that were memorable. This anecdotal uncertainty is what the hydrologic forecast and user community is attempting to supplant with objective information generated through ensemble techniques. More quantitative vehicles include:

- Generating a collection of ensemble-based products (text and graphics).
- Providing access to an interface that can create custom ensemble-based products (text and graphics).
- Providing ensemble members (data) that can be analyzed by end users in their own decision support architecture. Each approach has benefits and challenges and experience which has shown that all three, together, may represent a more reasonable approach.

2.1 Static Product Generation

Ensemble forecasts can be analyzed to address a seemingly infinite number of information requirements. This flexibility is beneficial, but it also creates challenges for forecast producers. What are the “best” sets of static graphics and text products that meet the greatest need for information? The fundamental questions associated with product generation are:

- What is the time period of interest (e.g., next 3 days, next 2 weeks, month of June)?
- What is the data aggregation period (e.g., hourly, daily, weekly, monthly, seasonal, annual)?
- What aspect of flow is of interest (e.g., summation (volume), mean, peaks, minimums, time to a threshold of interest)?

The more precisely the questions are addressed, the more useful the information for a specific application. Sounds simple enough, but ultimately, choices must be made if the number of routinely generated products is limited.

One of the most vexing issues encountered in generating ensemble-based graphics involves the impact of time aggregation on probability. Customers of hydrologic forecast information really want to see a “hydrograph” with associated uncertainty or “error bars.” What they really get is a series of histograms, at the time-step of analysis, placed side-by-side in sequential order. Interpretation of these sorts of graphics can very easily lead toward the wrong conclusions. Look first at Fig. 1. This graphic includes ten (10) 1-day histograms for flow and stage. It has the look of a hydrograph, but the 1-day time-step defeats that interpretation tendency to some degree. One interpretation of this graphic might be that “the river has a less than a ten percent chance of exceeding 12 feet over the next 5 days.” Now compare this with Fig. 2. This figure shows the distribution of peak flows within the coming 5-day period. This graphic suggests that the “river has a probability of between 25% and 50% of exceeding 12 feet over the next 5 days.” The proper interpretation of Fig. 1 is “the river has less than a ten percent chance of exceeding 12 feet on any of the next 5 individual days”; however, Fig. 2 indicates that collectively (all 5 days considered together), the probability of exceeding 12 feet is much higher. This phenomenon becomes more pronounced as the time-step of the analysis gets shorter. For example, if the time-step is reduced to 1 h, Fig. 1 really begins to look like a hydrograph and the likelihood of any extremes (peaks or minimums) is further reduced by appearance because their likelihood associated with any specific hour is lower than it is for any day or the entire period of consideration (e.g., 5 days). Best practices, therefore, limit the generation of graphics that are easily misinterpreted.

It is important for the users of hydrologic ensemble forecasts to continually remember that the generated products are simply an interpretation of the current set of ensemble members. For that reason, it remains good practice to provide a trace plot (spaghetti) among the set of routinely generated graphics. The 10-day trace plot that serves as the basis for the information in Figs. 1 and 2 is shown in Fig. 3. Note that while the analyzed probabilities shown in Figs. 1 and 2 may be well below the

NAVARRO RIVER - NAVARRO (NVRC1)

Latitude: 39.17° N Longitude: 123.67° W Elevation: 20 Feet
 Location: Mendocino County in California

River Group: Russian Napa

Issuance Time: Feb 28 2014 at 9:14 AM PST

10-Day Probability Plot

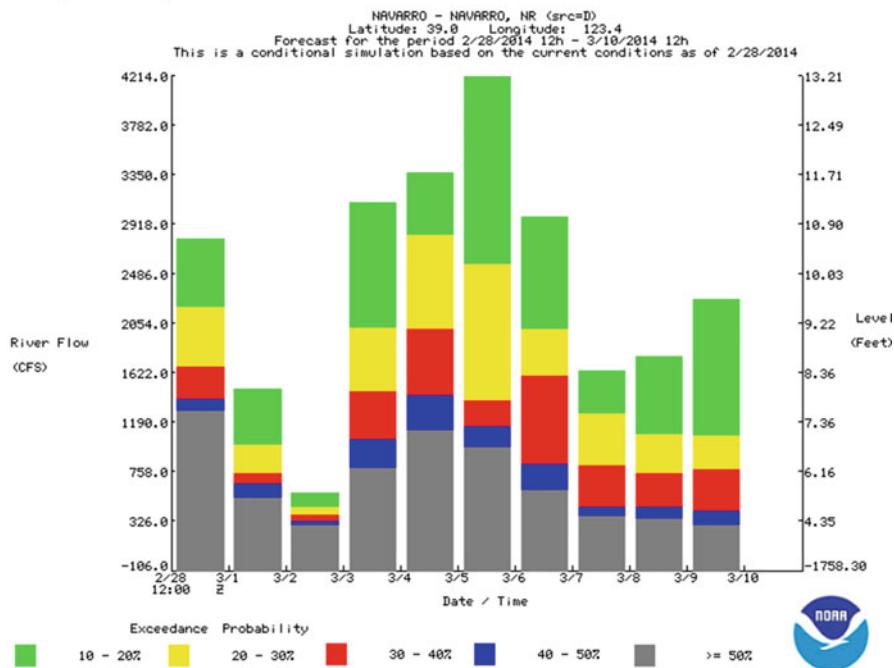


Fig. 1 Streamflow histogram, 1-day duration for Navarro River in California

region of concern for an emergency/resource manager, a limited number of traces may be very problematic and therefore very much worth being aware of.

Operators of reservoirs are normally more interested in volumetric (often multiday) forecasts of inflow rather than instantaneous or single day inflows. Understanding that the 1-day flows as shown in Fig. 1 cannot to be added together to form a probabilistic multiday volume, graphics such as is shown in Fig. 4 can serve a critical need of the reservoir management community. Again, without the provision of an accumulated volume plot (Fig. 4), a reservoir operator might be led to misinterpret a daily histogram (e.g., Fig. 1).

For some time, River Forecast Centers in the USA have generated 90-day graphics that depict weekly probabilities of maximum stage (Fig. 5) and the maximum stage probability distribution over the entire 90-day period (Fig. 6). These sorts of graphics are particularly valuable when preparing for spring snowmelt flooding as often occurs in the upper Midwest of the USA. As with all longer-range products, they make heavy reliance on the information content of the model states.

NAVARRO RIVER - NAVARRO (NVRC1)

Latitude: 39.17° N Longitude: 123.67° W Elevation: 20 Feet

Location: Mendocino County in California

River Group: Russian Napa

Issuance Time: Feb 28 2014 at 9:16 AM PST

5-Day Peaks Plot

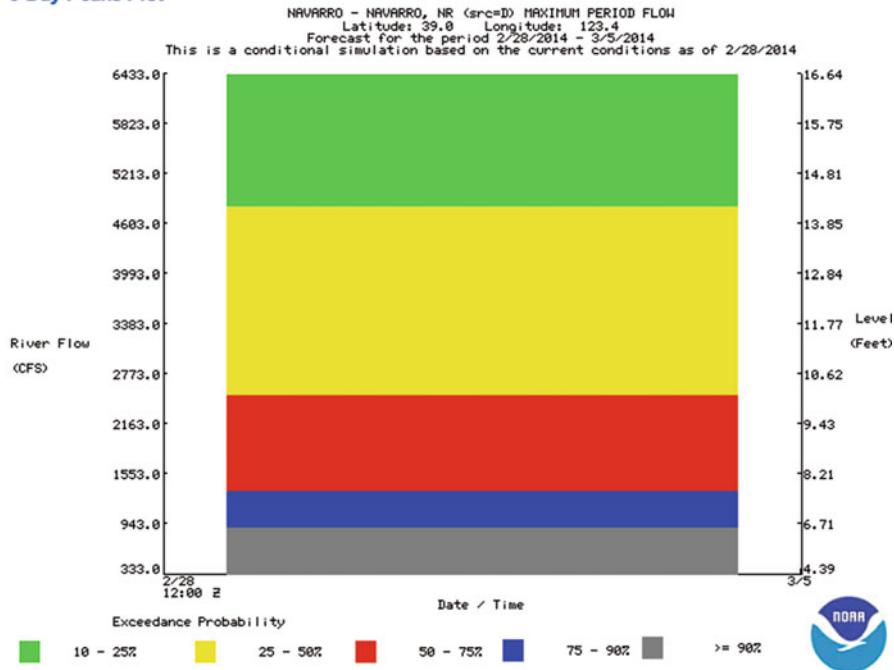


Fig. 2 Streamflow histogram, 5-day duration for Navarro River in California

2.2 Interactive “User” Product Generation

Given the diversity of interests in streamflow projections and the multitude of options for periods, durations, and flow attributes (maximum, minimum, mean, summation, and time to a threshold of interest), providing customers with a tool to analyze a set of ensembles using their specific criteria makes a lot of sense. This sort of feature does come with risk as it assumes that the user is well informed enough to make the required selections and properly interpret the results. This represents a minority of the total number of forecast customers, but to those who use it, it is a very important and powerful service. Figure 7 shows the interface supported by the California-Nevada River Forecast Center. A substantial “help” section is provided to assist users in navigating the options and interpreting the product generated. This sort of interface allows users to “narrow” the scope of their information need and generate products that directly address their requirements.

NAVARRO RIVER - NAVARRO (NVRC1)

Latitude: 39.17° N Longitude: 123.67° W Elevation: 20 Feet
 Location: Mendocino County in California River Group: Russian Napa

Issuance Time: Feb 28 2014 at 9:15 AM PST

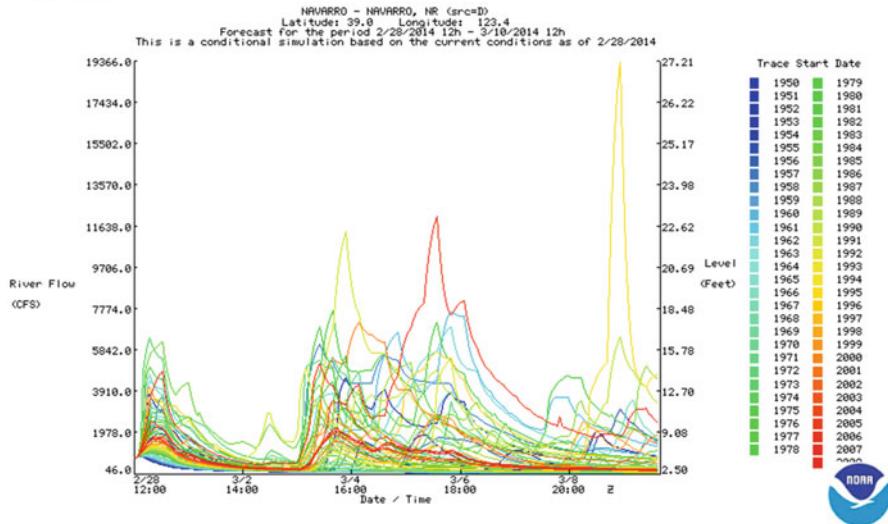
10-Day Traces Plot

Fig. 3 Ensemble streamflow traces for Navarro River in California

2.3 Provision of “Raw” and “Postprocessed” Ensembles

While forecasters and developers struggle with the “best” ways of describing the uncertainty of hydrologic ensemble forecasts with complex and often difficult to interpret graphics, the most effective practice may be to simply provide the data and allow the customer to perform an analysis that is meaningful to them. For sophisticated users with resources, this is clearly the most effective alternative. As an example, a reservoir operator with a model that can simulate operations can easily process each member of an ensemble set to evaluate the benefits/costs of a selected release strategy. Alternatively, that same operator will have to use their imagination to understand how a histogram of daily inflow probabilities will impact their regulation strategy. The difference is profound. Substantial progress is being made along this front. In California alone, the INFORMS project (Georgakakos et al. 2007) as well as the Yuba-Feather Forecast Coordinated Operations (FCO) project have engineered solutions to leverage the full potential of ensemble forecasts in a decision support model. Figure 8 shows the conceptual process schematic for the Yuba-Feather FCO ensemble-based decision support model.

Just as ensemble Numerical Weather Prediction (NWP) models exhibit biases and inappropriate spread, so too will the hydrologic forecasts without some sort of postprocessing methodology (Demargne et al. 2014). Sophisticated users, however,

DRY CREEK - LAKE SONOMA (WSDC1)

Latitude: 38.72° N Longitude: 123.01° W Elevation: 440 Feet

Location: Sonoma Country in California

River Group: Russian Napa

Issuance Time: Feb 28 2014 at 9:46 AM PST

10-Day Accumulated Volume Plot

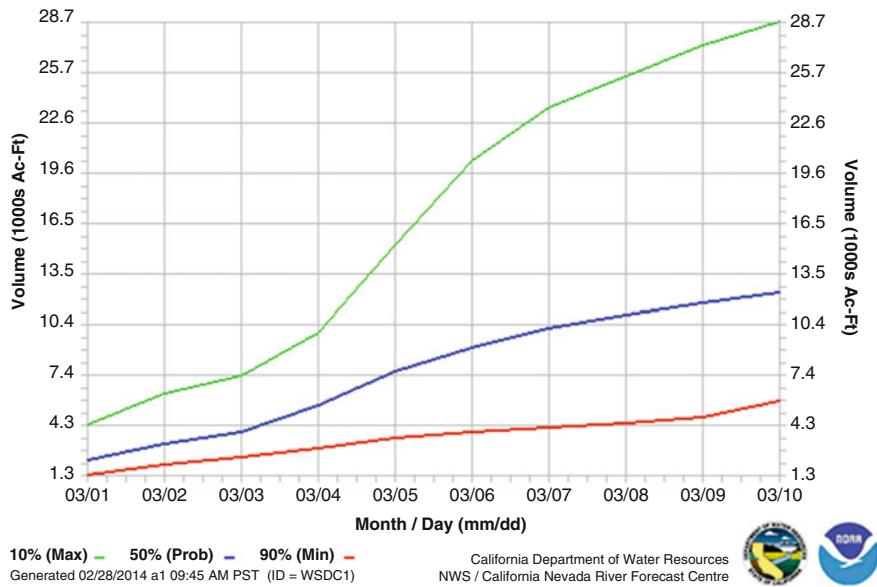


Fig. 4 10-day accumulated inflow volume for Lake Sonoma in California

have a choice in this process. They may choose to accept the “raw” ensembles without the benefit of postprocessing and instead apply their own error correction processes based on an adequately long history of performance. Such is the case for the ensemble forecast services provided to the New York City Department of Environmental Planning (NYCDEP) by the US National Weather Service. This may be the most efficient way of objectively accounting for ensemble reliability issues, but it also may result in additional workload for operational entities required to issue both raw and postprocessed ensemble information.

2.4 Managing Expectations

It is clear that, given all of the assumptions that must be made to generate an ensemble-based hydrologic forecast, there will be uncertainty in the estimates of uncertainty. The “discrimination” in the system may not be able to reliably differentiate between 85% and 90% probability of exceedance. Further, work is needed to help users understand that some risk must be assumed if one expects to leverage

AHPS / ESP Trace Analysis

The form consists of seven numbered sections:

- 1 Select a Location:** A dropdown menu showing "FEATHER RIVER - LAKE OROVILLE (ORDC1)".
- 2 Select an Accumulation Type:** Radio buttons for Mean (selected), Minimum, Maximum, and Summation.
- 3 Select an Interval:** Radio buttons for Day (selected), Week, Month, and Entire Period.
- 4 Select a Distribution Type:** Radio buttons for Empirical (selected) and Wakeby.
- 5 Select a Starting Date:** Month: Mar, Day: 01, Year: 2014. A note below states: "Please Note: For the Klamath River - Klamath **Excluding Reservoir Releases**, Klamath River - Iron Gate Reservoir, Klamath River - Below JC Power Plant, and Klamath River - Keno locations, a date one day in the future is the earliest that can be used to build a product. (Example: If today is November 15th, 2013 then the start date must be either November 16th, 2013 or later)"
- 6 Select an Ending Date:** Month: Jun, Day: 01, Year: 2014.
- 7a Select a Plot Option and Generate:** Radio buttons for Traces (selected), Probability, Expected Value, and Exceedance. A "Generate a Plot" button is present.
- 7b Select a Table Option and Generate:** Radio buttons for Forecast Info (selected), Quantiles, and Flood Quantiles. A "Generate a Table" button is present.

Help Making Selections and Interpreting Results (Click Help Button)

Fig. 5 Web interface for “Create Your Own” ensemble product

uncertainty in the long run. If you need to be 99% sure before you will take action, you will likely miss a lot of opportunities.

3 Applications to Water Resources Forecasting

Ensemble applications to water resources forecasting are not new but have gained substantial growth and acceptance over the past decade. Early work appeared in the 1970s and the National Weather Service formalized a process within their

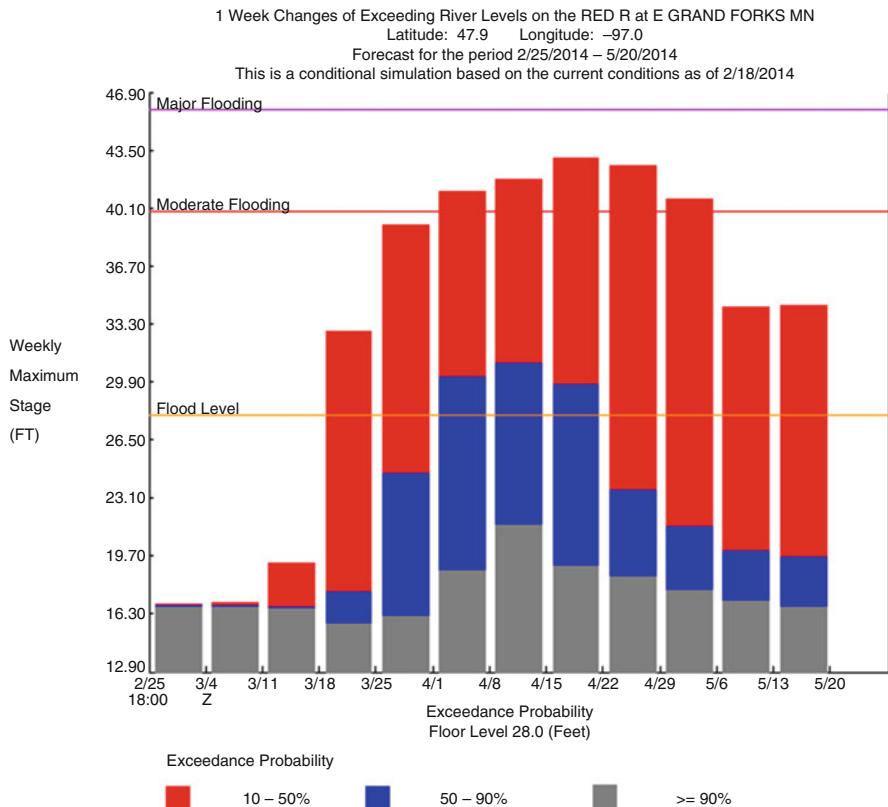


Fig. 6 Weekly histogram of maximum stage for the Red River in Minnesota

forecasting system in the mid-1980s (Day 1985). Despite this, the predominate approach for seasonal streamflow forecasting in the Western USA has remained some form of regression modeling driven with data available on a monthly basis (Garen 1992). More recently, forecasters have begun to integrate daily observations, and the US National Weather Service is in the process of shifting toward full reliance upon ensemble processes evaluated every day and year-round.

The attributes of relying on ensemble process for longer-range water resources forecasting include:

- Integration with short-term hydrologic forecasting procedures
- Use and integration of near real-time observations (e.g., precipitation, air temperature, streamflow)
- Integration of current weather and climate forecast information
- Ability to update on a daily basis

As this transition takes place, forecasters are experimenting with graphical products that describe both the uncertainty as well as how the forecasts have changed

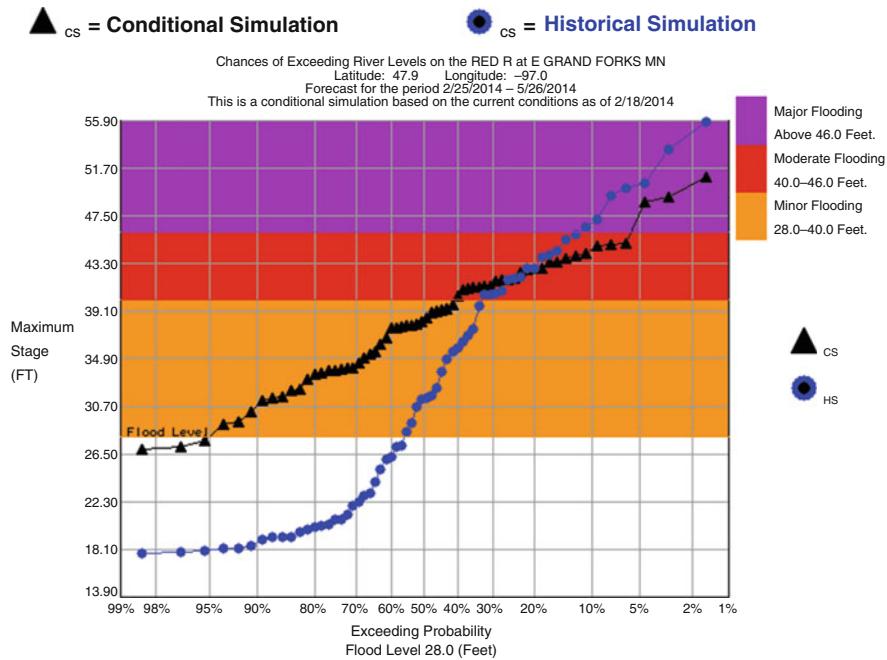


Fig. 7 90-day Exceedance probability plot for the Red River in Minnesota

or “trended” over time. Figures 9, 10, 11, and 12 show potential candidate graphics that describe expected volume over time. The “trend plots” in Figs. 10 and 11 do not show uncertainty but do show how the 50% exceedance probability forecast over seasonal volume has changed over time. Note the precipitous drop in expected water supply volume that took place during the fall and early winter as California received only a very small percentage of normal precipitation during this period. Assuming that the ensemble forecasts leverage the skill in the weather and climate forecasts, one can quickly see that the forecasts were not effective in detecting the coming drought with much or any lead time. This may seem discouraging, but it does highlight the need, value, and process for leveraging improved seasonal weather predictions through a hydrologic ensemble forecasting framework.

In the water resources services domain, the streamflow forecast alone is not adequate to provide customers with the complete picture of the water supply situation. In many areas, reservoirs provide a buffer for interannual variation as well as a means for shifting runoff from the time of occurrence to the time of need (e.g., irrigated agriculture). Information that summarized and combines the expected runoff with the existing reservoir storage is critical for assessment purposes. In addition, water supply customers have historically expressed a need to see information that supports the streamflow forecast itself, such as monthly and seasonal precipitation and snowpack. Comparison of precipitation and snowpack when

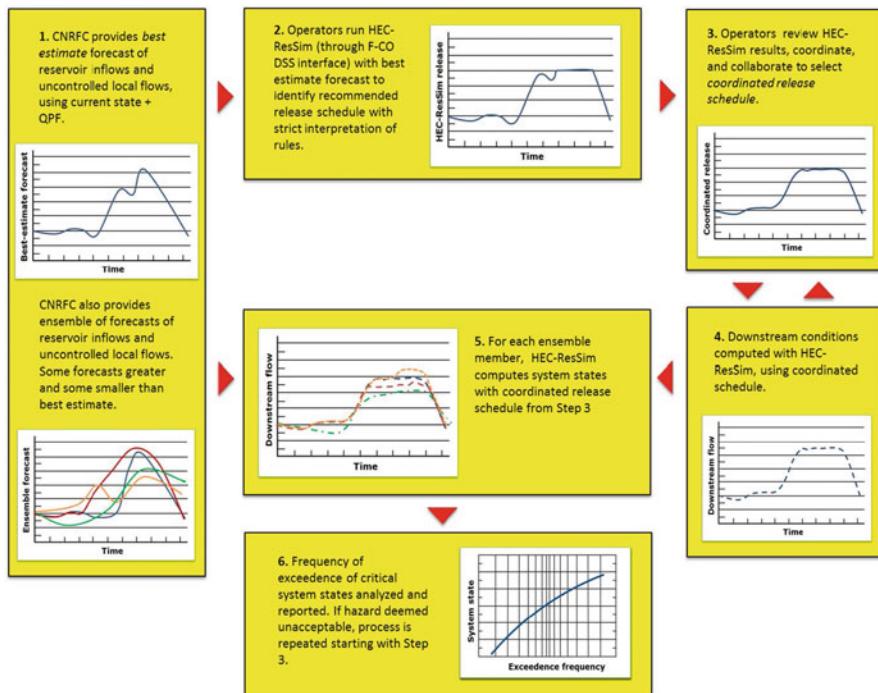


Fig. 8 Schematic of data flow for Yuba-Feather Forecast Coordinated Operations ensemble-based decision support model (by permission of David Ford Consulting Engineers, Sacramento, CA)

expressed as a percent of normal provide excellent context for understanding and establishing confidence in a specific volumetric seasonal streamflow forecast.

4 Hindcasting and Validation

The key attribute of ensemble-based probabilistic hydrologic forecasts that makes them useful is reliability. Reliability means that the ensemble members, as a package, are (1) unbiased and (2) have appropriate spread. If the ensembles have not been demonstrated to be adequately reliable, the user incurs a great deal of risk when applying the information contained in the ensembles to their specific decision-making process.

The process of “hindcasting” is well established (Demargne et al. 2014). In essence, the complete hydrologic forecasting system is run in a retrospective process to effectively create the set of forecasts that would have been generated over an adequately long period of time. That period of time is normally constrained by the availability of numerical weather prediction (NWP) models, used to force the hydrologic model set, to the last 25 or 30 years (Hamill et al. 2013). The process of generating the NWP hindcasts requires a great deal of computer resources and is

FEATHER RIVER - LAKE OROVILLE (ORDC1)

Latitude: 39.53° N Longitude: 121.52° W Elevation: 992 Feet

Location: Butte Country in California

River Group: Lower Sacramento

Issuance Time: Feb 28 2014 at 9:13 AM PST

Monthly Probability Plot

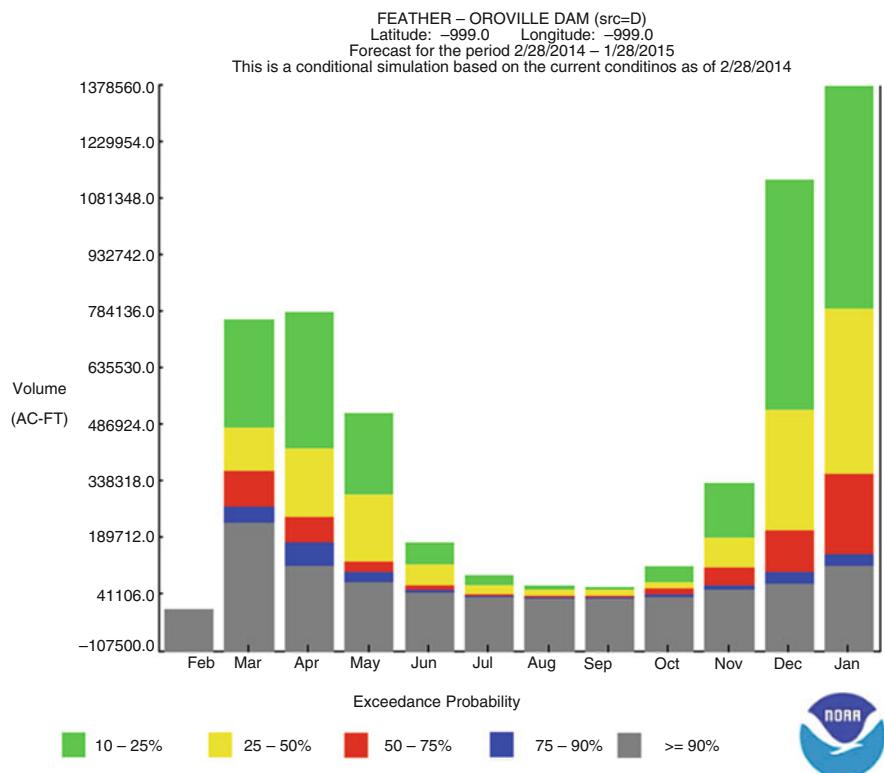


Fig. 9 Monthly volume histogram for 1-year for the Feather River inflow to Lake Oroville in California

therefore expensive. Once generated, the NWP hindcast serves as a rich dataset that can be used to understand the behavioral climatology of the specific NWP. Everyone accepts that NWPs are not perfect. They are biased to some extent and exhibit uncertainty. The hindcasts allow for measurements of the bias (difference between forecasts and observations) and uncertainty (correlation between forecasts and observations). This hindcast analysis information allows for the proper interpretation “today’s” NWP model run and the effective integration of the NWP forecast information into the hydrologic ensemble forecast process. One approach for doing that is well described by Demargne et al. (2014).

FEATHER RIVER - LAKE OROVILLE (ORDC1)

Latitude: 39.53° N Longitude: 121.52° W Elevation: 922 Feet

Location: Butte County in California

River Group: Lower Sacramento

Issuance Time: Feb 28 2014 at 9:49 AM PST

2014 Seasonal Trend Plot (Year View)

[Tabular View](#) | Select a Different Water Year:

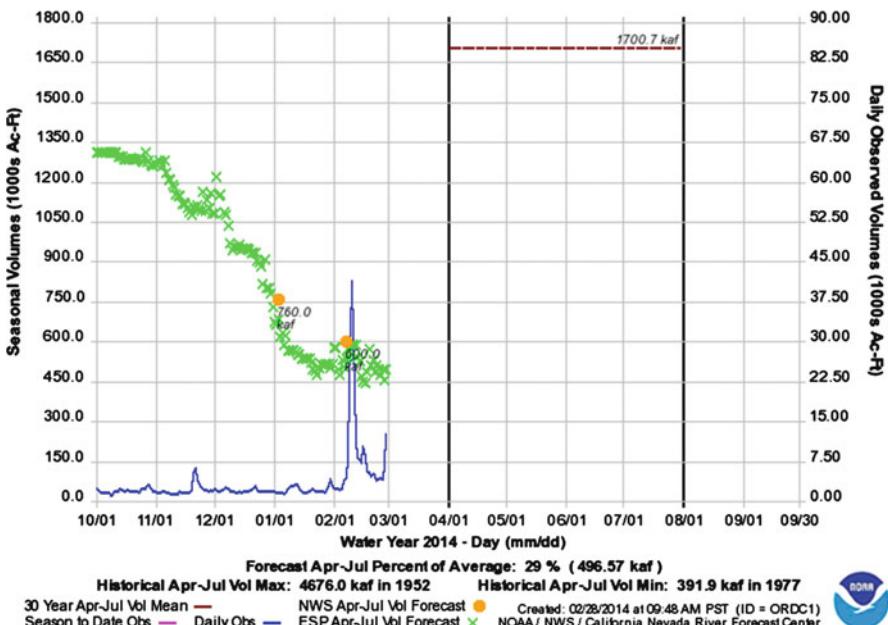


Fig. 10 Seasonal (April-July) accumulated volume trend plot for the Feather River inflow to Lake Oroville in California

It is important to note that the NWP hindcasts are specific to a model and the parameterization at the time of hindcast generation. If changes are made to the NWP, the hindcasts would simply no longer apply and would need to be rerun, reanalyzed, and reintergrated into the hydrologic ensemble forecast process. Further, customers of the hydrologic ensembles may need to make adjustments to their decision models to accommodate any resulting shifts. For these reasons, it is critically important that “frozen versions” of NWPs are operationally supported when the user community is dependent upon representative hindcasting information. Further, it is important to recognize that both the hydrologic forecast community and the user community need time (months) to integrate new NWP hindcast information before a “frozen version” is operationally discontinued.

While the hindcast process provides a way to understand the behavior of the complete forecasting system and resulting information, it is not perfect. Fully replicating the somewhat interactive hydrologic forecasting process in practice today is not feasible. It is generally accepted that hydrologic forecasters add value

FEATHER RIVER - LAKE OROVILLE (ORDC1)

Latitude: 39.53° N Longitude: 121.52° W Elevation: 922 Feet

Location: Butte County in California

River Group: Lower Sacramento

Issuance Time: Feb 28 2014 at 10:11 AM PST

2014 Water Year Trend Plot

[Tabular View](#) | [Select a Different Water Year:](#) 2014

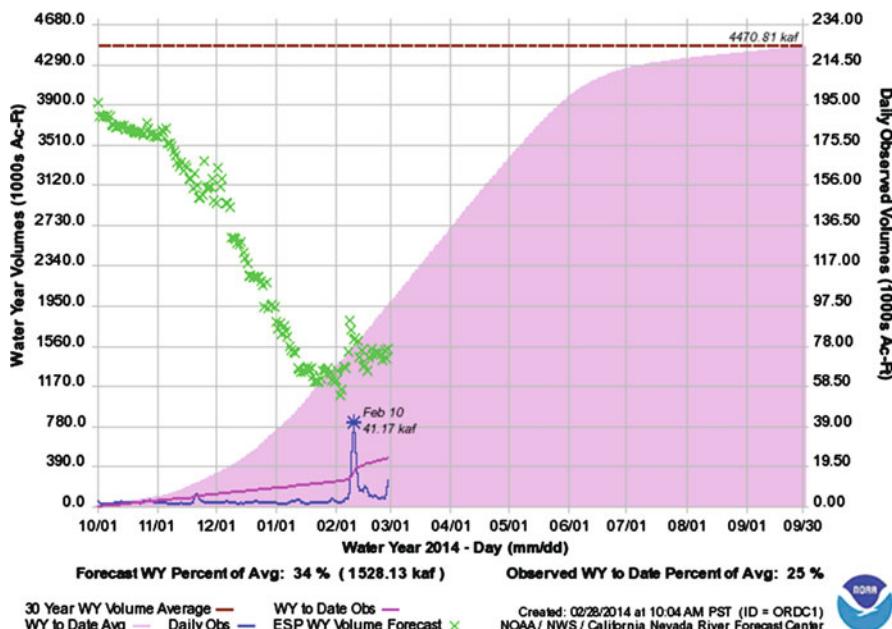


Fig. 11 Water Year (October-September) accumulated volume trend plot for the Feather River inflow to Lake Oroville in California

(reduce errors) through interaction with the hydrologic forecasting modeling system. This might take the form of small adjustments to forcing data (precipitation or air temperature) or adjustments to model states to better align model simulations with observations of streamflow during the recent observed period. Even if you were able to insert a forecaster into the hindcast process (very labor intensive), it would be extremely difficult to replicate the full forecasting environment that influences human decision-making. As such, the hydrologic ensemble hindcasts are an approximation of what we should expect from the current forecast process, but they may exhibit slightly more uncertainty as they do not benefit from forecaster experience and interaction.

4.1 Validation and Associated Services

With all their conditions and issues, hydrologic ensemble forecast hindcasts offer keen insight into the value of current probabilistic hydrologic forecasts. They provide the body of information that allows for the development of trust. Ensembles allow for a

FEATHER RIVER - LAKE OROVILLE (ORDC1)

Latitude: 39.53° N Longitude: 121.52° W Elevation: 922 Feet

Location: Butte County in California

River Group: Lower Sacramento

Issuance Time: Feb 28 2014 at 10:25 AM PST

2014 Water Year Accumulated Volume Plot

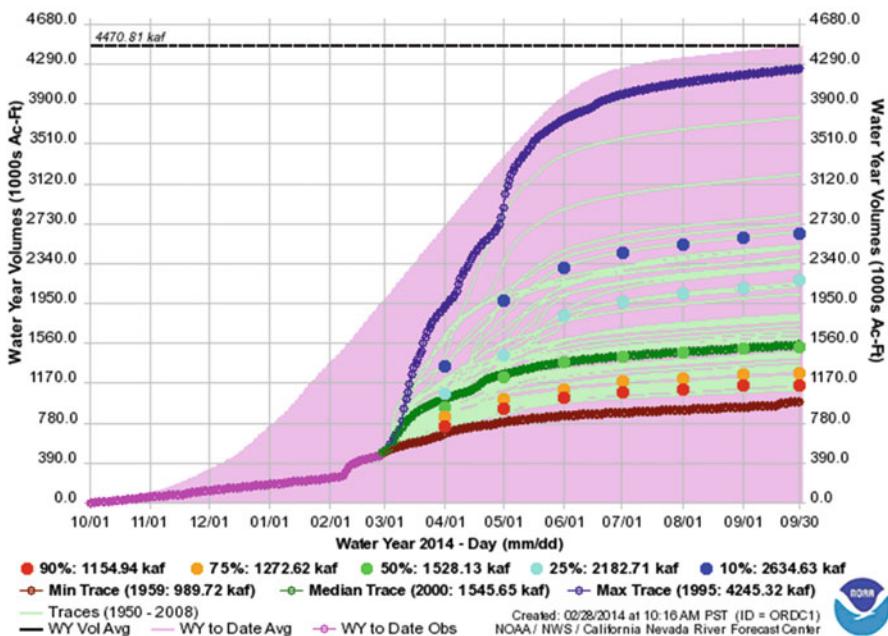


Fig. 12 Probabilistic water year (October-September) accumulated volume trend plot for the Feather River inflow to Lake Oroville in California

nearly infinite number of questions to be addressed. In turn, the hindcast analysis allows one to assess the reliability of the information used to address those very same questions.

Ensemble verification capability such as those described by Brown et al. (2010) provide the flexibility and rigorous statistical testing needed. Substantial education and training are needed to help consumers of this information to fully understand the implications of the ensemble forecast verification metrics and how they affect their specific decision-making process. Substantial work will be required to create validation information that is general enough to apply to most cases and specific enough to build/demonstrate value and trust.

5 Conclusion

The pace of improvement in hydrologic forecasts is steady, but very slow. Rather than waiting for the perfect forecast, a great deal of value and insight can be gained by understanding and leveraging the uncertainty associated with today's forecast.

Resource managers are pressed harder every day to work “smarter.” Integrating risk into every aspect of decision-making is warranted as long as the information being used is reliable and understood. While hydrologic forecasters and water resource managers have years of experience in using probabilistic seasonal streamflow volume forecasts, the notion and technology of short- and medium-range probabilistic hydrologic forecasts is quite new. Developers, forecasters, and users are challenged to create, provide, and integrate probabilistic information that will yield understanding and improved outcomes for end users.

References

- J.D. Brown, J. Demargne, D.-J. Seo, Y. Liu, The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorologic and hydrologic variables at discrete locations. *Environ. Model Software* **25**, 854–872 (2010)
- G.N. Day, Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan Manag.* **111**, 157–170 (1985)
- J. Demargne, L. Wu, S.K. Regonda, J.D. Brown, H. Lee, M. He, D.-J. Seo, R. Hartman, H.D. Herr, M. Fresch, J. Schaake, Y. Zhu, The science of NOAA’s operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* **95**, 79–98 (2014)
- D.C. Garen, Improved techniques in regression-based streamflow volume forecasting. *J. Water. Resour. Plan Manag. Am. Soc. Civil Eng.* **118**(6), 654–670 (1992)
- K.P. Georgakakos, N.E. Graham, A.P. Georgakakos, H. Yao, Demonstrating Integrated Forecast and Reservoir Management (INFORM) for Northern California in an operational environment. *IAHS Publ.* **313**, 1–6 (2007)
- T.M. Hamill, G.T. Bates, J.S. Whitaker, D.R. Murray, M. Fiorino, T.J. Galarneau, Y. Zhu, W. Lapenta, NOAA’s second generation global medium-range ensemble reforecast dataset. *Bull. Am. Meteorol. Soc.* **94**, 1553–1565 (2013)
- National Research Council of the National Academies, *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts* (The National Academies Press, Washington, DC, 2006), p. 124



Saving Lives: Ensemble-Based Early Warnings in Developing Nations

Feyera A. Hirpa, Kayode Fagbemi, Ernest Afiesimam, Hassan Shuaib,
and Peter Salamon

Contents

1	Introduction	1110
2	Regional Disparities of Disaster Impacts	1112
2.1	Flood Disaster	1112
2.2	Drought Problem	1114
2.3	Early Warnings from Ensemble-Based Systems Could Save Lives	1115
3	The State of Early Warning Systems in Regions with Fragmented Infrastructure: Nigeria Case	1116
3.1	The 2012 Flood in Nigeria	1117
3.2	The Need for Early Warning Systems in Nigeria	1117
3.3	Ensemble Forecast for the 2012 Flood	1119
4	Global Partnerships to Address Local Problems: Joining Forces	1123
4.1	Global Flood Partnership	1123
4.2	Integrated Drought Management Programme	1126
5	Conclusions	1127
	References	1127

F. A. Hirpa (✉) · P. Salamon

European Commission, Joint Research Centre (JRC), Institute for Environment and Sustainability (IES), Climate Risk Management Unit, Ispra, VA, Italy
e-mail: feyera-aga.hirpa@jrc.ec.europa.eu; peter.salamon@jrc.ec.europa.eu

K. Fagbemi

National Emergency Management Agency (NEMA), Abuja, Nigeria
e-mail: kayodef@hotmail.com

E. Afiesimam

Nigerian Meteorological Agency (NiMet), Abuja, Nigeria
e-mail: ernafies@yahoo.com

H. Shuaib

University of Abuja, Abuja, Nigeria
e-mail: hassanalabo2@gmail.com

Abstract

Natural disasters disproportionately affect the developing nations due to the lack of effective early warning systems. In this chapter, we present the need, challenges, and opportunities of early warning systems in developing nations for decision making in disaster risk management and demonstrate the added value of ensemble forecasting in particular in data- and infrastructure-scarce regions. First, we review the global extent of flood and drought disaster damages in the last few decades on human lives and the economy and demonstrate that a disproportionately high rate of death (per event) occurred in developing regions, where there is no (or ineffective) operational early disaster warning systems. Next, we present the everyday needs and challenges of preparing for and responding to natural disasters in Nigeria, a typical developing country with fragmented data infrastructure and limited national early warning system capability. Particularly, we share experiences from the most recent major flood disaster and demonstrate a potential value of ensemble-based flood early warnings, using streamflow forecasts from the Global Flood Awareness System (GloFAS).

However, forecasting of disasters alone is not sufficient if the information is not translated into actionable advice at a local community level. This is particularly important for ensemble forecasting which requires training for the forecasters as well as the receiving authorities. In order to achieve this, technical knowledge and communication infrastructure are needed to deliver the early warning information to the relevant communities and concerned authorities. Multi-stakeholder partnerships bringing together scientific community, policy, and decision makers and end users from international to local level could facilitate humanitarian aid organizations, and decision makers understand and use the ensemble predictions on timely basis before, during, and after disaster strikes. The chapter concludes with highlighting the multi-stakeholder partnership initiatives on floods (Global Flood Partnership (GFP)) and droughts (Integrated Drought Management Programme (IDPM)), established with the common goal of reducing flood and drought risk across the globe.

Keywords

Ensemble forecasting · Early warning · Risk reduction · Disaster response · Global Flood Partnership · Integrated Drought Management Programme

1 Introduction

Floods and droughts range among the world's deadliest and costliest natural disasters. Since the turn of the twentieth century, almost 5000 major hydrological disasters have occurred (CRED 2015); more than seven million people have been killed; over 3.5 billion affected; and estimated damages of 650 billion USD are reported. During the same period, droughts have been responsible for over 11 million deaths and 134 million USD of economic damages (CRED 2015; Hirpa et al. 2016). While these disasters occur worldwide, they disproportionately affect the poor, with

more than 97% of the associated deaths occurring in the developing world and the disasters amounting to a significant portion of their GDP (Pilon 2002; CRED 2015).

Therefore, national authorities as well as the international community are seeking to cope with and adapt to the phenomena and to reduce the impact through risk prevention policies. Effective risk prevention policy extends over a set of long- to short-term actions. Long-term measures include action on prevention and adaptation such as planning (control of urban development and construction following legal rules and risk prevention master plans), construction works resilient to disaster risk, as well as educating experts, decision makers, and the public. Shorter-term measures include the monitoring, forecasting, and warning before disaster strikes. The importance of strengthening early warning systems and to tailor them to the needs of the users (including quantification and communication of uncertainty) remains one of the priorities of the *Sendai Framework for Disaster Risk Reduction 2015-2030* (UNISDR 2015).

Forecasting hazards and the associated uncertainty are an important element of any early warning system. However, to fully understand the risk and consequently translate it into actionable warning information, it needs to be complemented with information on exposure, vulnerability, and impact over a wide range of spatial and temporal scales. For instance, major flooding in a rural area with no population may not appear critical while the same magnitude flood could cause significant damage in urbanized area with large population. However, a flooding in a generally non-populated area can turn into a disaster in case temporary camps or public events are organized when the disaster strikes. An effective risk reduction procedure, thus, should take all the three components – hazard, exposure, and vulnerability – into consideration.

Many of the methods, tools, and applications described in this handbook require a functioning data collection and sharing infrastructure, remote sensing tools, hydrometeorological models with different levels of sophistication, and broad expert knowledge in scientific methods and data handling. Furthermore, platforms for communicating the information to end users and decision makers and, at the end of the chain, skilled end users that understand the hazard information are needed for an effective risk reduction. While numerous initiatives and services exist, still such prerequisites are not available in several, particularly developing, countries, and, consequently, there are often a large number of human lives lost due to natural disasters as indicated in this chapter.

To this end, we discuss the regional variations of disaster (flood and drought) impacts using data from global disaster archives. Next, using the example of Nigeria, a country with limited early warning infrastructure, we present the challenges of coping with disasters in developing countries. Furthermore, we demonstrate the potential value of an ensemble prediction system (ENS) using flood forecasts from the Global Flood Awareness System (www.globalfloods.eu) for the unprecedented 2012 flood disaster in the country. Finally, an outlook to new initiatives on global flood and drought partnerships for risk assessment and reduction are highlighted. The partnerships have goals of closing the gaps in disaster risk management with regard to floods in particular for the developing countries but with added value products also for developed countries.

2 Regional Disparities of Disaster Impacts

2.1 Flood Disaster

Since the year 2000 (as of September 2015), close to two thousand riverine flood disasters were recorded worldwide (CRED 2015; see Table 1). More than one third (37%) of the disasters during this period occurred in Asia; a quarter of those were recorded in Africa; and close to one fifth (18%) of the events were reported in the Americas (excluding Canada and the USA). This means that four out of five of the total flood disasters reported worldwide occurred in the developing regions (denoted “Region I” in this chapter for brevity), while the remaining part of the globe (Region II: Europe, Oceania, Canada, and the USA) experienced 20% of the total count of riverine flood disaster during the same period. Note that this broad regional classification is used differently from the World Meteorological Organization’s (WMO) six regional associations (WMO 2014).

In terms of impact, more than 67,000 people have lost their lives, and a total economic damage of more than 34.5 billion USD was caused worldwide due to riverine floods since the turn of this century. Asian continent was the most affected: accounting for 69% of the total deaths, 93% of the total people affected (e.g., injured or displaced), and 62% of the total economic damage since the year 2000. The Caribbean and Central and South (CCS) America and Africa had the second and the third most recorded deaths at 10,484 (15.5%) and 9,263 (13.7%), respectively. As shown in Fig. 1, 97.6% of the total deaths attributed to riverine floods occurred in the developing regions (Region I). This is significantly higher than the proportion of the reported flood occurrence (80.1%) in these parts of the globe (see Fig. 1). Similarly, a disproportionately high number of people were affected by flood (e.g., injured or displaced) in Region I. This disparity could be attributed to, among others, the lack of early warning systems in the developing nations. Several countries in the developed regions (Region II) have operational flood forecasting services that help them

Table 1 Global flood disaster occurrence and its impact since year 2000 (as of September 2015) (Data was obtained from CRED (2015))

Regions		Flood occurrence	Total deaths	Total affected (millions)	Total damage (million USD)
Region I	Asia	712	46,452	1148.8	214,237
	Africa	473	9,263	38	4,591
	Caribbean and Central and South (CCS) America	346	10,484	31.2	19,097
Region II	Europe	255	1,108	6.9	63,589
	USA and Canada	82	362	11.6	33,025
	Oceania	43	125	0.59	11,062
Total		1911	67,794	1237.09	34,5601

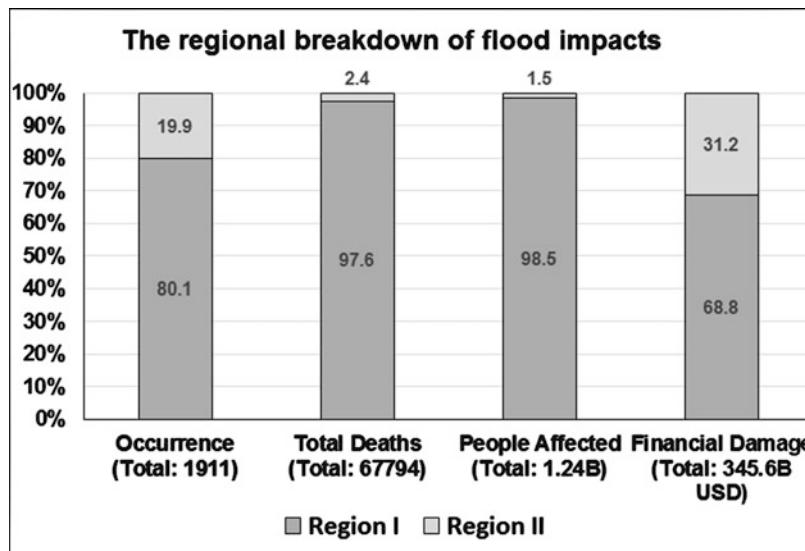


Fig. 1 The regional breakdown of flood occurrence and impact for the last 15+ years (2000–September 2015). Region I includes Asia, Africa, South and Central America, and the Caribbean. Region II includes Europe, Oceania, Canada, and the USA (Data was obtained from CRED ([2015](#)))

prepare for an upcoming flood disaster. For example, in Europe, local, regional, and national as well as continental flood forecasting services complement each other ([www.efas.eu](#); Thielen et al. 2009) to provide early warning information; Meteorological Model Ensemble River Forecasts (MMERF) produce operational streamflow forecasts across the USA (NWS [2015](#)); several flood forecasting centers provide operational forecasts across Canada (EC [2015](#)); and Australia has a national-scale operational flood forecasting and warning service (BoM [2015](#)).

Further, looking at the impact of flood per occurrence across different regions reveals that for every flood event that occurred since year 2000, an average of 65 people lost their lives in Asia, while 30 in CCS America and 20 in Africa were killed (see Fig. 2). To relatively a lesser extent, every flood event during the same period caused deaths of four persons in Europe (similar to the USA and Canada combined). This means that about 16 times as many deaths in Asia were caused by a flood disaster as the number of deaths recorded in Europe. The total number of people affected by a flood event tells a similar story (also shown in Fig. 2). There were substantially more people affected in Asia per flood event compared to other regions (e.g., 60 times as many as in Europe and 11 times as many as in the USA and Canada). The larger impacts on human life seem to be a direct reflection of (besides the lack of early warning systems) the high vulnerability and the lack of coping capacity to the flood hazard in developing countries (De Groot et al. [2014a](#)).

Interestingly, the reported economic damages due to flood are in a stark contrast to the loss of human lives. They were higher (compared to the proportion of the

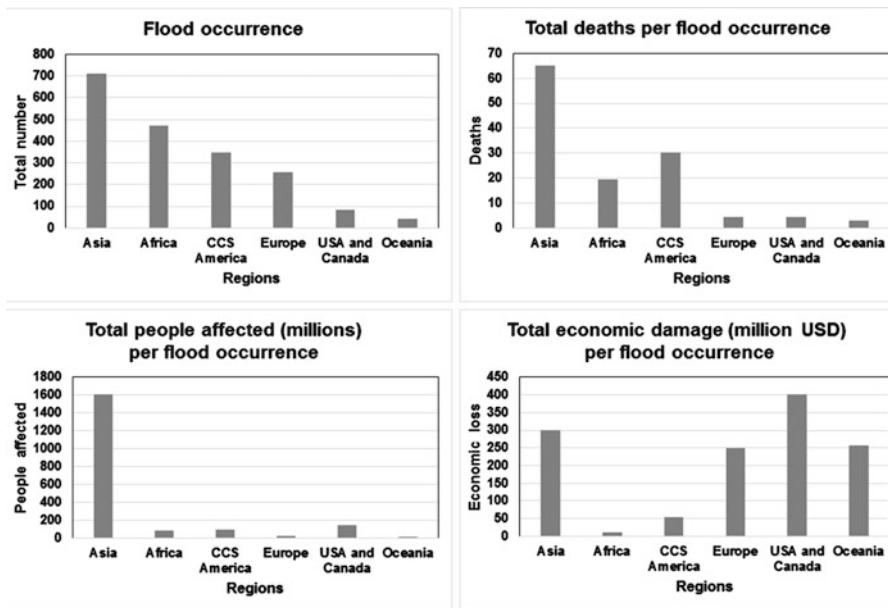


Fig. 2 The impacts of riverine flood per occurrence (since 2000) for different regions and continents (Data was obtained from CRED (2015))

occurrence) in the developed regions (i.e., Europe, Oceania, USA, and Canada): 31.2% of the total damage compared to the 19.9% of the total flood disaster count globally (see Fig. 1). This can be attributed to the relatively higher economic assets and infrastructure exposed to flood hazard in those developed regions (Region II) compared to the Region I. The average cost of a flood event in the USA and Canada is the highest at 400 million USD. Moreover, the damage in each of Europe and Oceania per flood event is 250 million USD. While Asia experienced an average loss of 300 million USD per one event, Africa (10 million USD) and CCS America (55 million USD) suffered comparatively lower damages per flood event (Fig. 2). The data from CRED shows that the economic damages of the flood hazard are more concentrated in developed countries.

2.2 Drought Problem

Even though it occurs less frequently than flood worldwide, drought causes significantly higher socioeconomic damages and humanitarian crisis per event. It affects large population mainly due to its large spatial extent (sometimes extending to a continental scale) and long duration (possibly lasting multiple years) (Haile 2005; Sheffield and Wood 2011; Dutra et al. 2012). Since 1980, a total of 519 drought events worldwide have killed more than half a million people and caused a loss of 129 billion USD (see Fig. 3). A large majority (86.3%) of these events were reported in developing nations (Region I), and almost all people affected belong to these regions. Notably,

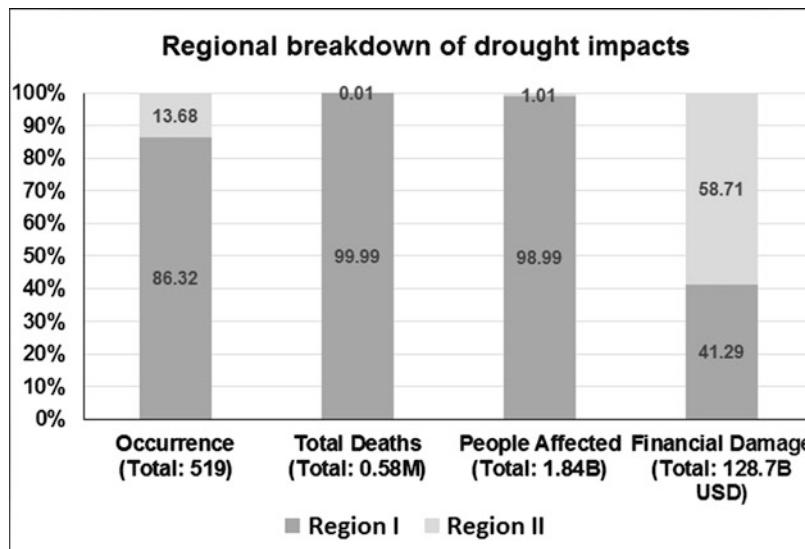


Fig. 3 The regional breakdown of the global drought events and their impacts since 1980 (as of September 2015). Region I includes Asia, Africa, South and Central America, and the Caribbean. Region II includes Europe, Oceania, Canada, and the USA (Data was obtained from CRED ([2015](#)))

the 1983/1984 drought in Ethiopia and Sudan accounted for 450,000 deaths, which remain to be the biggest share of causalities caused by a drought event (CRED [2015](#)). The 2010/2011 drought in the Horn of Africa resulted in humanitarian crisis affecting more than 10 million people. Similarly, between 2002 and 2009, different drought events affected hundreds of millions of people in China and India.

Since 1980, a drought event has killed, on average, more than 1120 people and has affected 3.5 million people worldwide (CRED [2015](#)). A continental break down reveals that almost all of the deaths were recorded in Africa (2560 persons killed per drought event) and Asia (45 persons). The African continent depends largely on rain-fed agriculture which is highly affected by the climate variability (Sheffield et al. [2013](#); Dixon et al. [2001](#)), and, therefore, drought has a direct devastating impact on the food production and supply of the continent. While the overall economic loss due to drought was the highest in the USA (3.1 compared to 0.01 billion USD per event in Africa), there was no reported loss of human life as a direct consequence of drought in the USA. This clearly indicates that drought causes humanitarian crisis (due to shortage of food supply) disproportionately in the poorer nations.

2.3 Early Warnings from Ensemble-Based Systems Could Save Lives

The importance of disaster early warning systems in developing nations cannot be overstated. As demonstrated above, the flood disasters were more frequent and they

disproportionately killed more persons per event in regions with developing countries (Asia, Africa, Caribbean, and South and Central America). Similarly, the largest share of droughts occurred in these regions, and a strikingly large portion of causalities (99%) over the last 35 years were reported in Africa.

Early flood warning systems can raise awareness of, and increase preparedness for, an upcoming disaster and hereby reduce the risk on human lives. The regions with (some sort of) operational early flood warning systems experienced significantly lower loss of human lives to natural disasters. A large majority of the countries in Region II have operational hydrologic ensemble prediction systems (HEPSs; Pappenberger et al. 2015; Cloke and Pappenberger 2009) providing near real-time ensemble streamflow forecasts at national or continental level. While increasing the local technical knowledge and management capacity for communicating early warning information to the concerned authorities and the public are equally important, the advanced warnings play a key role in predicting the extent of the foreseen hazard. To this end, implementing effective flood forecasting and early warning systems in developing regions will save a considerable number of human lives. This is particularly significant in light of the number of persons losing life for every flood event (e.g., 65 persons in Asia) mainly due to the lack of effective advanced warnings.

Drought early warning systems and timely seasonal forecasts may greatly help farmers and pastoralists plan to protect their crops and livestock in developing nations (Sheffield et al. 2013; Pozzi et al. 2013). However, most of the developing countries do not have advanced seasonal forecasting due to lack of reliable monitoring network, low institutional capacity, and lack of national drought policies (Sheffield et al. 2013; Naumann et al. 2014). This is particularly evident in sub-Saharan Africa where seasonal climate forecasts are generated by regional climate outlook forums (Ogallo et al. 2008) based primarily on statistical regression without detailed information needed for agricultural adaptations (Sheffield et al. 2013; Patt et al. 2007). Skillful seasonal climate forecasts with long lead times (up to 6 months) could help the national authorities plan for storing and transporting food to the vulnerable communities and inform the farmers in advance to relocate their livestock to a water-available area. Moreover, early drought forecast may help humanitarian aids (e.g., World Food Programme (WFP)) to plan for rapid disaster response. The ensemble prediction system provides a range of hazard scenarios which allow emergency managers to consider all likelihoods and prepare for actions.

3 The State of Early Warning Systems in Regions with Fragmented Infrastructure: Nigeria Case

A practical implementation of an effective flood early warning system in developing regions with fragmented infrastructure has several challenges. The lack of adequate data, advanced modeling and communication infrastructures, and the limited technical knowledge for understanding the extent of the disaster obstruct the forecasting

and monitoring efforts. Using the experiences from the devastating flood of 2012 and in Nigeria, we discuss the need for and the challenges of an effective advanced flood warning in developing nations. Furthermore, we demonstrate the value of the ENS flood forecasting using forecasts from the Global Flood Awareness System.

3.1 The 2012 Flood in Nigeria

Floods are the most common and recurring disaster in Nigeria (PDNA 2013). While they impact the country each year, the damage and losses from the 2012 floods were unprecedented. Heavy rainfall between July and October 2012 combined with rising water contributed to the flooding of human settlements located downstream of the Kainji, Shiroro, and Jebba dams on the Niger River; the Lagdo Dam in Cameroun on the Benue River; the Kiri Dam on the Gongola River; and several other irrigation dams. In some cases, the dams were damaged; in others, water had to be released at full force to avert a dam overflow.

The reports from the National Emergency Management Agency (NEMA) showed that between July and the end of October 2012, about 7.7 million people were affected by the floods, and over 2.1 million persons were displaced (IDPs), while 363 persons were reported dead. In addition, more than 600,000 houses were damaged or destroyed (PDNA 2013). The worst flooding in Nigeria for over four decades was experienced as a result of high rainfall intensity and the emergency opening of the spillway gates of the Kainji Dam on the Niger River and the Lagdo Dam on the Benue River. This disaster was partly man induced as an uncontrolled release of the waters upstream from dams compounded the problem.

3.2 The Need for Early Warning Systems in Nigeria

Currently an advanced, basin-wide early warning system is not available in Nigeria and in many parts of the developing world mainly due to the lack of data and modeling infrastructures. The lack of early warning makes it difficult to have sufficient time for individuals and communities to act to reduce flood risk, loss of life, and damage to property and the environments. For the flood warning to be effective, it should be issued sufficiently early prior to the potential inundation to allow adequate preparation (Kundzewicz 2013). The appropriate time frame is affected by the catchment size relative to the vulnerable zones. The warning should also be expressed in a way that persuades people to take appropriate action to reduce damage and costs of the flood. An assessment of risks provides the basis for an effective warning system by identifying potential threats from hazards and establishing the degree of local vulnerability and resilience to extreme weather conditions.

In Nigeria efforts are being made in enhancing monitoring weather and climate facilities and providing early warning advisories, but it requires far reaching steps in contributing to the mitigation of the extreme weather impacts

on the community. However, there are several limitations in the data, modeling, and communication infrastructures which obstruct the monitoring efforts. For example, ground data networks in Nigeria for monitoring weather and climate elements are sparse with low station coverage density; the upper air monitoring network is even more limited. It is also rather sad that the station networks in many cases were originally installed to support aviation services (due to the international requirements) without expansion to wider coverage in the country – a situation that now limits the applicability of the limited data to other applications such as flood and drought monitoring, water resources, and disaster risk reduction. Without adequate ground-based measurement data, meteorological services rely more on circulation model outputs from outside the country whose parameterization may not take into account the specificity of the country region, or they rely on satellite products despite the lack of accuracy verification associated to these products, thus having large uncertainty in the spatiotemporal distribution and intensity of rainfalls.

Seasonal forecast tools have also been tailored more toward volumetric rainfall totals (or probability of rainfall totals expressed as a percentage of the long-term mean) that do little to offer guidance to users requiring more specific information. For example, in order to be informed comprehensively, the vulnerable communities would not only need information about the probability of the total rainfall of a rainy season to be above or below the long-term mean but more importantly on the timing, intensity, and volume of rainfalls. A farmer is interested in the timing of the onset of the rainy season (the separation between the real and false onset dates), the distribution of the rains within the season relative to the germination, maturity, and harvest stages of the crops, more than the cumulative rainfall of the season. In Nigeria, the seasonal rainfall prediction is usually released between January and February every year. However, the prediction does not provide specifics with the magnitudes and timing of the rainfall, and there is no uncertainty information provided. For example, the 2014 seasonal rainfall prediction in Nigeria stated, “the 2014 growing season is predicted to experience normal onset across the country except in and around Gusau and Yola in the North, Shaki and Abeokuta in the Southwest which are predicted to experience delayed onset” ([NiMet 2014](#)). This will be inadequate for a farmer who will want to know the dates he can start planting.

A long lead-time (up to a month or longer) flood may allow those involved in agriculture and pastoral activities to make preparations to protect assets and reduce the impacts of a flood. For example, if it is known that a significant flood may occur near the end of a growing season, farmers may plan ahead of time to harvest crops early. Or if a significant flood is anticipated early in a season, planting may be delayed, and different varieties of seed may be chosen that have a shorter growing season or are better adapted to saturated soils.

The two major rivers that traverse the country are transboundary which complicates monitoring of potential floods. Furthermore, Nigeria has more than 20 dams ([FAO 2004](#); and increasing) which were constructed for multiple purposes. Some are designated as multipurpose dams to provide water for irrigation in the dry season and hydroelectric and domestic uses. But in reality, in some cases, the water is

impounded and is neither used for hydroelectric generation nor irrigation. Compounding this is the siltation that has reduced the carrying capacities of the dams. To avoid a situation where water will have to be released suddenly, a short-term (24 h to 3 days) skillful precipitation forecast and river discharge would be useful.

Dissemination of disaster information to the disaster management agencies, such as the National Emergency Management Agency (NEMA), and the communities at risk is also a crucial step for making plans for early disaster response. The early warning message-based ensemble forecast with attached uncertainty is preferable for disaster managers or civil protection since they can plan their actions based on the probability scenarios. For example, it is generally followed that forecast probability of less than 70% should be communicated to the NEMA and state agencies, while forecast probabilities above 70% should be communicated to all stakeholders across the board from the national to the local communities. However, without ensemble-based prediction systems in Nigeria, it is difficult to implement these policies.

Understanding forecast for flood and drought, no doubt, requires training of the users. The agencies responsible for the communication of early warning messages must train users of the information to avoid misinterpretation in forecast messages. For example, the 2012 flood in Nigeria caused so much damage to farmers that when Season Rainfall Prediction was released in 2013, some farmers were unwilling to invest in crop production for fear of losing their investment to flooding because they did not clearly understand the rainfall prediction. This was attributed to the way and manner the prediction was communicated to the communities through general mass media (TV and radio), but not through organizational communication with more attention to the needs of the vulnerable communities. Communication through television tends to be ineffective in reaching a rural community in Nigeria due to high proportion of the inhabitants who have no access to electricity.

To accommodate the diverse economical, language, social, and infrastructure levels in Nigeria, wide ranges of commutation channels need to be used. Disaster warning messages could be sent, when possible, over telephones and mobile devices, on the Internet, and through social media and printed media. There are also well-organized indigenous mass communication systems (e.g., influential community and religious leaders) that can be used to disseminate emergency information to the rural communities. To fully benefit from ensemble forecast, implications of the forecast for flood and droughts should be analyzed and communicated to experts in hydrology, agriculture, transport, health, civil protection, and tourism and dam managers. The ever improving means of communication, through advancement of information and communication technologies, provides new opportunities of making the communication of disaster to rural communities faster and more reliable (Ewolabi and Ekechi 2014).

3.3 Ensemble Forecast for the 2012 Flood

Here, we examine the ensemble flood prediction from the Global Flood Awareness System (GloFAS; www.globalfloods.eu/) for the 2012 flood in Nigeria. The GloFAS

(Alfieri et al. 2013) provides global ensemble-based flood early warning information for major rivers worldwide. It produces probabilistic flood forecasts based on a modeling setup combining the land surface model (LSM) Hydrological Tiled ECMWF Scheme for Surface Exchanges over Land (HTESSEL) (Balsamo et al. 2009) with a river routing scheme used by LISFLOOD (van der Knijff et al. 2010). HTESSEL is used operationally in the Integrated Forecast System (IFS) of the European Centre for Medium-Range Weather Forecasts (ECMWF) for describing the evolution of soil, vegetation, and snow over the continents at diverse spatial resolutions. The GloFAS combines state-of-the-art weather forecasts with a hydrologic model set up at a $0.1^\circ \times 0.1^\circ$ grid resolution to produce once a day 51-member ensemble streamflow forecasts and probability flood occurrence (in terms of threshold exceedance probability). The flood thresholds were extracted from long-term reforecasts (see Hirpa et al. 2016).

We investigate the GloFAS forecasts during the devastating 2012 flood at three locations (see Fig. 4) on the main streams of the Niger River (A) and Benue River (C) and on the downstream of the intersection of the rivers (B). The Post-Disaster Needs Assessment report from the Federal Government of Nigeria (PDNA 2013) indicated that states along the Benue River and lower reaches of the Niger River (below point B in Fig. 4) were considerably affected by the flood disaster with the most downstream states suffering the heaviest damages. Conversely, the damages reported in the states along the upper reaches of the Niger River (including the Niger



Fig. 4 Selected locations on the Benue and Niger rivers in Nigeria. The government report indicated severe flood damage in the states along the Benue River (location C), Kogi (location B), and downstream states (PDNA 2013)

State that includes point A) were relatively small. To see how the HEPS captures the streamflow in various regions with (and without) reported flood damages, we selected the three points to analyze the GloFAS forecasts.

Figure 5a–c shows the streamflow time series of a 5-day lead forecast produced by the GloFAS over a period of more than 2 months (from 5 August to 6 October 2012). The time period was selected based on the recorded start and end date of the flood event in Dartmouth Flood Observatory (DFO, Brakenridge 2015). We have extended both the start (20 days ahead) and the end (10 days further) dates of the flood to fully capture that the rising and falling limbs of the GloFAS forecast. In the figures, we also show 2-year, 5-year and 20-year flood levels.

At all three locations, the discharge magnitude increases with the peak occurring between 9 and 20 September 2012 (see Fig. 5a–c). However, the extent of the increase varied from one location to another. For the location in the state of Niger (location A), the ENS mean did not reach 2-year flood at any time throughout the duration of the flood. This is consistent with the Nigerian government's post-disaster report (see PDNA 2013) which documented no flood damage in this part of the river. The GloFAS indicated large flow magnitudes for the other two locations considered. The ENS mean at the point on the Benue River (in the

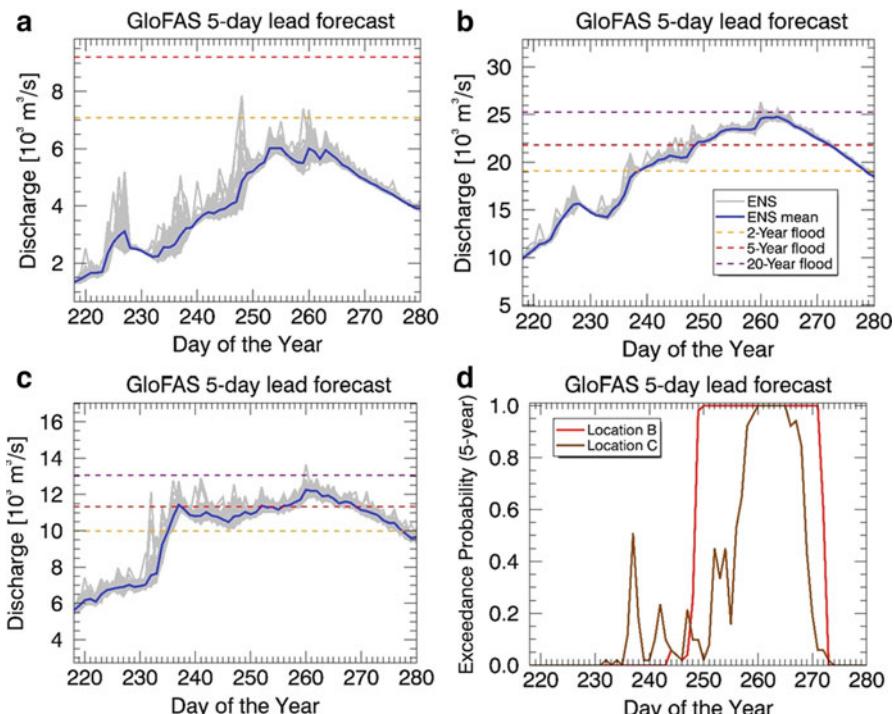


Fig. 5 GloFAS forecasts for the 2012 flood in Nigeria. Five-day lead forecasts at location (a), (b), and (c) are shown (see Fig. 4). And the 5-year flood threshold exceedance probabilities at location (b) and (c) are indicated

state of Benue, C) exceeded the 2-year flood threshold for 43 days (beginning from 21 August), and it came close to or exceeded the 5-year threshold for 36 days with the peak forecast on 16 September. The forecast at the downstream point (location B) showed a consistent increase in the flow volume until 20 September when it approached the 20-year threshold. This high-to-severe flow forecast is in agreement with the regions of the country (especially the downstream states) severely damaged by the flood event.

The most attractive feature of ENS compared to deterministic forecast is that, besides proving a prediction of just flood magnitude, it provides the uncertainty attached to it in a form of probability of occurrence of a certain flood magnitude. The ensemble prediction gives the ability to provide ranges of scenarios that may occur, and can enable the forecasters to identify the most likely outcome. We calculated a probability of the streamflow forecast exceeding three (2-, 5-, and 20-year flood) threshold levels. Figure 5d shows the probability of the 5-day lead forecast exceeding the 5-year return level for locations B and C (location A is excluded since no high flood was indicated). For the downstream location (B), a high 5-year exceedance probability ($p = 1.0$) was indicated over a 3-week period between 5 (rising) and 26 September (recession period of the flow). The forecast for the Benue River (C) indicated a high probability of 5-year exceedance after mid-September.

These probabilities would be valuable information for the National Emergency Management Agency since it decides on whether to disseminate early warning information based on these values (when the probability $>70\%$). For instance, if these forecasts were used by the NEMA and if there were GloFAS reporting points at the locations A, B, and C, the agency would have communicated flood alerts for 5 September to all stakeholders in the country from national authorities to the local communities in the Kogi State and all downstream of point B. Similarly, communities in the Benue State and surrounding would have been alerted for 12 September five days ahead. The 5-day lead forecasts may provide enough time to temporarily transfer vulnerable communities to low risk places and help save some lives.

It is known that the quality of the GloFAS forecasts, in the same manner as any forecast, needs to be verified through comparing ground-based observations. Nevertheless, these forecasts for the 2012 Nigerian flood demonstrate the potential value of an ENS and how it could assist disaster risk reduction efforts in developing countries without their own early warning systems. The challenge is, then, to translate the forecast information (including the freely available GloFAS) into actions that save human lives and the economy. This may require an integration of forecasted datasets (e.g., flood hazard) into the national early warning system and a creation of an effective channel of communication from the scientific communities and forecast centers to the local authorities, the vulnerable communities, and international disaster response organizations. Fragmented infrastructure and limited technical knowledge pose great challenges to preparedness, adaptation, and response to extreme weather and climate conditions in developing countries. This can be mitigated through partnerships with different stakeholders and raising the knowledge of local stuffs on the understanding and use of already available data from forecasting and observation (e.g., near real-time satellite detection) systems.

4 Global Partnerships to Address Local Problems: Joining Forces

As illustrated in the previous sections, floods and drought are global problems that require global, regional, and national coordination in order to mitigate its impacts. The increasing availability of advanced remote sensing technologies for earth observations and improving global climate and land surface models (GCMs and LSMs) provide an attractive platform for skillful flood and drought forecasting at global and regional scales; however, it is evident that not all services have the technological capacity yet to put such systems into practice. Therefore, regional and global solutions are key to overcome such limitations and can become particularly important for transboundary river basins where no or limited cooperation between the upstream and downstream countries exists (Nishat and Faisal 2000; Turton et al. 2003; Gerlak et al. 2011; Veilleux 2013).

In such cases, global and regional flood and drought early warning systems can provide national and local authorities with added value information they would otherwise not have access to and thus strengthen disaster preparedness and improve recovery efforts. The local agencies which naturally have easier access to on-site observations (both real time and historical) of a hazard and possess a better understanding of disaster risk (e.g., in terms of population and economic assets exposed) can provide a useful feedback for the forecasting and early warning centers. The feedback helps improve their performance and, in doing so, gain confidence from the local stakeholders in the usability of the forecast information.

Effective knowledge sharing about an upcoming disaster between scientists, operational forecasters, and local government agencies and communication to end users are necessary in order to optimally manage the associated risks. It is evident that this is much more challenging in countries with fragmented data infrastructures, low computational capacities, and lack of coherent early warning systems and disaster risk management policies than in highly developed industrialized countries. To overcome these challenges, global partnerships can help fill the gap that exists in the scientific data, modeling infrastructure, and effective knowledge sharing. To this end, there are global initiatives related to flood and drought disasters created with collaboration of multidisciplinary groups of scientists, operational forecasters, international disaster response organizations, and national stakeholders: Global Flood Partnership and Integrated Drought Management Programme.

4.1 Global Flood Partnership

The Global Flood Partnership (GFP, <http://portal.gdacs.org/Global-Flood-Partnership> and De Groot et al. 2014b) has been launched in March 2014 with a vision of bringing together the scientific community, decision makers, and end users for a common goal of raising global flood awareness and reducing disaster risks. The partnership is aimed at closing the gaps with regard to data availability, monitoring, detection, and forecasting both for flood hazard and risk assessments as well as

communication between all parties concerned. The overall objective of the Global Flood Partnership is “the development of flood observational and modeling infrastructure, leveraging on existing initiatives for better predicting and managing flood disaster impacts and flood risk globally” (GFP 2013). The partnership promotes sharing of flood-related information among the partners, and it is stated that “open data policy, where partners have access to data, tools, and services, [is] considered to be a cornerstone of the Partnership” (De Groot et al. 2014b).

The GFP has five components which work hand in hand to increase flood risk preparedness and organize effective disaster responses as summarized in De Groot et al. (2014b):

- Flood service and toolbox component provides operational flood detection, forecasting, mapping, and risk assessment services and shares tools (software, algorithm, and models) used for supporting such services. Besides, it creates a platform for historical analysis, validation, and comparison studies.
- Flood observatory creates a link between flood services and decision makers to facilitate the easy dissemination of flood information to stakeholders.
- Flood record collects, archives, and shares flood data (mainly from remote sensing but also ground observations) and related impacts which are helpful for understanding flood risk.
- User guidance and capacity building support the developing regions with limited or no flood early warning systems by providing timely information, tools, and expertise from the partnership.
- User forum creates a platform for knowledge exchange among scientists, local authorities, and disaster response organizations through trainings and user conferences.

There are several organizations actively involved in the GFP. These include academic and scientific organizations (e.g., universities, research centers) developing global-scale flood tools and services, national and continental hydrometeorological modeling and operational centers, private companies involved in flood risk analysis (insurance and reinsurance), remote sensing centers (e.g., NASA), disaster response organizations (e.g., World Food Programme, Red Cross, and World Bank), and the World Meteorological Organization. In contrast to a purely research-driven initiatives such as HEPEX, the Global Flood Partnership aims at providing operational services providing daily, global information on floods including medium-range ensemble prediction-based flood forecasts from different models, multiple satellite-based flood detection, short-term flood forecasting, and risk assessment tools, all of which will provide the disaster response organizations such as the World Food Programme with a wealth of information to assess the probabilities of critical situations coming up.

One of the systems providing global ensemble-based flood early warning information is the Global Flood Awareness System (GloFAS, www.globalfloods.eu). The GloFAS has been running preoperational since June 2011. Another central element of the GFP is the global flood risk assessment and mapping based on flood hazard,

vulnerability, and exposure (e.g., Winsemius et al. 2013; Ward et al. 2013). The global flood risk under current and future climate scenarios is produced and provided freely with Aqueduct Global Flood Analyzer (<http://www.wri.org/resources/maps/aqueduct-global-flood-analyzer>).

Regions with limited or no early flood warning systems are expected to highly benefit from the GFP. A first demonstration of the benefit of GFP was provided at the onset of the flood event in Malawi in January 2015. From 12 to 17 January, heavy rainfalls were observed in Mozambique and Malawi, affecting largely the Zambezi River Basin. These heavy rains were forecast well by the numerical weather prediction models, e.g., the probabilities of exceeding 300 mm of rain accumulated over a 10-day forecast period indicated heavy precipitation at the border of Mozambique and Malawi (Fig. 6) which coincided well with later observations.

Due to the heavy rains and previous flooding, the President of Malawi declared a state of emergency on 13 January 2015 for 15 districts. A number of humanitarian aid organizations started acting to provide assistance to the affected population. One day after the declaration of the state of emergency, the World Food Programme posted a request for more information to the partnership. On the same day, information from GloFAS and the public portal of the Emergency Response Coordination Centre (<http://ercportals.jrc.ec.europa.eu>) was shared with the community. On 15 January an extreme rainfall assessment for Malawi was presented by ITHACA (www.ithacaweb.org), and Vienna University of Technology communicated information on available satellite information which was then shared by WFP with the

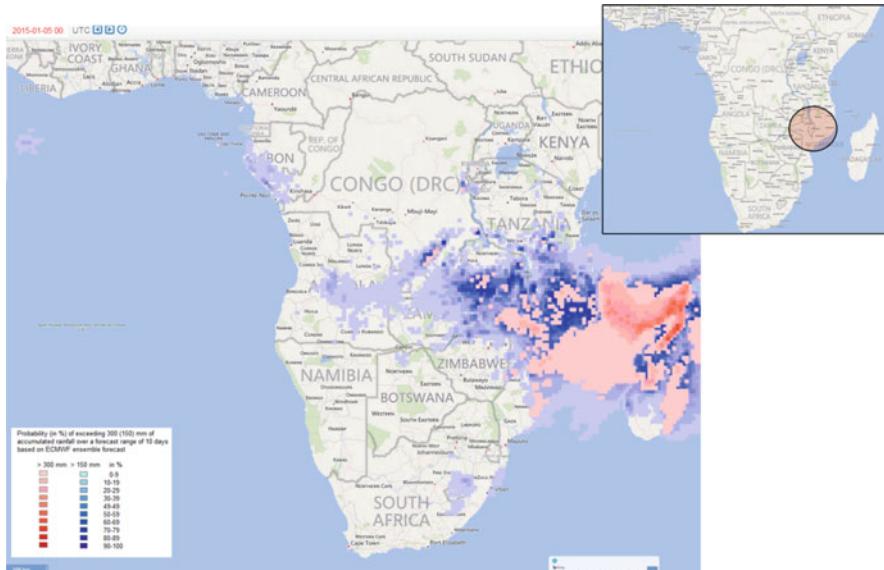


Fig. 6 Probabilities of exceeding 300 mm (150 mm) over a 10-day forecast shown in monochromatic red (blue) shading for the time period of 5–15 January 2015 based on ensemble prediction rainfall of ECMWF 00:00 UTC forecast

country offices. Also the Global Facility for Disaster Reduction and Recovery (<https://www.gfdrr.org/>) informed GFP that they got involved in aid management. On 16 January further information was exchanged by the Dartmouth Flood Observatory (www.dartmouth.edu/~floods), NASA, and UNOSAT on available satellite imagery and the activation of the UN charter. Thus within a few days a wealth of information ranging from ensemble forecasts over flood detection to information on aid management was shared within the community. This illustrates clearly how information flow can be enhanced, relevant content shared, and knowledge increased through the partnership.

In the future, GFP is planning to develop a more proactive approach with daily analysis of forecasts and flood detection information as well as the building of a flood record database. It is important to highlight that the partnership works in close collaboration with the national authorities and does not interfere with authoritative information such as national warnings nor override the national policies and communication channels.

4.2 Integrated Drought Management Programme

Effective drought risk reduction requires coordination of several stakeholders at different levels ranging from seasonal forecasting and risk assessment to designing proactive drought policies. To support the coordination efforts, the Integrated Drought Management Programme (IDMP) has been established as a joint initiative of the World Meteorological Organization (WMO) and the Global Water Partnership (GWP, <http://www.gwp.org/>). The central objective of the IDMP (<http://www.droughtmanagement.info/>) is “to support stakeholders at all levels by providing policy and management guidance and by sharing scientific information, knowledge and best practices for Integrated Drought Management” (WMO/GFP 2011). The coordination of already existing national and regional drought-related agencies and centers revolves around four central issues: better scientific understanding of drought and knowledge sharing; drought risk assessment, monitoring, prediction, and early warnings; policy and planning for drought preparedness and mitigation; and drought risk reduction and response.

Drought early warning, one of the core elements of the IDMP, plays a central role for planning and preparedness in drought risk reduction. A Global Drought Early Warning System (GDEWS, Pozzi et al. 2013) has been developed as an outcome of a partnership between several international organizations. The early warning system consists of monthly to seasonal (up to 15 months lead) ensemble drought forecasts worldwide provided by GDEWS partner organizations such as the European Centre for Medium-Range Weather Forecasts (ECMWF) and the US National Centers for Environmental Prediction (NCEP)/Climate Prediction Center (CPC). The skill of the ensemble-based forecast is continuously enhanced by model merging and downscaling techniques (Pozzi et al. 2013). The European Drought Observatory (EDO, Sepulcre-Canto et al. 2012) has demonstrated a good forecast skill over the European continent. A drought monitoring and forecasting for sub-Saharan Africa has been proposed and evaluated (Sheffield et al. 2013).

The IDPM operates in three regional programs (IDPM 2015) in Central and Eastern Europe (IDPM CEE) located in Bratislava, Slovakia; in the Horn of Africa (IDPM HoA) centered in Entebbe, Uganda; and in West Africa (IDPM WAF) located in Ouagadougou, Burkina Faso. Besides, there are two regional initiatives in South Asia (Colombo, Sri Lanka) and Central America (Tegucigalpa, Honduras). There are also a number of drought management efforts supported by the IDMP (IDPM 2015).

5 Conclusions

Natural hazards disproportionately affect people in the developing countries mainly due to the lack of an early warning system. Forecasting and early detection of hazards is an important element in preparedness and disaster risk reduction. Early detection of natural hazards allows the disaster response organizations to more effectively provide assistance to the affected and manage the recovery efforts. However, due to the inherent complexity and uncertainty of natural hazards, this remains a complex task. Ensemble forecasts indicate a set of possible hydrometeorological scenarios (e.g., flood predictions and possible hurricane tracks) and provide the confidence in the likelihood of these scenarios. These make the ensemble models attractive for assisting national governments and international organizations in planning for a wide range of disaster scenarios and have the potential to allow them to allocate resources more carefully.

However, effective early detection and warning systems require a functioning data collection and sharing infrastructure, hydrometeorological models with different levels of sophistication, and broad expert knowledge in scientific methods and data handling as well as platforms for communicating the information to end users and decision makers. Developing nations tend to have limited or no infrastructure for early warning systems (as discussed above with the Nigeria case) and, as a result, are less prepared and consequently disproportionately affected by disasters.

To fill these gaps, global or/and regional partnerships have been established to share scientific knowledge, models, and tools with regard to upcoming flood and drought disasters. This can be achieved by taking advantage of the increasing availability of remote sensing data (for disaster detection and monitoring), advanced hydrometeorological models for short-term (flood) and seasonal (drought) forecasting, and interoperable data sharing platforms on the Internet to reach the concerned authorities. The GFP and IDMP are two of the typical showcases of such global partnerships.

References

- L. Alfieri, P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, F. Pappenberger, GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.* **17**, 1161–1175 (2013). <https://doi.org/10.5194/hess-17-1161-2013>
- G. Balsamo, A. Beljaars, K. Scipal, P. Viterbo, B. van den Hurk, M. Hirschi, A.K. Betts, A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the integrated forecast system. *J. Hydrometeorol.* **10**, 623–643 (2009)

- BoM, Australian Government Bureau of Meteorology, National Flood forecasting and warning services (2015). Available online at <http://www.bom.gov.au/water/floods/index.shtml>. Last accessed 29 Sept 2015
- G.R. Brakenridge, Global Active Archive of Large Flood Events, Dartmouth Flood Observatory (University of Colorado, 2015). Available online at <http://floodobservatory.colorado.edu/Archives/index.html>. Last accessed 13 July 2015
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009)
- CRED, Centre for Research on the Epidemiology of Disasters Emergency Database (2015). Available online at <http://www.emdat.be/>. Last accessed 14 Sept 2015
- T. De Groot, K. Poljansek, L. Vernaccini, Index for Risk Management - INFORM: Concept and Methodology. Luxembourg: Publications Office of the European Union. <https://doi.org/10.2788/78658> (2014a)
- T. De Groot, J. Thielen, R. Brakenridge, R. Adler, L. Alfieri, D. Kull, F. Lindsay, O. Imperiali, F. Pappenberger, R. Rudari, P. Salamon, N. Villars, K. Wyjad, Joining forces in a global flood partnership. *Bull. Am. Meteorol. Soc.* (2014b). <https://doi.org/10.1175/BAMS-D-14-00147.1>
- J. Dixon, A. Gulliver, D. Gibbon, *Farming Systems and Poverty: Improving Farmers' Livelihoods in a Changing World* (FAO & World Bank, Rome/Washington, DC, 2001)
- E. Dutra, F. Di Giuseppe, F. Wetterhall, F. Pappenberger, Seasonal forecasts of drought indices in African basins. *Hydrol. Earth Syst. Sci. Discuss.* **9**, 11093–11129 (2012). <https://doi.org/10.5194/hessd-9-11093-2012>
- EC, Environment Canada, Flood Forecasting Centers across Canada (2015). Available online at www.ec.gc.ca/eau-water/default.asp?lang=En&n=7BF9B012-1. Last accessed 29 Sept 2015
- T.O.S. Ewolabi, C.O. Ekechi, Communication as critical factor in disaster management and sustainable development in Nigeria. *Int. J. Dev. Econ. Sustain.* **2**(3), 58–72 (2014)
- FAO, Review of the public irrigation sector in Nigeria, Food and Agricultural Organization of the United Nations. Report No: 0009/TF/NIR/CPA/27277-2002/TCOT (2004). Available online at <ftp://ftp.fao.org/AGL/AGLW/ROPISIN/ROPISINReport.pdf>. Last accessed 31 Mar 2015
- A.K. Gerlak, J. Lautze, M. Giordano, Water resources data and information exchange in transboundary water treaties. *Int. Environ. Agreements* **11**, 179–199 (2011). <https://doi.org/10.1007/s10784-010-9144-4>
- GFP, Global Flood Partnership, Draft paper (2013). Available online at <http://portal.gdacs.org/Portals/0/GFP/Concept%20Paper%20Global%20Flood%20Partnership%20v4.3.pdf>. Last accessed 1 Sept 2015
- M. Haile, Weather patterns, food security and humanitarian response in sub-Saharan Africa. *Philos. Trans. R. Soc. B* **360**(1463), 2169–2182 (2005). <https://doi.org/10.1098/rstb.2005.1746>
- F.A. Hirpa, P. Salamon, L. Alfieri, J. Thielen, E. Zsoter, F. Pappenberger, The effect of reference climatology on global flood forecasting. *J. Hydrometeorol.* **17**(4), 1131–1145 (2016). <https://doi.org/10.1175/JHM-D-15-0044.1>
- IDPM, Integrated Drought Management Programme (2015). Available online at <http://www.droughtmanagement.info/idmp-activities/>. Last accessed 18 May 2015
- Z.W. Kundzewicz, Floods: lessons about early warning systems, in *Late Lessons from Early Warnings: Science, Precaution, Innovation 347 Emerging Lessons from Ecosystems* (European Environment Agency, 2013) Luxembourg: Publications Office of the European Union. <https://doi.org/10.2800/73322>
- G. Naumann, E. Dutra, P. Barbosa, F. Pappenberger, F. Wetterhall, J.V. Vogt, Comparison of drought indicators derived from multiple data sets over Africa. *Hydrol. Earth Syst. Sci.* **18**, 1625–1640 (2014). <https://doi.org/10.5194/hess-18-1625-2014>
- NiMet, Seasonal Rainfall Prediction for 2014 provided by the NiMet (2014). Available online at <http://nimet.gov.ng/sites/default/files/publications/SRP%20BROCHURE%20FINAL.pdf>. Last accessed 6 Oct 2015
- A. Nishat, I. Faisal, An assessment of the institutional mechanisms for water negotiations in the Ganges–Brahmaputra–Meghna system. *Int. Negot.* **5**(2), 289–310 (2000)

- NWS, United States National Weather Service, Meteorological Model Ensemble River Forecast (2015). Available online at <http://www.erh.noaa.gov/mmeefs/>. Last accessed 29 Sept 2015
- L. Ogallo, P. Bessemoulin, J.P. Ceron, S. Mason, S.J. Connor, Adapting to climate variability and change: the climate outlook forum process. *WMO Bull.* **57**, 93–102 (2008)
- F. Pappenberger, L. Stephens, S.J. van Andel, J.S. Verkade, M.H. Ramos, L. Alfieri, J. Brown, M. Zappa, G. Ricciardi, A. Wood, T. Pagano, R. Marty, W. Collischonn, M. Le Lay, D. Brochero, M. Cranston, D. Meissner, HEPEX—Operational HEPS systems around the globe (2015). Available online at <http://hepex.irstea.fr/operational-heps-systems-around-the-globe>. Last accessed 3 Oct 2015
- A.G. Patt, L. Ogallo, M. Hellmuth, Learning from 10 years of climate outlook forums in Africa. *Science* **318**(5847), 49–50 (2007). <https://doi.org/10.1126/science.1147909>
- PDNA, Nigeria post-disaster needs assessment 2012 floods (2013). Available online at https://www.gfdrr.org/sites/gfdrr/files/NIGERIA_PDNA_PRINT_05_29_2013_WEB.pdf. Last accessed 6 Oct 2015
- P.J. Pilon, Guidelines for reducing flood losses. United Nations Office for Disaster Risk Reduction (UNISDR) (2002). Available online at http://www.unisdr.org/files/558_7639.pdf. Last accessed 10 Mar 2015
- W. Pozzi, J. Sheffield, R. Stefanski, D. Cripe, R. Pulwarty, J.V. Vogt, R.R. Heim Jr., M.J. Brewer, M. Svoboda, R. Westerhoff, A.I.J.M. van Dijk, B. Lloyd-Hughes, F. Pappenberger, M. Werner, E. Dutra, F. Wetterhall, W. Wagner, S. Schubert, K. Mo, M. Nicholson, L. Bettio, L. Nunez, R. van Beek, M. Bierkens, L.G.G. de Goncalves, J.G.Z. de Mattos, R. Lawford, Toward global drought early warning capability: expanding international cooperation for the development of a framework for monitoring and forecasting. *Bull. Am. Meteorol. Soc.* **94**, 776–785 (2013). <https://doi.org/10.1175/BAMS-D-11-00176.1>
- G. Sepulcre-Canto, S. Horion, A. Singleton, H. Carrao, J. Vogt, Development of a combined drought indicator to detect drought in Europe. *Nat. Hazards Earth Syst. Sci.* **12**, 3519–3531 (2012)
- J. Sheffield, E.F. Wood, *Drought: Past Problems and Future Scenarios* (Earthscan, Routledge, London, 2011), p. 192
- J. Sheffield, E.F. Wood, N. Chaney, K. Guan, S. Sadri, X. Yuan, L. Olang, A. Amani, A. Ali, S. Demuth, A drought monitoring and forecasting system for sub-Saharan African water resources and food security. *Bull. Am. Meteorol. Soc.* **95**, 861–882 (2013). <https://doi.org/10.1175/BAMS-D-12-00124.1>
- J. Thielen, J. Bartholmes, M.-H. Ramos, A. de Roo, The European flood alert system—part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125–140 (2009). <https://doi.org/10.5194/hess-13-125-2009>
- A. Turton, P. Ashton, E. Cloete, An introduction to the hydropolitical drivers in the Okavango River basin, in *Transboundary Rivers, Sovereignty and Development: Hydropolitical Drivers in the Okavango River Basin*, ed. by A. Turton et al. (African Water Issue Research Unit/Green Cross, Pretoria, 2003), pp. 7–30
- UNISDR, Third United Nations World Conference on Disaster Risk Reduction, Geneva, 17–18 November 2014 (2014). Available at <http://www.wcdrr.org/uploads/Zero-draft-post2015-framework-for-DRR-20-October-.pdf>. Last accessed 10 Aug 2015
- UNISDR, Sendai Framework for Disaster Risk Reduction 2015–2030. United Nations Office for Disaster Risk Reduction (UNISDR), 32p (2015)
- J.M. van der Knijff, J. Younis, A.P.J. de Roo, LISFLOOD: a GIS – based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.* **24**(2), 189–212 (2010)
- J.C. Veilleux, The human security dimensions of dam development: the Grand Ethiopian Renaissance Dam. *Glob. Dialogue* **15**(2), 41 (2013)
- P.J. Ward, B. Jongman, F.S. Weiland, A. Bouwman, R. van Beek, M.F.P. Bierkens, W. Lictvoet, H.C. Winsemius, Assessing flood risk at the global scale: model setup, results, and sensitivity. *Environ. Res. Lett.* **8**(4), 044019 (2013)

- H.C. Winsemius, L.P.H. Van Beek, B. Jongman, P.J. Ward, A. Bouwman, A framework for global river flood risk assessments. *Hydrol. Earth Syst. Sci.* **17**(5), 1871–1892 (2013)
- WMO, World Meteorological Organization, Composition of the WMO (2014). Available at <http://www.wmo.int/wmocomposition/documents/wmocomposition.pdf>. Last accessed 19 Oct 2015
- WMO/GFP, Integrated Drought Management Programme, A joint WMO-GFP Programme, Concept Note, Version 1.2 (2011)



Communicating and Using Ensemble Flood Forecasts in Flood Incident Management: Lessons from Social Science

David Demeritt, Elisabeth M. Stephens, Laurence Crétton-Cazanave,
Céline Lutloff, Isabelle Ruin, and Sébastien Nobert

Contents

1	Introduction	1132
2	Visualizing HEPS Forecasts	1133
3	Perception and Understanding of Probabilistic Forecast Information	1137
4	Case Studies of HEPS in Operational Use	1141
4.1	France	1142

D. Demeritt (✉)

Department of Geography, King's College London, Strand, London, UK

e-mail: david.demeritt@kcl.ac.uk

E. M. Stephens

School of Archaeology, Geography and Environmental Science, University of Reading,
Whiteknights, Reading, UK

e-mail: elisabeth.stephens@reading.ac.uk

L. Crétton-Cazanave

Université Paris Est Marne-la-Vallée, Labex Futurs Urbains (LATTES, LEESU, Lab'Urba),
Marne-la-Vallée, France

e-mail: lcrettoncazanave@gmail.com

C. Lutloff

Université Grenoble 1, PACTE UMR 5194 (CNRS, IEPG, UJF, UPMF), Grenoble, France
e-mail: celine.lutloff@ujf-grenoble.fr

I. Ruin

Laboratoire d'étude des Transferts en Hydrologie et Environnement (LTHE), CNRS, Grenoble,
France

e-mail: isabelle.ruin@ujf-grenoble.fr

S. Nobert

School of Earth and Environment, University of Leeds, Leeds, UK

e-mail: s.nobert@leeds.ac.uk

4.2 United Kingdom	1145
4.3 European Flood Awareness System (EFAS)	1149
4.4 The Global Flood Awareness System (GloFAS)	1152
5 Summary and Conclusions	1155
References	1156

Abstract

This chapter explores the practical challenges of communicating and using ensemble forecasts in operational flood incident management. It reviews recent social science research on the variety and effectiveness of hydrological ensemble prediction systems (HEPS) visualization methods and on the cognitive and other challenges experienced by forecast recipients in understanding probabilistic forecasts correctly. To explore how those generic findings from the research literature work out in actual operational practice, the chapter then discusses a series of case studies detailing the development, communication, and use of HEPS products in various institutional contexts in France, Britain, and internationally at the EU and global levels. The chapter concludes by drawing out some broader lessons from those experiences about how to communicate and use HEPS more effectively.

Keywords

Risk communication · Risk perception · Cognitive biases · Heuristics · Visualization · Spaghetti plots · HEPS, public saliency of · Decision-making · Probability of precipitation forecasts · Uncertainty, public understanding of · Civil protection · EFAS · GloFAS · SCHAPI · UK Met Office · Environment Agency

1 Introduction

Recent years have seen rapid advances in flood forecasting. With developments in supercomputing, data assimilation, and modeling, real-time ensemble flood forecasting is increasingly now an operational reality for forecasting agencies. However, the technical capacity of fully coupled hydrological ensemble prediction systems (HEPS) to extend forecast lead times and provide quantitative information about the uncertainties associated with those forecasts has often galloped ahead of the institutional capacity to communicate and use those forecasts effectively. For all its promises, HEPS will be of little practical benefit in reducing the toll from flood disasters if the information they generate cannot be successfully communicated, interpreted, and used by those immediately at risk and the emergency services charged with protecting them.

Accordingly, this chapter explores the practical challenges of communicating and using ensemble forecasts in operational flood incident management. It summarizes recent social science research on the variety and effectiveness of HEPS visualization methods and on the cognitive and other difficulties experienced by forecast

recipients in understanding probabilistic forecasts correctly. To explore how those generic findings from the research literature shape operational practice, the chapter then draws on a series of case studies from the published literature about the development, communication, and use of HEPS products in various institutional contexts. The chapter concludes by drawing out some broader lessons from those experiences about how to communicate and use HEPS more effectively.

2 Visualizing HEPS Forecasts

Risk communication research highlights the ways in which the design and framing of uncertainty information can shape its perception (Visschers et al. 2009; Spiegelhalter et al. 2011). HEPS is a new technology, and there is, as yet, no expert consensus on the best methods for visualizing its forecasts, or even, as Demeritt et al. (2010) and Ramos et al. (2010) document, on what is the most important information to extract and transmit from a probabilistic flood forecast. However, it is increasingly recognized that there can be no “one-size-fits-all” solution: different users have different decision-making needs and will require different information visualized in different ways to meet those needs (Pappenberger et al. 2013).

For the most part, operational HEPS have tended to rely on fairly conventional visualization techniques (Lumbroso et al. 2009; Cloke and Pappenberger 2009; Bruen et al. 2010). So-called spaghetti-style hydrographs representing the temporal evolution of different ensemble members are quite common, as they can illuminate the degree of dispersion among ensemble members and thus provide an at-a-glance assessment of forecast uncertainty (Fig. 1). Particularly for less expert users, these may be displayed in a more simplified form. Sometimes they are converted to percentiles and displayed as a plume chart (Fig. 2) or box plot, or alternatively presented in reduced form as the mean of the ensemble members and their maximum and minimum values (Fig. 3). Another method for summarizing the ensemble – and one that does not require postprocessing – is in tabular form summarizing the number of ensemble members for a given location exceeding various thresholds over time (Fig. 4). This enables the timing and persistence over time of any signal to be assessed (Pappenberger et al. 2011). However, neither hydrographs nor these tabular displays provide any spatial information about the location of the risk and its spatial patterning, which can be important for some users.

To address that gap, HEPS forecasts in Sweden are presented on the public website in a choropleth map in which the shading of 1001 sub-basins in a national overview map represents the probability of exceeding the very lowest warning threshold for that sub-basin (Fig. 5). This very high-level summary gives a good overview but does not provide much local detail. The European Flood Alert System (EFAS) has developed a hybrid visualization that combines threshold exceedance maps, in which map pixels are color coded to represent the number of ensemble members at each point exceeding various thresholds, with additional clickable tabular summaries for any pixel (Fig. 6) as well as, for some selected gauging stations where data has been made available, postprocessed hydrographs of the

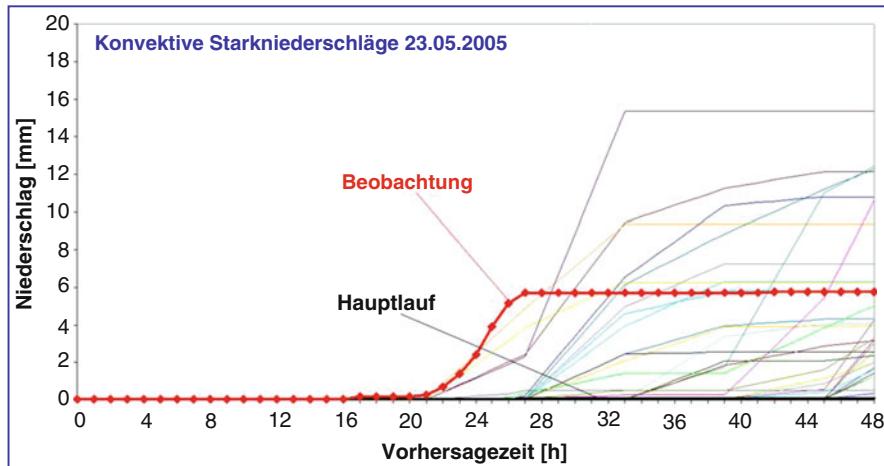


Fig. 1 In Austria, emergency services personnel working in the Abteilung Feuerwehr und Zivilschutz Landeswarnzentrale (the Fire Service and Civil Defence Early Warning Centre) have additional access to much richer EP outputs, including “spaghetti” plots of the 51-member ALADIN-LAEF of convective rainfall, which are the light colored lines in this plot which also shows the deterministic forecast (*Hauptlauf*) in black and the observed in red

ensemble (Fig. 2). Rather than just conventional hydrographs, Zappa et al. (2013) recommend the benefits of what they call a peak-box approach to draw attention to the temporal spread and magnitude of peak flood flow forecasts in different ensemble members. As with many such best practice recommendations for communicating HEPS (Martini and De Roo 2007), it is not backed by any empirical testing or systematic evaluation of its communicative effectiveness, which is now recognized as a research priority (Spiegelhalter et al. 2011; Demeritt and Nobert 2014).

One of the few studies to empirically evaluate HEPS visualization was conducted by Pappenberger et al. (2013). Based on focus group discussions with operational forecasters and other experts, they found strong support for the idea that HEPS needed to provide expert users with information about the following:

- Discharge
- Lead time in the form of a fully written date and time
- Warning/alert levels and/or return period
- Observations and past model performance
- Representation of uncertainty, either in percentiles or more simplified quantiles
- Worst/best case scenarios
- Metadata about station location and model, forecast provider, and whom to contact for further information
- A risk measure (expected cost, population affected, etc.)

However, it was also recognized that displaying such a rich array of information may be more appropriate for hydrological experts than for many other HEPS forecast

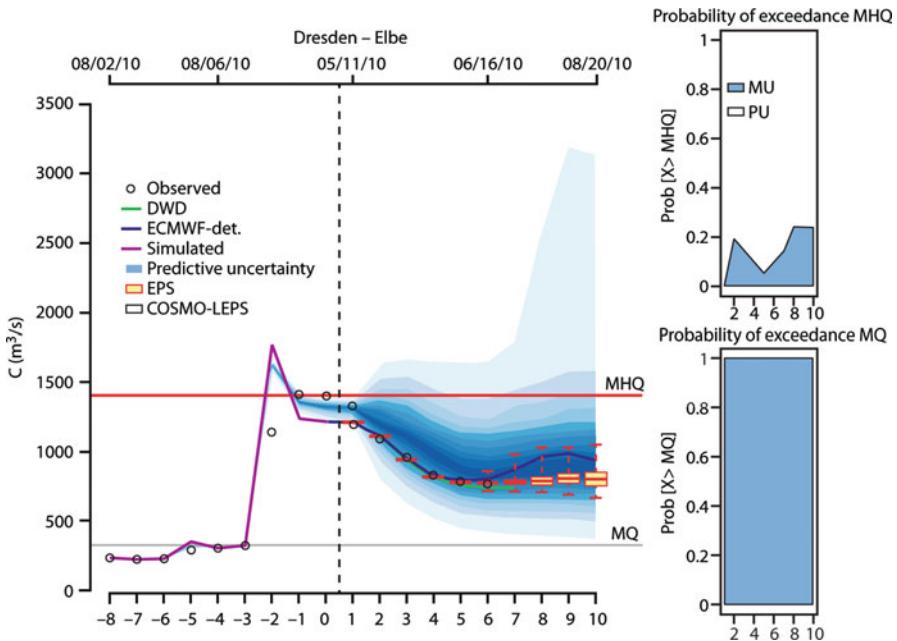


Fig. 2 Fan chart display of HEPS forecast from the European Flood Awareness System. The x-axis displays the time in days. The y-axis displays the flow discharge. MHQ stands for mean high discharge (taken from observed). MQ stands for mean flow (from observed). The plot shows the simulated forecast (*pink line*) and the observations (*open circles*). It also displays the three main meteorological forecasting systems driving EFAS (DWD, ECMWF, and COSMO-LEPS). It gives a full account of all uncertainties (model and predictive). Probability of threshold exceedances is shown in the *right-hand panel*. A detailed explanation can be found in the work of Bogner and Pappenberger (2011). Courtesy of EFAS, Joint Research Centre, European Commission, Ispra, Italy

recipients lacking hydrological expertise, such as members of the emergency services, elected officials, or members of the general public. Since user needs are likely to differ by customer type, level of expertise, and decision-making context, Pappenberger et al. (2013) cautioned that firm conclusions must await further research focused specifically on evaluating the effectiveness of different HEPS visualizations at meeting those needs.

Nevertheless, research on other domains of ensemble forecast communication does suggest some potential lessons for the communication of HEPS. A report from the US National Research Council (NRC 2006) provides an extensive review of the literature on communicating the uncertainty of meteorological forecasts. Among other things, it highlights the need for communication to be tailored to user needs. For instance, in the USA, the National Hurricane Center has been using “cone of probability” visualizations to depict the uncertainty about forecasted storm tracks (Fig. 7). However, Broad et al. (2007) found that the meaning of these visualizations was not always understood correctly by the general public because of ambiguity about what the cone represents: is it uncertainty around the track of the storm or the

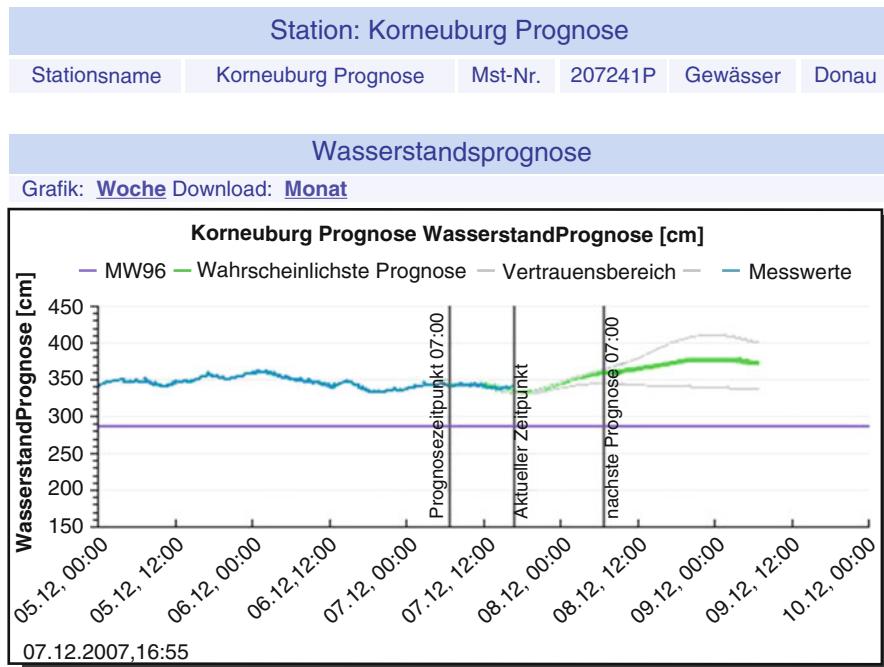


Fig. 3 In Austria, the public has access to simplified HEPS forecasts of streamflow, with the *blue line* showing observation, the *green line* a “best-guess” forecast, and the two *gray lines* the 10 and 90% confidence intervals

area that will be impacted by the storm, which is very much broader than the path of the storm eye itself. Research on the cone of uncertainty points to the importance of being explicit about what, exactly, is being represented in HEPS visualizations and what other features are being abstracted away. In this case the graphic clearly shows the uncertainty about the storm track, but at the cost of omitting other information that is arguably more important for public safety, such as the uncertainties about maximum wind speeds or about the scale and extent of secondary flooding, which the general public often fails to appreciate is in fact the leading cause of hurricane fatalities in the USA (Morss and Hayden 2010).

Distilling lessons for the communication of climate ensembles, Stephens et al. (2012) conclude that when designing a means of communication attention needs to be paid to balancing three priorities (Fig. 8). The first is robustness: making sure that the visualization reflects the scientific confidence in the prediction. Second, richness is the amount of detail, data dimensionality, and contextualizing information that are represented. Finally, saliency is the degree to which the information is relevant to the decision needs of the user. These priorities are often in tension with one another. As Stephens et al. (2012) explain “Some users may demand increases in informational richness (e.g., a full probability distribution rather than a range) that impact the ability of others to understand or use the information. Likewise concerns

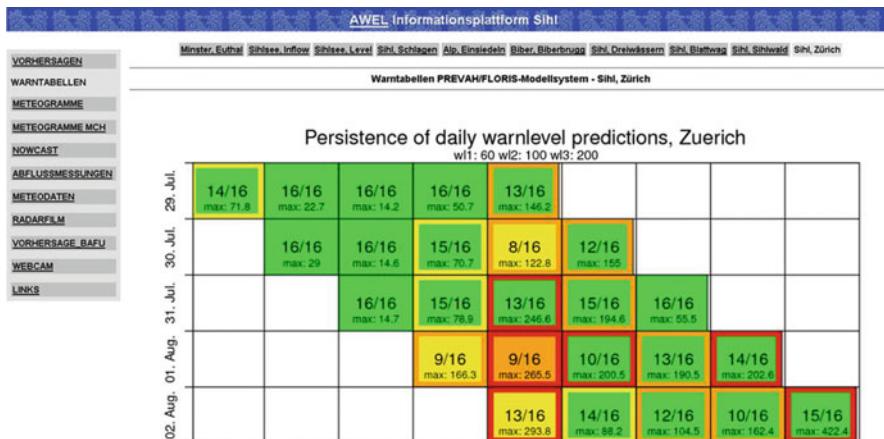


Fig. 4 Persistence in the number of COSMO-LEPS driven HEPS ensemble members for River Sihl, Zurich, Switzerland. The columns represent the days being forecasted and the rows the 5 day forecasts generated on a given date. The numbers inside each cell give the maximum forecasted flow and the number of ensemble member at the threshold level represented by the color of the cell, while the color of the cell border shows the highest warning level exceeded by any single ensemble member. Further details on this HEPS product are provided by Bruen et al. (2010)

with robustness (e.g., limitations and ambiguities of the EP) might require reduced informational richness, given that highly contested or incomplete predictions should not be communicated with unwarranted precision. Such alterations in richness, in turn, also affect perceptions of saliency, potentially decreasing it for users who want access to particular predictions, or increasing it for those who prefer simple, unambiguous results.” Beyond highlighting the general point that the design of HEPS products needs to be sensitive to the needs of the particular audiences for which they are intended, the framework developed by Stephens et al. (2012) provides a way for articulating the trade-offs and compromises involved in selecting among alternative HEPS visualizations.

3 Perception and Understanding of Probabilistic Forecast Information

While some, very simplified HEPS forecast products are now being disseminated directly to the public in a few countries, including Austria (Fig. 3), and Sweden (Fig. 5), with France, Canada, and several others planning to do so as well, there has been no research as yet about whether and how they are understood by public audiences.

There is, however, a rich body of research on public understandings of other forms of forecast and uncertainty information, and it provides a sound basis for inferring how the public is likely to perceive HEPS forecasts. Psychological research

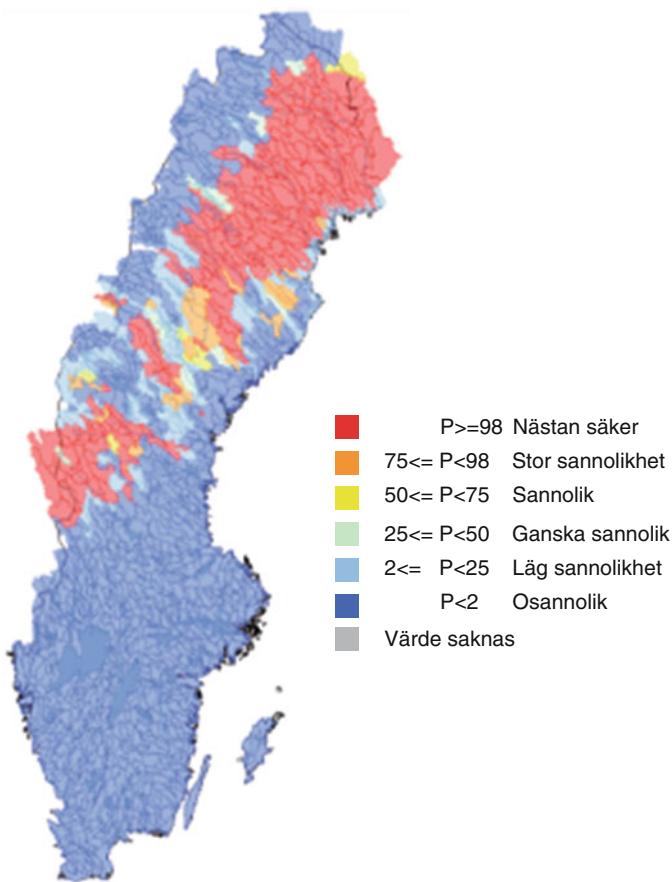


Fig. 5 Choropleth map of the probability of water levels in the 1001 sub-basins in Sweden exceeding the very lowest warning threshold for that sub-basin. Further details on this HEPS product are provided by Bruen et al. (2010)

has shown that people often resort to mental shortcuts, or heuristics, that bias their perceptions of and responses to uncertainty (Tversky and Kahneman 1974). For instance, people tend to anchor their judgments around whatever information is immediately available to them and then when seeking new information to heed any new information that reinforces those initial prejudices and discounts anything that contradicts them. A recent study attributed the higher death toll from feminine-named hurricanes to gender stereotypes that lead people to take fewer precautions for feminine storms on the false assumption that they are weaker than masculine ones (Jung et al. 2014). Although the validity of that study is contested (GrrlScientist and O'Hara 2014), it does suggest how cultural biases can shape the perception of forecast information. Cognitive biases are increasingly well recognized in the meteorological community (Nicholls 1999; Doswell 2004; NRC 2006), and they

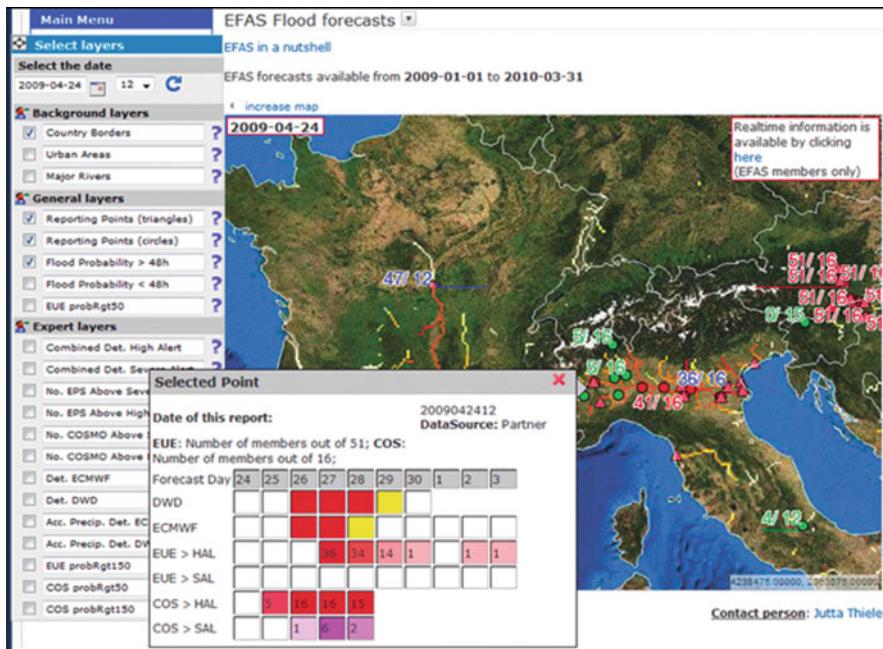


Fig. 6 Hybrid visualization of EFAS forecast for a selected point at 12:00 h UTC on 24 April 2009 when an external EFAS alert was issued warning of a high probability of flooding for the River Po from 27 April onwards with the peak wave between the 28 and 29 April

suggest reasons for caution about the oft-repeated assertion that probabilistic weather forecasts will inevitably lead to improved decision-making (Krzysztofowicz 2001; Palmer 2002).

Evidence for the value added by probabilistic forecasts is somewhat mixed. Several experimental studies with students have found that providing probabilistic information improves decision-making (Joslyn and Nichols 2009; Roulston and Kaplan 2009), though there is also evidence about the influence of message framing and formatting on the perception of and response to forecast uncertainty (Joslyn et al. 2009). In contrast to experimental studies, survey research has found widespread public misunderstanding of the meaning of probability of precipitation (PoP) forecasts (Murphy et al. 1980; Gigerenzer et al. 2005; Handmer and Proudley 2007), which are the most commonly provided probabilistic forecast product. In particular many people fail to understand the reference class of PoP forecasts, i.e., thinking that a 70% chance of rainfall means that it will rain for 70% of the time or over 70% of an area. This finding suggests the importance for robust decision-making of communicating the reference class for probability forecasts. However, despite the public's failure to grasp the precise technical meaning of PoP forecasts, Morss et al. (2008) argue that their understanding is still sufficiently good to meet everyday decision-making needs, based on a large-scale survey ($n = 1520$) of the general public in the USA.

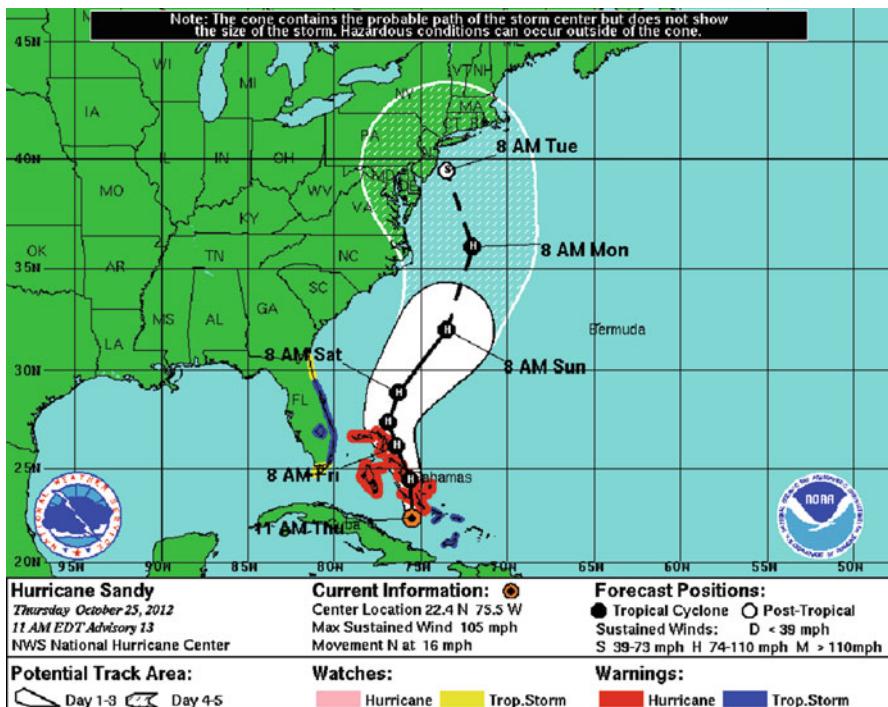
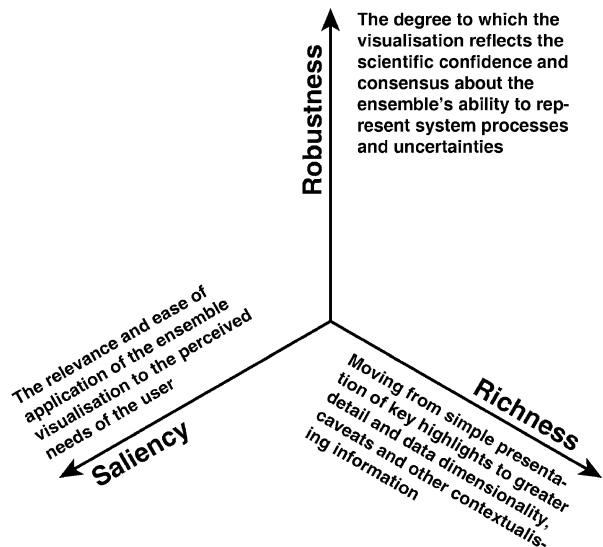


Fig. 7 Cone of uncertainty visualization issued by US National Hurricane Center at 11:00 EST on 25 October 2012. Further detail about the definition of the “cone” is published online: <http://www.nhc.noaa.gov/aboutcone.shtml>

There is also considerable debate over whether the uncertainty in the forecast should be communicated as a probability (i.e., 70%) or as a natural frequency (i.e., 7 in 10) (Spiegelhalter et al. 2011). In the field of medical risk communication, Gigerenzer (2003) has shown that medical professionals and lay publics alike understand risk better when it is communicated in terms of natural frequencies rather than probabilities, though Joslyn and Nichols (2009) found the opposite to be true in the case of weather forecasts. Similar questions arise around how best to communicate the concept of flood return periods. Research shows that people respond differently depending on whether flood risk is communicated as a return period or a probability (Bell and Tobin 2007; Highfield et al. 2013).

While social science research has tended to focus on whether the general public can understand probabilistic information correctly, research suggests that expert users are no less prone to cognitive biases in their perception and understanding of probabilities (Demeritt et al. 2007). Morss (2009) details a series of case studies from the USA in which expert failures to appreciate forecast uncertainties correctly led to disaster. Demeritt et al. (2010) found widespread unfamiliarity with EPS and quantitative probability forecasts, even among professional forecasters, and conflicting views among experts about their informational value as a measure of total forecast uncertainty.

Fig. 8 The three imperatives for visualization after Stephens et al. (2012): richness (amount of information communicated), robustness (the fidelity of the EP and the degree to which this is communicated), and saliency (interpretability and usefulness of the communication to a particular user). These may be viewed as a three-dimensional space in which the location of any given communication method depends on both design choices made and the limitations of the underlying EP



Other studies have looked at the ability of emergency services personnel and other more expert recipients of probabilistic flood forecasts to understand and use them for decision-making. In a simulation exercise in the Thames Estuary in England, McCarthy et al. (2007) found that flood managers often struggled to understand probabilistic flood forecasts correctly without support from forecast providers. Although Demeritt et al. (2010) documented widespread skepticism among operational flood forecasters about the ability of emergency services personnel to understand probabilistic forecasts or cope with forecast uncertainty more generally, interviews with civil protection officials suggest that in fact they have a greater appetite for and ability to receive and intelligently use uncertainty information than forecasters sometimes imagine. Similarly Kox et al. (2014) found Germany emergency responders to be generally confident about their ability to deal with probabilistic forecasts. Nobert et al. (2010) argue that the understanding and use of probabilistic flood forecast products could be improved through training and collaboration between forecast providers and users in the design and dissemination of HEPS to insure that they are salient for and meaningful to their intended users. Those conclusions are partly borne out by Frick and Hegg's (2011) study of the effectiveness of D-PHASE uncertainty visualizations in communicating forecast uncertainty to civil protection officials in Switzerland (Fig. 4).

4 Case Studies of HEPS in Operational Use

To understand how these general findings from the social science literature might play out in operational practice, we turn now to a series of case studies exploring the development and practical use of HEPS in a variety of national and international settings.

4.1 France

Recent French experience highlights some of the institutional challenges involved in communicating and using HEPS for operational flood incident management. Over the last 15 years, forecasting and early warning systems in France have been extensively reformed in response to successive failures to predict and manage severe weather events (Créton-Cazanave 2010; Vinet et al. 2012; Dedieu 2013; Daupras et al. 2014). The former 52 warning services (Services d'Annonce des Crues, SAC) have been consolidated into what are now just 19 regional flood forecasting services (Services de Prévision des Crues, SPC), whose local flood monitoring, forecasting, and warning activities are now overseen by a new national agency (the Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondation, SCHAPI) with responsibility for coordinating those local river forecasts within a single national visualization and communication platform, the Vigicrue.

These institutional reforms have been accompanied by significant technical innovation, including the development, by MétéoFrance, of a HEPS providing 10-day streamflow forecasts throughout France (Thirel et al. 2010a, b). As well as these medium-term HEPS forecasts, SCHAPI, along with many but not all SPC flood forecasters, also have access to various other ensemble products, including probabilistic weather forecasts from MétéoFrance and from the European Centre for Medium Range Weather Forecasting (ECMWF) as well as HEPS forecasts from the European Flood Awareness System (EFAS). Some SPC offices with responsibility for transboundary watersheds also have access to outputs from German and Swiss forecasting agencies as well. Typically these externally provided HEPS products are consulted informally by on-duty forecasters to build their situational awareness and complement the various internally developed and locally calibrated deterministic models and real-time observations that provide the primary sources of information relied on by SPC and SCHAPI forecasters in fulfilling their statutory duty to issue river level forecasts over the next 24 h. Despite government efforts to consolidate and standardize their forecasting practices, the models and information available to and regularly used by the 19 SPC offices remain quite heterogeneous. This reflects both the different forecasting needs of France's variegated physical geography and a tradition of professional autonomy and expert discretion in France and in forecasting more generally (Fine 2010; Créton-Cazanave and Lutoff 2013).

In contrast to the heterogeneity of the information used to underpin flood forecasting in France, the form of the forecasts themselves is now nationally standardized. All flood forecasts in France are now issued through a single national platform, the Internet-based Vigicrue (Fig. 9). It uses a color-coded "traffic-light" system to communicate four levels of "vigilance" required over the next 24 h. Whereas transmitting a "green" or escalating to "yellow" is at the discretion of the local SPC, issuing an "amber" or a "red" is a collective decision negotiated between the SPC, with their local models and expertise, and the SCHAPI with access to longer-term HEPS products and overall responsibility for forecasting and warning nationwide. In that capacity, SCHAPI is particularly concerned with harmonizing the map: if one section of river is red, the downstream sections should not be green. However,

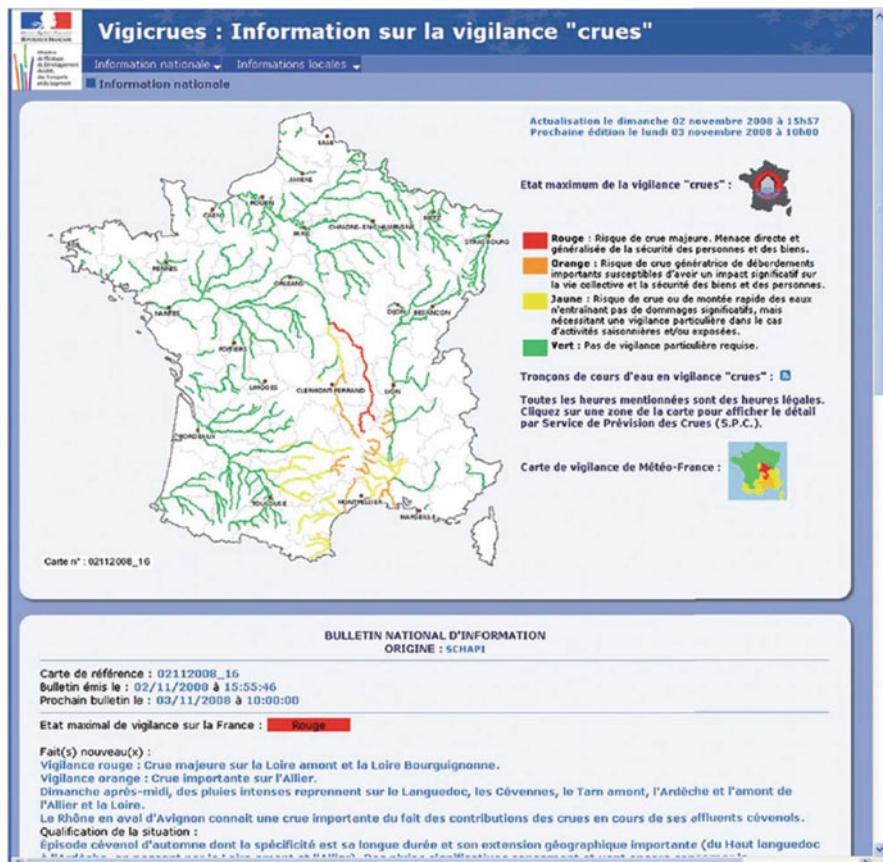


Fig. 9 Vigicrue flood risk mapping information as found on their website for the SPC *Littoral Atlantique* for the 3rd of May 2012. The four color-coded warning levels correspond to different levels of danger requiring increasing vigilance. *Red*: “Risk of a major flood, posing an imminent and widespread danger to life and property”; *Amber*: “Risk of substantial flooding with potential for significant impact on community life and the security of people and property”; *Yellow*: “Risk of flooding or of rapidly rising waters not causing significant damage but requiring some vigilance from those exposed”; *Green*: “No particular vigilance required”

given the fragmentation of the SPCs and the differences in tools, information, and expertise at their disposal, such inconsistencies are inevitable. The reforms help clarify who has the final say. In the end, once the color is chosen, SCHAPI updates the online map and the SPC writes the local bulletin, which also then becomes available to external users on the Vigicrue website.

These centralizing reforms have certainly helped clarify forecasting responsibilities, but in several other ways they have widened the gap between forecasting agencies and those responsible for civil protection. Under the old system (Houdré 2001), local forecasters in the 52 SACs were responsible, in the event of a flood, for

directly alerting the Préfet, the representative of the central state charged with overseeing the state's administrative services and ensuring public order in each of the 96 départements of metropolitan France (Cole 2011). The Préfet was then responsible for mobilizing the relevant state services and cascading the warning downwards to the elected mayors of any affected municipalities who were responsible for triggering their local emergency plans and for warning the population of any imminent danger. Both because of this regular personal contact and because with 52 SACs and various local subunits associated with them, forecasters were often located in close physical proximity to the officials they were responsible for alerting, it was possible to build close working relationships of trust (Daupras et al. 2014). However, without any central oversight or performance standards, this rather devolved system also meant that there were sometimes problems with the consistency, timeliness, and quality of the advice provided by the SACs (Huet et al. 2003). In response to these concerns, forecasting responsibilities were consolidated into a few bigger SPC offices (Vinet 2007). Consolidation facilitated higher levels of training and technical support, but it also resulted, during the transition period especially, in some loss of the tacit local knowledge so often said to be crucial to good flood forecasting (Blöschl 2008). Compared to the SACS, the SPCs were responsible for preparing forecasts for bigger territories with which they, necessarily, had less experience and tacit personal knowledge (Daupras et al. 2014).

Centralization of forecasting also altered the routines for forecast dissemination. Under the new system, forecasts are disseminated centrally through the new Vigicrue platform on the web. Mayors and civil protection agencies are now expected to consult the Internet without being prompted directly, as before, by the local flood forecasters. Research has shown that mayors and others with civil protection responsibilities often fail to do so, such that the forecasting-warning-emergency response chain can often break down at the very first link (Demeritt et al. 2013a; Daupras et al. 2014).

Other gaps arise because of mismatches between the technical design of the Vigicrue and the expectations of its users in civil protection. The thresholds used to trigger Vigicrue alert levels are not transparent, and so it is not clear whether an escalation from amber to red represents a judgment about the potential magnitude of flooding, an increase in its probability of occurring, both of which HEPS now makes it possible to quantify, the likely impacts of it doing so, or some composite sense of risk as the mathematical product of that probability and impact. Daupras et al. (2014) found that the meaning of these thresholds was often poorly understood by mayors and others responsible for civil protection. Furthermore Vinet (2007) has shown that Vigicrue thresholds are often poorly correlated to the local water levels used in local emergency plans (*plans communaux de sauvegarde*). Consequently they are sometimes disregarded by local civil protection authorities (Demeritt et al. 2013a; Daupras et al. 2014).

Scale is another issue. Particularly in the mountainous southeast of France, localized flash flooding is a major concern. In France, warnings of extreme rainfall likely to cause flooding are issued through MétéoFrance vigilance maps at the resolution of a French administrative department, which are typically several thousand km², while the Vigicrue system transmits separate flood warnings for

catchments larger than 100 km². Unlike the UK (see below), where the joint Flood Forecasting Centre was created precisely in order to ensure coordination between institutionally distinct rainfall forecasts from the Met Office and flood forecasts from the Environment Agency, there is no such mechanism in France for coordinating the meteorological vigilance alerts issued by MétéoFrance with the Vigicrue alerts put out by SCHAPI. Moreover, the scales for both these alerts are often too coarse to inform localized action in response to rising flood risk. Even at 24 h lead time, it is often difficult to forecast flash flooding with much confidence. As a result such forecasts are often discounted by local civil protection officials and members of the public alike (Créton-Cazanave 2009; Creutin et al. 2013).

Fluvial flooding on major rivers is more predictable than flash flooding, and here HEPS can provide substantially longer forecast lead times while also quantifying the uncertainty about them to enable more proportionate and cost-effective responses to risk. Indeed for this reason Électricité de France, the government-owned but autonomously managed electric utility company, has long used its own HEPS to help it manage water levels for nuclear plants and hydroelectric generation facilities more efficiently.

Despite that potential, there is limited appetite, either among flood forecasters or civil protection officials in France, for using HEPS in flood incident management. The rigidly hierarchical and top-down structure of crisis management in France, with strictly defined institutional responsibilities and individuals often held criminally liable for failures, deters organizations from acknowledging uncertainty externally. While forecasters are often keen to see HEPS themselves to increase their intelligence about the synoptic situation, they are reluctant to issue external warnings on a probabilistic basis or at low levels of confidence. Partly this reflects the empiricist “epistemic culture” of hydrological and flood control engineering, in which measured flows and historically calibrated return periods are preferred as the grounds for truth over mathematical formalism and physically based simulation modeling (Odoni and Lane 2010). However, it also reflects some wider institutional concerns about blame in the event of error, as well as tacit forecaster beliefs about the need for certainty in civil protection. For their part, civil protection organizations are reluctant to accept responsibility for dealing with forecast uncertainty and for managing the risk of error involved in acting on the basis of uncertain information from forecasters. They insist on accuracy above all else and blame forecasters for failing to provide it (Demeritt et al. 2013a). These institutional dynamics reinforce the tendency for public warnings to be issued at relatively short time horizons (typically <24 h) and coarse spatial scales, so as to reduce the chance of error, and hence blame. While SCHAPI is now contemplating the public dissemination of uncertainty information, more work will probably be required to develop the appetite for it amongst frontline SPC and civil protection staff.

4.2 United Kingdom

While the UK Met Office has been using its own MOGREPS ensemble for operational weather forecasting for a number of years, fluvial flood forecasting and

warning in Britain was traditionally based on various locally calibrated deterministic models run by the Environment Agency and its Welsh and Scottish equivalents. (There is no operational fluvial flood forecasting service for Northern Ireland, which relies instead on rainfall warnings provided by the Met Office.) However, in response to the Pitt (2008) Review of the 2007 floods, which criticized the lack of coordination between institutionally separate rainfall and flood forecasts, the government created a new joint Flood Forecasting Centre (FFC) (as well as a separate Scottish equivalent) to bridge the gap between these two types of forecasting and provide government and the emergency response community with an authoritative national overview of immediate flood risks (Haines and Stephens *in review*). To deliver the “step change in the quality of flood warnings” demanded by Pitt (2008) and “enable the most likely and the most extreme scenarios to be identified and shared with emergency responders to facilitate better preparedness” (para 4.10), FFC has developed and is now using its own HEPS as well as related ensemble storm surge products (Stephens and Cloke 2014).

Quite apart from the many technical difficulties involved in developing a fully coupled HEPS (Cloke and Pappenberger 2009), communicating and using the resulting quantitative probability forecasts has challenged existing institutional arrangements for managing flood risk in Britain. Traditionally, the Environment Agency was focused on short-term deterministic forecasting so as to meet a government performance target of providing people at risk of flooding with at least a two-hour advanced warning (Penning-Rowsell et al. 2000). While HEPS provides the technical capacity to extend that lead time, developing the institutional appetite and capacity to use it has proven more difficult.

To support the use of HEPS products internally by its own flood incident managers, the Environment Agency commissioned a decision-support framework to help operational staff use probabilistic forecasts to optimize the expected net benefits of their decisions given the known costs of different mitigation options and the uncertainty about the benefits (in the sense of flood losses avoided) that they would yield (Dale et al. 2014). Optimizing risk management through the use of such economic cost-loss functions is often touted as one of the principle benefits of HEPS (Palmer 2002). However, it can raise hackles by seeming to put a price on human life, and with the British Prime Minister David Cameron promising that “money would be no object” in responding to the winter 2014 floods (BBC 2014), senior management at the Environment Agency are understandably cautious about rolling out a tool that makes so explicit the cost-benefit trade-offs that are an inevitable part of planning for and responding to flood incidents.

There have also been difficulties in deciding whether and how to communicate probabilistic HEPS forecasts externally. At present HEPS products are not communicated directly to the public through the Environment Agency’s Flood Warning Service. This is partly because probabilistic forecasts are seen as too complicated to be understood by the lay public, though the Agency does now provide a three-day forward look on its public website that is based on HEPS. There are also concerns that providing public access to HEPS-based warnings issued at low thresholds of probability might reduce public responsiveness to the Flood Warning Service, for

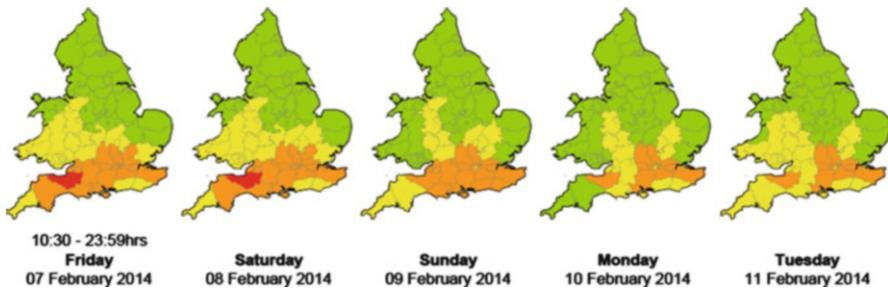


Fig. 10 Flood Guidance Statement overview map of flood risk from all sources for England and Wales issued by the joint Met Office/Environment Agency Flood Forecasting Centre on Friday 7 February 2014. Flood risk is quantified at county scale, while accompanying text and further detailed maps of areas of concern provide further information about the forecast

which warnings are triggered at much higher thresholds. Even the forecasters in the Agency's local area offices responsible for issuing those flood warnings to the public do not presently have direct access to the HEPS developed and used by FFC. Instead they still rely upon locally calibrated deterministic models to issue their warnings to the public and account for uncertainty about them by drawing on their local knowledge and experience to formulate "reasonable worst case scenarios" to run through their deterministic models rather than by using a HEPS to generate a more objective and quantitative assessment of the likelihood of such an event.

HEPS do, however, provide the basis for the FFC's Flood Guidance Statement, a 5 day forward look sent daily to so-called Cat 1 and 2 organizations in the emergency response community. In addition to detailed textual description of the synoptic situation and the expected impacts of any flooding, the FGS provides a cartographic overview (Fig. 10) in which English counties are color coded according to a risk matrix (Fig. 11) that takes account both of the probability of flooding and of its anticipated consequences. This "impacts-based" approach to hydrometeorological forecasting was introduced in March 2011. Both the FGS and the severe weather warnings issued by the Met Office use the same risk matrix. Prior to being issued formally, the precise position in the FGS risk matrix, along with the detailed wording about potential impacts, are discussed in a daily teleconference between the FFC, regional Environment Agency offices, and other key FGS customers in the emergency services to build consensus about what is, informally at least, something of a shared decision. This reliance on prerelease negotiation about the precise contents of the FGS is thus quite similar to the French practice in which SCHAPI and the SPC consult about the level at which the Vigicrue will be set.

The new impact-based format for forecasts and warnings is very popular with emergency responders, who praise both its simplicity and its benefits in reducing the number of inconsequential alerts. However, there is also evidence that the traffic light model of yellow-, amber-, and red-coded warnings is often misunderstood simply as a measure of probability rather than as a composite of probability *and* impact (Demeritt 2012). Indeed, it was this concern that led the Environment

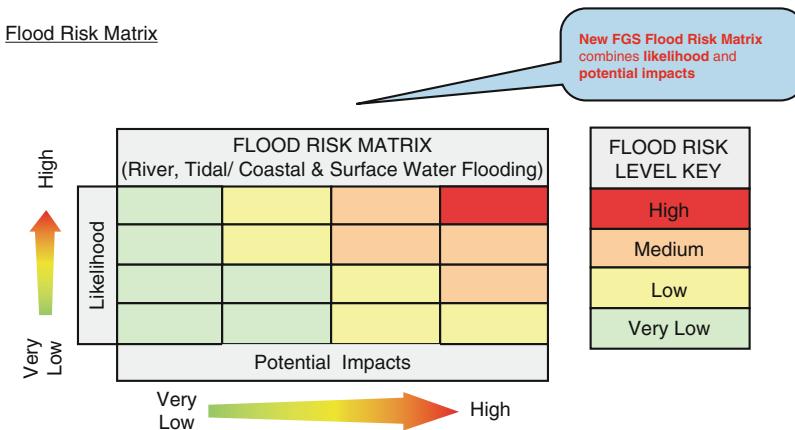


Fig. 11 Risk matrix used by the Flood Guidance Statement to specify the combination of estimated probability and potential impacts from flooding. This same risk matrix is now also used by the UK Met Office as part of its National Severe Weather Warning Service to communicate the likelihood and impact of severe weather events

Agency to move away from a color-coded traffic light model for its public Flood Warning Service after the Bye and Horner (1998) report into the Easter 1998 floods concluded “Colour coded warnings appear to be misunderstood by nearly all who receive them. This is because the colours are spontaneously linked with the escalating probability of flooding actually occurring.” While some emergency responders are quite cognizant of the different kinds of “amber” (i.e., “high” likelihood of “significant” impact vs. “low” likelihood of “severe” impacts), it is more typical for forecasters and recipients alike to speak of them as aligned along a single, somewhat inchoately defined dimension from a “yellow” to an “amber” to “red” without recognizing the different combinations of estimated probability and consequence that those colors can embody (Demeritt 2012).

Even if those warnings are correctly understood, there are also challenges in getting recipients to respond proactively to early warnings issued at low levels of confidence. Although the FFC performed well in the recent winter 2013–2014 floods and was widely praised for providing effective early warnings (Stephens and Cloke 2014), the response from government was often chaotic and was widely criticized for it (Demeritt 2014). Within the “blue light” emergency services, there is an embedded professional culture of responding to emergencies rather than preventing them. As such the tendency is sometimes to wait for confirmation rather than respond early in the face of uncertainty. This hesitancy is reinforced by various other institutional factors, including resource constraints and an ingrained culture of institutional risk aversion and blame avoidance in the public sector in Britain (Hood 2010; Huber and Rothstein 2013), which deters preemptive action in response to probabilistic forecasts unless issued at probabilities of at least 50% (Demeritt 2012).

4.3 European Flood Awareness System (EFAS)

EFAS was created in the immediate aftermath of the devastating European floods of 2002. As part of a communication about creating the Solidarity Fund to release structural development funds to finance disaster relief and reconstruction, the European Commission also pledged “to provide scientific support for a European flood warning system containing information on the main European basins and with real-time access to medium-term meteorological forecasts” (CEC 2002). Developed by the Commission’s Joint Research Centre (JRC) in Ispra, EFAS is intended to compliment national flood forecasting capacities and to support a European-level response to flood emergencies across the European continent. As of August 2014, more than 30 hydrological and civil protection services across Europe were signed up to receive EFAS alerts. In addition to receiving EFAS alerts about potential flooding occurring 3–10 days ahead on large ($>4000 \text{ km}^2$) catchments, they also have access to a password-protected website providing Pan-European overview maps (Fig. 6) of flood probabilities up to 10 days in advance as well as other background information and reports.

The design of EFAS alerts has evolved considerably over the years through regular interaction between EFAS developers at the JRC and their users (Demeritt et al. 2013b). The annual user meeting has proven to be an effective venue for delivering training and soliciting feedback about the design of EFAS alerts and other issues about their accuracy, dissemination, and operational use by recipients (De Roo et al. 2011). The initial format for EFAS alerts combined textual information about the synoptic situation with threshold exceedance maps (Fig. 12) in which the color coding for each pixel represented whether, with different input combinations, some threshold had been exceeded for that location. In response to user feedback, EFAS alerts were modified to incorporate additional tabular information summarizing the number of EFAS ensemble members for a given pixel exceeding various thresholds over time (Fig. 6).

EFAS users also wanted access to the raw hydrographs, and the difficulties involved in providing them nicely illustrate the institutional complexities of multilateral collaboration in HEPS development. Particularly in the early years, some national hydrological services were sensitive about EFAS infringing on national forecasting prerogatives and thus were reluctant to provide the EFAS team with data needed for model validation and error correction (Demeritt and Nobert 2011). Consequently, EFAS thresholds were derived from a model climatology and its forecasts often differed substantially from observed values (Thielen et al. 2009). This discrepancy is of only secondary importance to the EFAS mission of providing early warnings, which can be triggered by reference to a simulated reference period of model runs, but it was not one that the EFAS team was keen to highlight by disseminating uncorrected hydrographs, which might be misinterpreted. As well as protecting their own credibility, the EFAS team was also conscious that EFAS hydrographs might be seen as somehow competing with the local river level forecasts issued by EFAS partners in the member states (Demeritt et al. 2013b).

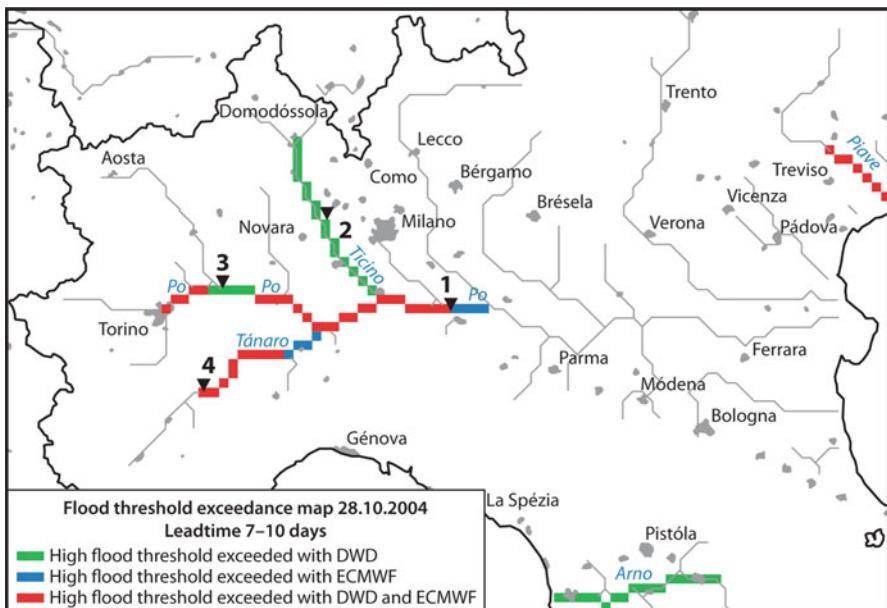


Fig. 12 Early EFAS prototype threshold exceedance map of the type included in the first preoperational EFAS alerts. It shows those areas that exceed high flood thresholds using Deutscher Wetterdienst (DWD) forecast only (green, lead time 7 days), ECMWF forecast only (blue, lead time 10 days), and both DWD and ECMWF (red). Triangles identify locations for which the temporal evolution of flooding over the forecasting range is given in tables provided elsewhere in the alert

These institutional considerations led EFAS to develop its innovative tabular display of the number of EFAS threshold exceedances (Fig. 13), which has since been copied by other forecasters, notably the visualization (Fig. 5) designed by the Swiss for the Map DPhase project (Frick and Hegg 2011). This tabular display provided a way to highlight the crucial information needed for EFAS early warnings without requiring extensive postprocessing to correct systematic biases in its local forecasts. Since it was not liable to being confused with a local forecast of river flows at a given point, the tabular display also reinforced the institutional distinction between the EFAS function of supporting national forecast centers with a complementary system of early warnings based on medium-term ensemble forecasts and those national centers, which are legally responsible for producing 0–48 h forecasts of river flows at given points. Having now won the confidence of national services through the sort of end user engagement recommended in the literature (Nobert et al. 2010), EFAS is currently collaborating with its partners to assimilate and bias correct real-time discharge data so as to be able to present its forecasts for some selected discharge station points as hydrographs incorporating local levels for certain return periods (Fig. 2).

Within national centers, EFAS alerts are used in several ways. Their primary role is in raising awareness of an emerging potential risk and prompting greater vigilance

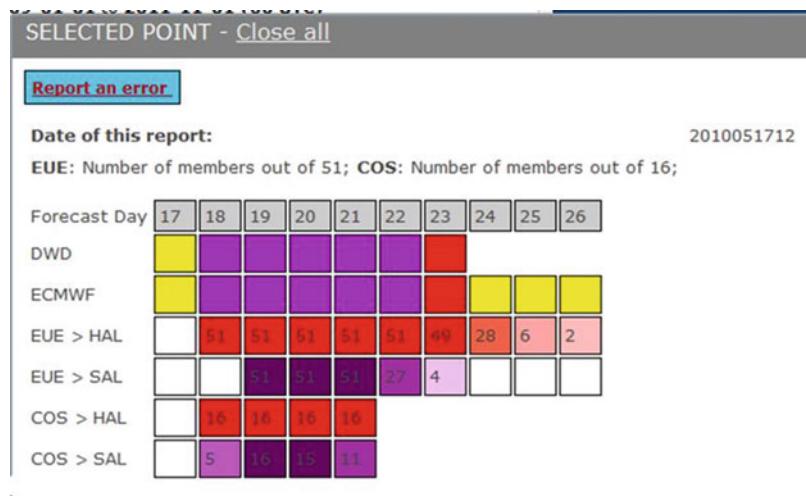


Fig. 13 Further tabular detail for a selected point in southern Poland from the EFAS forecast for 12:00 h UTC on 17 May 2010 when severe flooding resulted in 2.5 billion euros in damages. The first two rows classify the various EFAS river flow forecasts produced for that point using the deterministic rainfall forecasts from DWD (Deutscher Wetterdienst, the German national meteorological service) and ECMWF, with *purple* indicating flows in excess of the EFAS Severe Alert Level (SAL) corresponding to a simulated flood event with a return period of >20 years, *red* indicating flows in excess of the EFAS High Alert Level (HAL), and *yellow* in excess of the Medium Alert Level. The numbers in the subsequent rows indicate the number of EFAS ensemble members produced using the ECMWF ensemble (EUE) and the COSMO-LEPS limited area ensemble (COS) that exceed the HAL and Severe Alert Levels (SAL)

and closer surveillance of the evolving synoptic situation. EFAS alerts are sometimes also used to trigger increased preparedness within the forecasting center itself, for instance, by ensuring the availability of duty staff in the event an emergency is declared. In general, however, there is a reluctance to act on the basis of EFAS alerts alone. Instead, centers typically wait for local confirmation, either from their own in-house models or in other cases from direct empirical measures of rising flood waters, before triggering an alert (Demeritt et al. 2013b). Partly this reluctance to accept EFAS alerts reflected particular concerns about its skill and local relevance. As in the French case described above, forecasters in mountainous regions of Europe concerned primarily with flash flooding were often more skeptical about the value added by EFAS than those responsible for forecasting and warning on large rivers, where HEPS has greater skill (Demeritt et al. 2013a). Such complaints were not uncommon during the preoperational testing and development phase of EFAS, but with the ongoing work to postprocess and error correct EFAS forecasts, it is now possible to look at hydrographs presented alongside locally derived threshold information, which in turn should enhance users' ability to relate them to local sources of information (Blöschl 2008).

However, the tendency to wait for local confirmation before acting on medium-term EFAS alerts also reflects some deeper institutional considerations. As with

forecasters in Britain and France, the forecasters receiving EFAS alerts were often concerned about whether their users in civil protection would be able even to understand, let alone act on, a probabilistic forecast or a warning issued with less than total confidence. Particularly in countries with strongly legalistic traditions of expert management, forecasting centers often found it difficult to know what to do with the kind “prewarning” provided by EFAS (Demeritt and Nobert 2011; Demeritt et al. 2013b). Institutional arrangements for flood management in such countries are essentially deterministic and framed by an absolute distinction between safety and a state of emergency requiring the mobilization of civil protection and other extraordinary measures by the state. This is one reason why, historically, European flood forecasting agencies have tended to set quite high confidence thresholds for issuing flood warnings. Their focus has been on short-term (0–24 h) warnings to support public evacuations, rather than on medium-term forecasting in support of flood damage mitigation (Penning-Rowsell et al. 2000).

The hesitancy of national flood forecasting agencies to act on medium-term forecasts is magnified, in the case of EFAS alerts, because of their external provenance. If the EFAS alert proves to be wrong, it is the national agency that will be blamed. This was the cause of considerable anxiety about the very existence of EFAS and of a European-level flood forecasting capacity that might compete with national level competencies (Demeritt and Nobert 2011). Whereas the EFAS team has taken a strongly collaborative approach to engaging their users in the basic design of the EFAS system, the number and physical dispersion of EFAS partners, combined with the relative infrequency with which any one of them might receive an alert from EFAS, means that there are not the same close personal relationships and prerelease negotiation about issuing an EFAS alerts that there are in the French and British cases discussed above. Those interactions help to build user confidence and to close the “gap” between early warning and response identified in the literature (Créton-Cazanave 2010; Meyer et al. 2010).

4.4 The Global Flood Awareness System (GloFAS)

Based on the same basic modeling architecture as EFAS and building on many of the lessons about user engagement learned in its development (De Roo et al. 2011), the Global Flood Awareness System (GloFAS) is a global ensemble forecasting and warning system (Fig. 14). It currently runs daily in a preoperational experimental mode and provides global coverage, but at a 0.1° spatial resolution it is focused only on the largest river basins (Alfieri et al. 2013). Whereas EFAS was designed to complement national forecasting capabilities by providing probabilistic forecasts over longer time horizons, GloFAS is intended not only to enhance national forecasting capabilities but also to support humanitarian organizations who do not have their own hydrometeorological forecasting capabilities.

This diverse combination of users creates some unique challenges for system development. In keeping with the research evidence about the value of engaging with users “upstream” as part of the model development process (Wilsdon and Willis

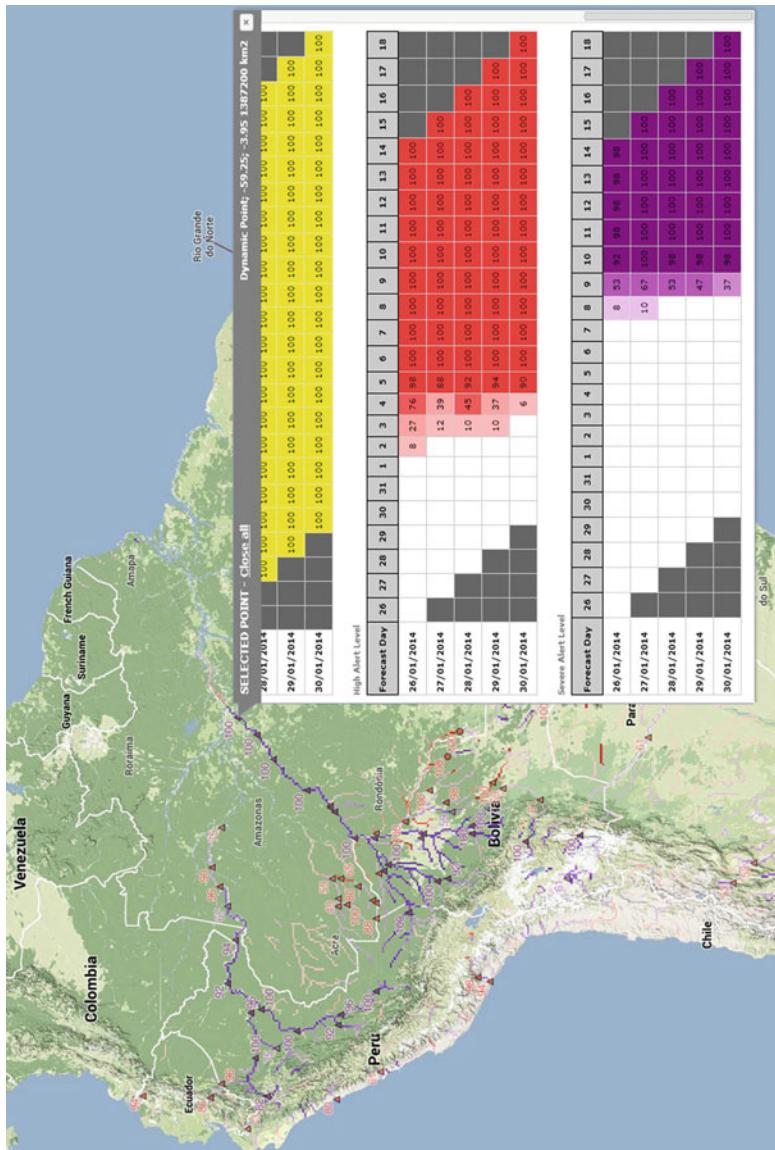


Fig. 14 Prototype visualization for GloFAS HEPS. Modeled on the hybrid EFAS visualization (see Fig. 6), this cartographic overview represents the probability of different GloFAS threshold exceedances using colored pixels, while further information for any given point is available in tabular form by clicking on the relevant pixel

2004; Nobert et al. 2010), GloFAS developers have looked to the lessons from EFAS about the importance of engaging with users early on in the process. With EFAS, the preoperational phase of development gave plentiful opportunity for potential users to provide feedback on the system and contribute substantively to its technical development (De Roo et al. 2011). EFAS users were all from national hydrological services and so had both technical expertise to contribute to the design process as well as clear and relatively consistent views about how EFAS could and should complement their existing national capabilities. By contrast GloFAS users are much more diverse, and the humanitarian agencies in particular lack the same “contributory expertise” (Collins and Evans 2002) as users from national forecasting centers. Bridging these gaps between science and operational need is a common challenge in the humanitarian community. In the climate services field, specialist agencies like the Red Cross Red Crescent Climate Centre try to provide science “brokerage” services to link weather and climate science more closely to the wider Red Cross and Red Crescent Movement. The Climate Center and the GloFAS team have invested time in developing the common language and “interactional expertise” necessary to deliberate about what humanitarian decisions GloFAS is – and isn’t – able to support, as well as working together to ensure that GloFAS is complementary to national capabilities.

GloFAS developers and humanitarian practitioners are working together to support the use of forecasts for humanitarian action. Anticipation of imminent disasters is an overlooked field, falling in the gap between the two distinct communities involved in long-term risk reduction and in the immediate emergency response to disasters that are already underway. For the latter community especially, there is little or no funding available to finance actions that might reduce the immediate risk in the run-up to a flood or other event. Additionally, despite the known benefits of acting before a disaster strikes – e.g., prepositioning aid by truck is much less costly than helicopter airdrops during a flood – donors are reluctant to release funds if there is a chance they might act in vain. Of course given the uncertainty inherent to forecasting it might never be possible to provide a 100% guaranteed forecast of flooding. There is recognition that somehow the barriers to using uncertain forecasts for anticipatory action need to be overcome, and so “forecast-based financing” – releasing funds or other assets when a forecast meets a certain threshold – is an emerging area for both the humanitarian response and disaster risk reduction communities (Coughlan de Perez et al. 2015). The Red Cross is currently using GloFAS within forecast-based financing pilot projects in Uganda and Peru. Currently the model output is being compared to the timing of known floods, from which standard operating procedures with predetermined probability and magnitude thresholds are being created to enable an automatic transfer of funds to enable the Ugandan Red Cross Society to respond in advance of a flood. This pilot scheme is one small example of how GloFAS is being applied to inform operational decision-making within the humanitarian community.

One of the main challenges faced by GloFAS in building more such partnerships arises from the heterogeneity of its users, which complicates efforts to identify and then deliver user-relevant thresholds on a global scale. On smaller scales and with

single users it is much easier to develop locally appropriate critical thresholds for flood impact (and therefore warning). For example, the Red Cross forecast-based financing pilot projects allow detailed assessment of appropriate thresholds from examination of the model, while in contexts where flood defenses are built to a particular standard, warning thresholds in the forecast could be set to that local standard. However, GloFAS serves a wide range of different users with very different needs and critical thresholds for operational decision-making. Even if appropriate thresholds could be identified, GloFAS faces other practical challenges. In data-sparse countries or in situations where flood exposure and vulnerability are dynamic (e.g., in times of conflict or crop failure), it may be extremely difficult to define locally appropriate thresholds.

As forecast-based financing is an emerging field it will be important for hydrologists to work together with the humanitarian response and disaster risk reduction communities to develop the decision-making processes to make best use of the technical potential of HEPS. For example, the creation and use of standard operating procedures for humanitarian action will require the forecast model to demonstrate statistical reliability (Coughlan de Perez et al. 2015). Such collaborative learning during the development stage of the system, as shown for EFAS, should serve to enhance the success of the partnership and promote the use of GloFAS for decision-making.

5 Summary and Conclusions

The technical promises of HEPS will be of little practical benefit unless its forecasts can be successfully communicated to and used by decision makers to improve flood incident management. This chapter considered a range of HEPS visualizations in current operational use. There is an urgent need for more research to evaluate their effectiveness. In its absence, the chapter reviewed recent social science research on the communication of other probabilistic forecast products to identify lessons for HEPS. That research highlights a series of cognitive biases that shape the perception of and response to risk information and that HEPS designers need to appreciate in order to ensure successful communication. One important lesson for HEPS emerging from recent risk communication research is the importance of being clear about the reference class for probabilistic forecasts and about what, exactly, is being represented and what is being abstracted away in particular HEPS visualizations.

Another key insight emerging from the literature is the importance of tailoring HEPS visualizations to suit the decision-making needs and expectations of their intended audiences. Those needs are likely to vary and so it follows that there can be no one “best way” to visualize HEPS forecasts. The appropriate balance between the competing demands for greater richness, robustness, and user saliency in HEPS visualization will vary from one user group to another. Insofar as HEPS designers are unlikely to know about the needs of their users *a priori*, there is a clear need for HEPS designers to engage early and often with their users so as to understand their decision-making needs and design their systems with those performance

requirements clearly in mind. The EFAS experience highlights the benefits of such engagement, even as it also illustrates how the ordinary institutional challenges of HEPS development, such as data access, standardization, and funding, can be magnified by the challenges of international partnership working and by the heterogeneity of users potentially served by GloFAS.

While effective communication is clearly necessary if HEPS is to achieve its potential for improving flood incident management, the case studies discussed in the chapter show that it is by no means sufficient to ensure that they are actually used operationally. There is work to be done to develop the institutional capacity to use HEPS intelligently. Proactive responses to the longer lead times generated by HEPS can be inhibited by various institutional factors, including lack of trust, resource constraints, institutional defensiveness, and absolutist legal duties of protection, which make it difficult to acknowledge uncertainty. Training and user engagement can go some way to overcoming these obstacles by helping to build confidence in HEPS and identify decision-making processes that might benefit from the sort of probabilistic information that it can provide. But it is important to recognize how intractable they can be and to be realistic about the institutional challenges involved in getting flood risk management regimes to acknowledge, and even embrace, uncertainty.

References

- L. Alfieri, P. Burek, E. Dutra et al., GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.* **17**, 1161–1175 (2013). <https://doi.org/10.5194/hess-17-1161-2013>
- BBC, Money “no object” for flood relief. BBC News (2014). <http://www.bbc.com/news/uk-26131515>. Accessed 1 Aug 2014
- H. Bell, G. Tobin, Efficient and effective? The 100-year flood in the communication and perception of flood risk. *Environ. Hazards* **7**, 302–311 (2007). <https://doi.org/10.1016/j.envhaz.2007.08.004>
- G. Blöschl, Flood warning – on the value of local information. *Int. J. River Basin Manage.* **6**, 41–50 (2008). <https://doi.org/10.1080/15715124.2008.9635336>
- K. Bogner, F. Pappenberger, Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water Resour. Res.* **47**, W07524 (2011). <https://doi.org/10.1029/2010WR009137>
- K. Broad, A. Leiserowitz, J. Weinkle, M. Steketee, Misinterpretations of the “Cone of Uncertainty” in Florida during the 2004 Hurricane season. *Bull. Am. Meteorol. Soc.* **88**, 651–667 (2007). <https://doi.org/10.1175/BAMS-88-5-651>
- M. Bruen, P. Krahe, M. Zappa et al., Visualizing flood forecasting uncertainty: some current European EPS platforms-COST731 working group 3. *Atmos. Sci. Lett.* **11**, 92–99 (2010). <https://doi.org/10.1002/asl.258>
- P. Bye, M. Horner, *Easter 1998 Floods: Report by the Independent Review Team to the Board of the Environment Agency* (Environment Agency, Bristol, 1998)
- CEC, *The European Community Response to the Flooding in Austria, Germany and Several Applicant Countries: A solidarity-Based Initiative* (Commission of the European Communities, Brussels, 2002)

- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**, 613–626 (2009). <https://doi.org/10.1016/j.jhydrol.2009.06.005>
- A. Cole, Prefects in search of a role in a Europeanised France. *J. Public Policy* **31**, 385–407 (2011). <https://doi.org/10.1017/S0143814X11000122>
- H.M. Collins, R. Evans, The third wave of science studies: studies of expertise and experience. *Soc. Stud. Sci.* **32**, 235–296 (2002). <https://doi.org/10.1177/0306312702032002003>
- E. Coughlan de Perez, B. van den Hurk, M.K. van Aalst et al., Forecast-based financing: an approach for catalyzing humanitarian action based on extreme weather and climate forecasts. *Nat. Hazards Earth Syst. Sci.* **15**, 895–904 (2015). <https://doi.org/10.5194/nhess-15-895-2015>
- L. Crétton-Cazanave, Warning! The use of meteorological information during a flash-flood warning process. *Adv. Sci. Res.* **3**, 99–103 (2009)
- L. Crétton-Cazanave, Penser l’alerte par les distances. Entre planification et émancipation, l’exemple du processus d’alerte aux crues rapides sur le bassin versant du Vidourle, Université de Grenoble, 2010
- L. Crétton-Cazanave, C. Lutoff, Stakeholders’ issues for action during the warning process and the interpretation of forecasts’ uncertainties. *Nat. Hazards Earth Syst. Sci.* **13**, 1469–1479 (2013). <https://doi.org/10.5194/nhess-13-1469-2013>
- J.D. Creutin, M. Borga, E. Gruntfest et al., A space and time framework for analyzing human anticipation of flash floods. *J. Hydrol.* **482**, 14–24 (2013). <https://doi.org/10.1016/j.jhydrol.2012.11.009>
- M. Dale, J. Wicks, K. Mylne et al., Probabilistic flood forecasting and decision-making: an innovative risk-based approach. *Nat. Hazards* **70**, 159–172 (2014). <https://doi.org/10.1007/s11069-012-0483-z>
- F. Daupras, J.M. Antoine, S. Becerra, A. Peltier, Analysis of the robustness of the French flood warning system: a study based on the 2009 flood of the Garonne River. *Nat. Hazards* 1–27 (2014). <https://doi.org/10.1007/s11069-014-1318-x>
- A. De Roo, J. Thielen, P. Salamon et al., Quality control, validation and user feedback of the European Flood Alert System (EFAS). *Int. J. Digital Earth* **4**, 77–90 (2011). <https://doi.org/10.1080/17538947.2010.510302>
- F. Dedieu, *Une catastrophe ordinaire la tempête du 27 décembre 1999* (Editions de l’Ecole des Hautes Etudes en Sciences Sociales, Paris, 2013)
- D. Demeritt, The perception and use of public weather services by emergency and resiliency professionals in the UK. Report for the Met Office Public Weather Service Customer Group. 2 Mar (King’s College London, London, 2012)
- D. Demeritt, Spooked politicians are undermining flood defence policy with short term decisions. *New Civil Engineer* 9 (2014)
- D. Demeritt, S. Nobert, Responding to early flood warning in the European Union, in *Forecasting, Warning, and Transnational Risks: Is Prevention Possible?* ed. by C.O. Meyer, C. de Franco (Palgrave Macmillan, Basingstoke, 2011), pp. 127–147
- D. Demeritt, S. Nobert, Models of best practice in flood risk communication and management. *Environ. Hazards* 1–16 (2014). <https://doi.org/10.1080/17477891.2014.924897>
- D. Demeritt, H. Cloke, F. Pappenberger et al., Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environ. Hazards* **7**, 115–127 (2007). <https://doi.org/10.1016/j.envhaz.2007.05.001>
- D. Demeritt, S. Nobert, H. Cloke, F. Pappenberger, Challenges in communicating and using ensembles in operational flood forecasting. *Meteorol. Appl.* **17**, 209–222 (2010). <https://doi.org/10.1002/met.194>
- D. Demeritt, S. Nobert, M. Bachecker et al., *Assessing Risk Communication Strategies and Effectiveness in Early Warnings* (UNESCO-IHE Institute for Water Education, Delft, 2013a)
- D. Demeritt, S. Nobert, H.L. Cloke, F. Pappenberger, The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol. Process.* **27**, 147–157 (2013b). <https://doi.org/10.1002/hyp.9419>

- C.A.I. Doswell, Weather forecasting by humans – heuristics and decision making. *Weather Forecast*. **19**, 1115–1126 (2004)
- G.A. Fine, *Authors of the Storm: Meteorologists and the Culture of Prediction* (University of Chicago Press, Chicago, 2010)
- J. Frick, C. Hegg, Can end-users' flood management decision making be improved by information about forecast uncertainty? *Atmos. Res.* **100**, 296–303 (2011). <https://doi.org/10.1016/j.atmosres.2010.12.006>
- G. Gigerenzer, *Reckoning with Risk: Learning to Live with Uncertainty* (Penguin Books, London, 2003)
- G. Gigerenzer, R. Hertwig, E. Van Den Broek et al., "A 30% chance of rain tomorrow": how does the public understand probabilistic weather forecasts? *Risk Anal.* **25**, 623–629 (2005). <https://doi.org/10.1111/j.1539-6924.2005.00608.x>
- GrrlScientist, O'Hara B, Is widespread sexism making hurricanes more deadly than himmicanes? (2014). http://www.theguardian.com/science/grrlscientist/2014/jun/04/hurricane-gender-name-bias-sexism-statistics?CMP=twt_fd. Accessed 5 Aug 2014
- S. Haines, E.M. Stephens, Partnerships in weather forecasting: development, distance, and dialogue (in review).
- J. Handmer, B. Proudley, Communicating uncertainty via probabilities: the case of weather forecasts. *Environ. Hazards* **7**, 79–87 (2007). <https://doi.org/10.1016/j.envhaz.2007.05.002>
- W.E. Highfield, S.A. Norman, S.D. Brody, Examining the 100-year floodplain as a metric of risk, loss, and household adjustment. *Risk Anal.* **33**, 186–191 (2013). <https://doi.org/10.1111/j.1539-6924.2012.01840.x>
- C. Hood, *The Blame Game: Spin, Bureaucracy, and Self-Preservation in Government* (Princeton University Press, Princeton, 2010)
- F. Houtré, *L'Annonce des crues: Histoire et évolution des services de 1847 à nos jours* (Ministère de l'Aménagement du Territoire et de l'Environnement, Paris, 2001)
- M. Huber, H. Rothstein, The risk organisation: or how organisations reconcile themselves to failure. *J. Risk Res.* **16**, 651–675 (2013). <https://doi.org/10.1080/13669877.2012.761276>
- P. Huet, P. Foin, C. Laurain, P. Cannard, *Retour d'expé'rience des crues de septembre 2002 dans les départements du Gard, de l'Hérault, du Vaucluse, des Bouches-du-Rhône, de l'Ardèche et de la Drôme: rapport consolidé après phase contradictoire* (Service de l'inspection générale de l'environnement, Paris, 2003)
- S.L. Joslyn, R.M. Nichols, Probability or frequency? Expressing forecast uncertainty in public weather forecasts. *Meteorol. Appl.* **16**, 309–314 (2009). <https://doi.org/10.1002/met.121>
- S.L. Joslyn, L. Nadav-Greenberg, M.U. Taing, R.M. Nichols, The effects of wording on the understanding and use of uncertainty information in a threshold forecasting decision. *Appl. Cogn. Psychol.* **23**, 55–72 (2009). <https://doi.org/10.1002/acp.1449>
- K. Jung, S. Shavitt, M. Viswanathan, J.M. Hilbe, Female hurricanes are deadlier than male hurricanes. *Proc. Natl. Acad. Sci. U. S. A.* 201402786 (2014). <https://doi.org/10.1073/pnas.1402786111>
- T. Kox, L. Gerhold, U. Ulbrich, Perception and use of uncertainty in severe weather warnings by emergency services in Germany. *Atmos. Res.* (2014). <https://doi.org/10.1016/j.atmosres.2014.02.024>
- R. Krzysztofowicz, The case for probabilistic forecasting in hydrology. *J. Hydrol.* **249**, 2–9 (2001). [https://doi.org/10.1016/S0022-1694\(01\)00420-6](https://doi.org/10.1016/S0022-1694(01)00420-6)
- D. Lumbroso, B. von Christierson, *Communication and Dissemination of Probabilistic Flood Warnings: Literature Review of International Material* (Environment Agency, Bristol, 2009)
- F. Martini, A. De Roo, *EXCIFF Guide: Good Practice for Delivering Flood-Related Information to the General Public* (Joint Research Centre, European Commission, Ispra, 2007)
- S. McCarthy, S. Tunstall, D. Parker et al., Risk communication in emergency response to a simulated extreme flood. *Environ. Hazards* **7**, 179–192 (2007). <https://doi.org/10.1016/j.envhaz.2007.06.003>

- C.O. Meyer, F. Otto, J. Brante, C. De Franco, Recasting the warning-response problem: persuasion and preventive policy. *Int. Stud. Rev.* **12**, 556–578 (2010). <https://doi.org/10.1111/j.1468-2486.2010.00960.x>
- R.E. Morss, Interactions among flood predictions, decisions, and outcomes: synthesis of three cases. *Nat. Hazard. Rev.* **11**, 83–96 (2009)
- R.E. Morss, M.H. Hayden, Storm surge and “certain death”: interviews with Texas coastal residents following Hurricane Ike. *Wea. Climate Soc.* **2**, 174–189 (2010). <https://doi.org/10.1175/2010WCAS1041.1>
- R.E. Morss, J.L. Demuth, J.K. Lazo, Communicating uncertainty in weather forecasts: a survey of the U.S. public. *Weather Forecast.* **23**, 974–991 (2008). <https://doi.org/10.1175/2008WAF2007088.1>
- A.H. Murphy, S. Lichtenstein, B. Fischhoff, R.L. Winkler, Misinterpretations of precipitation probability forecasts. *Bull. Am. Meteorol. Soc.* **61**, 695–701 (1980). [https://doi.org/10.1175/1520-0477\(1980\)061<0695:MOPPF>2.0.CO;2](https://doi.org/10.1175/1520-0477(1980)061<0695:MOPPF>2.0.CO;2)
- N. Nicholls, Cognitive illusions, heuristics, and climate prediction. *Bull. Am. Meteorol. Soc.* **80**, 1385–1398 (1999)
- S. Nobert, D. Demeritt, H. Cloke, Informing operational flood management with ensemble predictions: lessons from Sweden. *J. Flood Risk Manage.* **3**, 72–79 (2010). <https://doi.org/10.1111/j.1753-318X.2009.01056.x>
- NRC [National Research Council], *Completing the Forecast: Characterizing and Communicating Uncertainty for Better Decisions Using Weather and Climate Forecasts* (National Academies Press, Washington, DC, 2006)
- N.A. Odoni, S.N. Lane, Knowledge-theoretic models in hydrology. *Prog. Phys. Geogr.* **34**, 151–171 (2010). <https://doi.org/10.1177/0309133309359893>
- T.N. Palmer, The economic value of ensemble forecasts as a tool for risk assessment: from days to decades. *Q. J. Roy. Meteorol. Soc.* **128**, 747–774 (2002)
- F. Pappenberger, H.L. Cloke, A. Persson, D. Demeritt, HESS Opinions “On forecast (in)consistency in a hydro-meteorological chain: curse or blessing?”. *Hydrol. Earth Syst. Sci.* **15**, 2391–2400 (2011). <https://doi.org/10.5194/hess-15-2391-2011>
- F. Pappenberger, E. Stephens, J. Thielen et al., Visualizing probabilistic flood forecast information: expert preferences and perceptions of best practice in uncertainty communication. *Hydrol. Process.* **27**, 132–146 (2013). <https://doi.org/10.1002/hyp.9253>
- E.C. Penning-Rowsell, S.M. Tunstall, S.M. Tapsell, D.J. Parker, The benefits of flood warnings: real but elusive, and politically significant. *Water Environ. J.* **14**, 7–14 (2000)
- M. Pitt, *The Pitt Review: Learning Lessons from the 2007 Floods* (Cabinet Office, London, 2008)
- M.-H. Ramos, T. Mathevot, J. Thielen, F. Pappenberger, Communicating uncertainty in hydro-meteorological forecasts: mission impossible? *Meteorol. Appl.* **17**, 223–235 (2010). <https://doi.org/10.1002/met.202>
- M.S. Roulston, T.R. Kaplan, A laboratory-based study of understanding of uncertainty in 5-day site-specific temperature forecasts. *Meteorol. Appl.* **16**, 237–244 (2009). <https://doi.org/10.1002/met.113>
- D. Spiegelhalter, M. Pearson, I. Short, Visualizing uncertainty about the future. *Science* **333**, 1393–1400 (2011). <https://doi.org/10.1126/science.1191181>
- E. Stephens, H. Cloke, Improving flood forecasts for better flood preparedness in the UK (and beyond). *Geogr. J.* **4**, 310–316 (2014). <https://doi.org/10.1111/geoj.12103>
- E.M. Stephens, T.L. Edwards, D. Demeritt, Communicating probabilistic information from climate model ensembles—lessons from numerical weather prediction. *Wiley Interdiscip. Rev. Clim. Chang.* **3**, 409–426 (2012). <https://doi.org/10.1002/wcc.187>
- J. Thielen, J. Bartholmes, M.H. Ramos, A. De Roo, The European Flood Alert System – part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125 (2009)
- G. Thirel, E. Martin, J.-F. Mahfouf et al., A past discharge assimilation system for ensemble streamflow forecasts over France – part 2: impact on the ensemble streamflow forecasts. *Hydrol. Earth Syst. Sci.* **14**, 1639–1653 (2010a). <https://doi.org/10.5194/hess-14-1639-2010>

- G. Thirel, E. Martin, J.-F. Mahfouf et al., A past discharges assimilation system for ensemble streamflow forecasts over France – part 1: description and validation of the assimilation system. *Hydrol. Earth Syst. Sci.* **14**, 1623–1637 (2010b). <https://doi.org/10.5194/hess-14-1623-2010>
- A. Tversky, D. Kahneman, Judgment under uncertainty: heuristics and biases. *Science* **185**, 1124–1131 (1974). <https://doi.org/10.1126/science.185.4157.1124>
- F. Vinet, Approches nationales de la prévention des risques et besoins locaux : le cas de la prévision et de l'alerte aux crues dans le Midi méditerranéen. *Géocarrefour* **82**, 35–42 (2007). <https://doi.org/10.4000/geocarrefour.1438>
- F. Vinet, D. Lumbroso, S. Defossez, L. Boissier, A comparative analysis of the loss of life during two recent floods in France: the sea surge caused by the storm Xynthia and the flash flood in Var. *Nat. Hazards* **61**, 1179–1201 (2012). <https://doi.org/10.1007/s11069-011-9975-5>
- V.H.M. Visschers, R.M. Meertens, W.W.F. Passchier, N.N.K. De Vries, Probability information in risk communication: a review of the research literature. *Risk Anal.* **29**, 267–287 (2009). <https://doi.org/10.1111/j.1539-6924.2008.01137.x>
- J. Wilsdon, R. Willis, *See-Through Science: Why Public Engagement Needs to Move Upstream* (Demos, London, 2004)
- M. Zappa, F. Fundel, S. Jaun, A “Peak-Box” approach for supporting interpretation and verification of operational ensemble peak-flow forecasts. *Hydrol. Process.* **27**, 117–131 (2013). <https://doi.org/10.1002/hyp.9521>



Overview of Communication Strategies for Uncertainty in Hydrological Forecasting in Australia

With a Focus on “Assessing Forecast Quality of the
National Seasonal Streamflow Forecast Service”

Narendra Kumar Tuteja, Senlin Zhou, Julien Lerat, Q. J. Wang,
Daehyok Shin, and David E. Robertson

Contents

1	Introduction	1162
2	Methodology Used to Assess Forecast Quality of the National Seasonal Streamflow Forecast Service	1163
2.1	Modeling Approach	1163
2.2	Seasonal Streamflow Forecast Verification	1165
2.3	Aggregated Forecast Performance	1169
3	Results and Discussion	1170
3.1	Forecast Reliability	1170
3.2	Forecast Accuracy	1172
3.3	Aggregated Forecast Performance	1174
4	Conclusion	1175
	References	1178

Abstract

The National Seasonal Streamflow Forecasting Service operated by the Bureau of Meteorology since 2010 delivers monthly updates of 3 month ensemble forecasts at 147 locations across 75 river basins using the statistical Bayesian joint probability (BJP). Seasonal forecasts are communicated to the public using statistical concepts such as “chances,” “ensembles,” “lower/higher than median,” etc. However, these concepts require advanced competencies in statistics, and they

N. K. Tuteja (✉) · J. Lerat
Bureau of Meteorology, Canberra, ACT, Australia
e-mail: n.tuteja@bom.gov.au

S. Zhou · D. Shin
Bureau of Meteorology, Docklands, VIC, Australia

Q. J. Wang · D. E. Robertson
CSIRO Land and Water, Clayton, VIC, Australia

cannot be conveyed to a general audience easily. This chapter focuses on the challenge of communicating forecast skill to a wide range of users more effectively. A simple forecast performance measure called the “Aggregated Forecast Performance Index (AFPI)” was introduced which captures key attributes such as forecast reliability and accuracy and combines them into a single easy-to-understand and well-informed aggregated measure. Based on this index, it was demonstrated that bureau’s seasonal streamflow forecasts are reliable. They also offer improved accuracy by narrowing down the forecast uncertainty (up to 25%) with respect to reference climatology and hence offer a value proposition for water managers to improve their decision-making.

Keywords

Seasonal streamflow forecasting · Ensemble forecasting · Uncertainty estimation · Forecast verification · Aggregated forecast performance · Forecast accuracy · Forecast precision · Forecast reliability · Continuous Rank Probability Score (CRPS) · Root mean squared error (RMSE) · Root mean squared error in probability space (RMSEP) · Hit rates · Bayesian joint probability model (BJP)

1 Introduction

Hydrological conditions in Australia are among the most variable on earth (McMahon et al. 1987). Its streamflow regime can go through prolonged periods of droughts such as the “Millennium drought” that occurred between 1995 and 2007 across most parts of eastern Australia (Chiew et al. 2008). This variability has a profound impact on the management of water resources in Australia and more specifically on the management of risks related to water supply for urban, irrigation, and environmental needs.

On 26 January 2007, following a prolonged period of severe drought and rapidly diminishing water supplies, the Australian Prime Minister announced the National Plan for Water Security, a 10-point plan significantly enhancing Commonwealth involvement in the nation’s water affairs (Vertessy 2013). One of the pillars of the reforms was a significant commitment to improving the quality and coverage of Australia’s water information, which includes a National Seasonal Streamflow Forecasting Service (SSF) provided by the Bureau of Meteorology since December 2010 (www.bom.gov.au/water/ssf). The SSF service now provides forecasts at 147 locations across 75 river basins located across Australia as of July 2016. Work is currently underway to include more locations not covered to-date.

Seasonal forecasting is probabilistic by nature due to large uncertainties in the dynamics of the atmosphere and uncertainties in the response of catchments to climate forcing. As a result, seasonal forecasts are communicated to the public using statistical concepts such as “chances,” “ensembles,” “lower/higher than median,” etc. However, these concepts require advanced competencies in

statistics and they cannot be conveyed to a general audience easily (Morss et al. 2008).

More precisely, seasonal forecasting involves two main challenges. The first is to issue probabilistic forecasts using communication methods that aim to *minimize the risks of misinterpretation*. An overview of the products provided by the Bureau of Meteorology to address this challenge is discussed here, and more details can be found in a recent stakeholder survey indicating a high level of satisfaction related to the SSF products (Wilson et al. 2014).

Here we focus on the second challenge in the seasonal forecasting domain, i.e., *communicating forecast skill*. Forecasting models have varying levels of skill depending on the forecast location and period of the year. Measures of skill can have a strong influence on how forecasts impact decisions related to water management, and they must be communicated to the users of the forecasts. Various forecast verification methods are available for assessing the multiple facets of forecast performance including notions such as accuracy and reliability (Murphy 1993). However, these methods remain fairly complex and target a scientific audience.

This paper describes a rigorous approach used by the Bureau of Meteorology to verify streamflow forecasts using a variety of complementary performance metrics which are then integrated into a single index. This index can be easily communicated to and discussed with a wide audience as described below.

2 Methodology Used to Assess Forecast Quality of the National Seasonal Streamflow Forecast Service

2.1 Modeling Approach

Probabilistic streamflow forecasts of 3 month ahead outlooks for this service are derived from a statistical modeling approach called the *Bayesian Joint Probability* model (BJP; Wang et al. 2009; Wang and Robertson 2011; Robertson and Wang 2012) using a modeling system called the *Water Forecasting System for Australian Rivers* (WAFARi; Shin et al. 2011). Currently, the service updates streamflow forecasts every month for 147 key water supply catchments in eastern Australia, with the objective of helping water suppliers and users make better decisions. In the period 2014–15, the service will be expanded across all jurisdictions in Australia (Fig. 1).

The BJP model uses statistical relationships between climate indices, recent catchment conditions, and historical rainfall and streamflow at a site to forecast streamflow for the next 3 months. The BJP approach assumes that a transformed set of streamflows and their predictors follow a multivariate normal distribution. A typical model comprises a predictor to represent antecedent catchment conditions, commonly recently observed streamflow, one climate index predictor, and a streamflow predictand (see Wang and Robertson (2011) for details). Such a model would require up to 13 parameters. Model parameters and their uncertainty are inferred using a Bayesian formulation, which is implemented through a Markov Chain Monte Carlo (MCMC) sampling method. Forecasts are made by conditioning

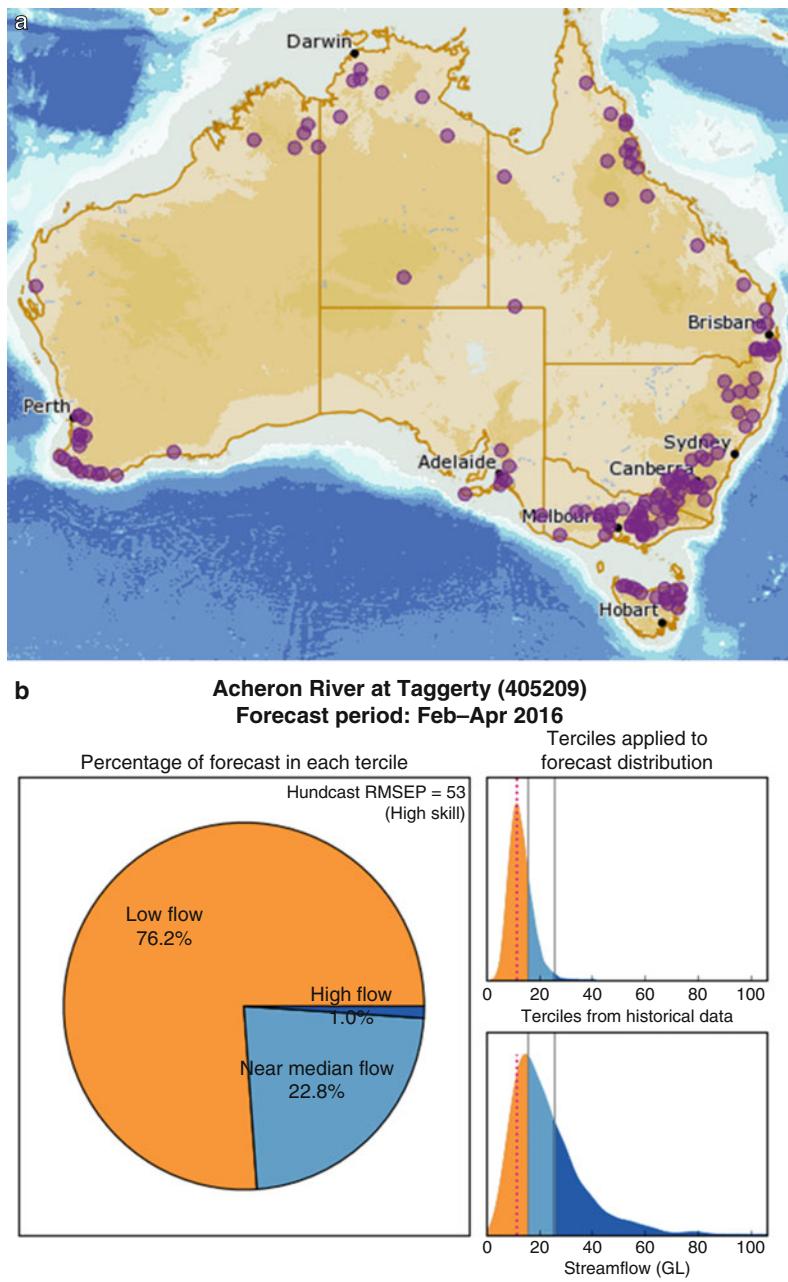


Fig. 1 (a) Spatial extent of the seasonal streamflow forecast locations across Australia (146 locations as of July 2016). (b) Forecast distribution derived from the Bayesian Joint Probability model, historical reference distribution (1950–2010) and tercile forecasts derived from the historical reference distribution for the Acheron River at Taggerty, Goulburn River Basin

the multivariate distribution on predictor values represented by a 5000-member ensemble.

The BJP model for forecasting 3 month streamflow volumes is calibrated and verified separately for each 3-month period (e.g., JFM, FMA... DJF). The cross-validation process, i.e., model calibration and forecast verification in retrospective mode at a given site and season, is illustrated in Fig. 2. The current cross-validation procedure is developed for the period 1980–2008 of available streamflow records in which 5 years are left out of the model calibration and are used for forecast verification.

2.2 Seasonal Streamflow Forecast Verification

The need to establish requirements for a comprehensive national system to verify hydrologic forecasts and guidance products which satisfy end user requirements is unequivocal. Key elements of the Bureau's approach to national seasonal streamflow forecast verification are:

- Scientific rigor in the development of stringent forecast verification metrics encompassing a multitude of methods using existing and/or new and evolving techniques that can be used to represent probabilistic forecast performance across all sites and seasons. Central to this requirement is the objective to publicly display a range of forecast performance metrics at a granular level in a transparent manner for the end users to have an informed view of the forecasting system performance prior to its use in decision making.
- The requirement is to simplify the communication of ensemble streamflow forecast performance for practitioners across the water industry, to support

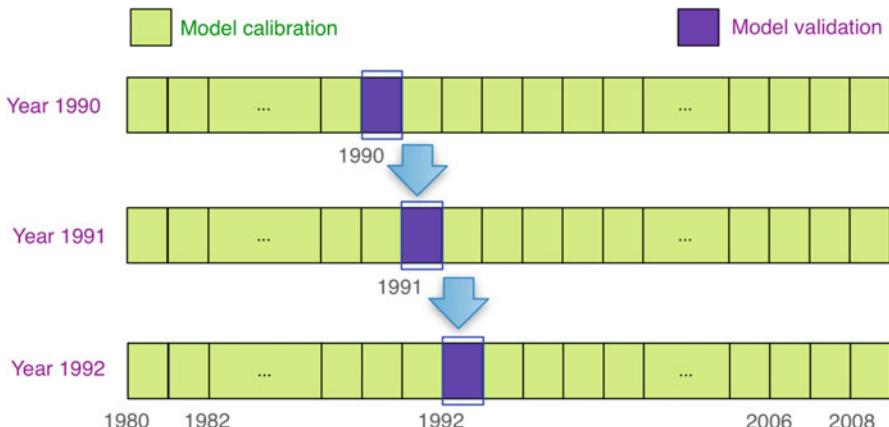


Fig. 2 Illustration of the BJP model calibration and validation procedure implemented in the WAFARI modeling system for each forecast month and location (146 locations as of July 2016). The calibration scheme is repeated for each year of the streamflow records

informed decision making against the background of imperfect forecasts within an ensemble-based risk assessment framework. While this latter issue is complex, it is nevertheless critical for the adoption of the seasonal streamflow forecasts to directly support operational planning and decision making in reservoir operations, water allocation planning and delivery for consumptive use, and environmental flow management.

Streamflow forecast performance could be classified in two broad categories: (a) reliability and (b) accuracy. Both of these categories are important for the end user. The Bureau of Meteorology in collaboration with its research partners from CSIRO and university sector is developing methods to describe streamflow forecast performance designed to address end user needs across the water industry in Australia. It is largely work in progress, and these methods will continue to evolve into the future as end user needs are better understood over the coming years.

2.2.1 Forecast Reliability

Forecast reliability is related to the correspondence between the distribution of forecasts and the distribution of observations. It is critically important so that water managers can confidently eliminate least plausible options in their water allocation and delivery planning which invariably involves extensive scenario analysis. The streamflow forecast distribution is conditioned on the assumptions made during the inference stage. Unsupported assumptions may lead to inadequate and unreliable forecast distributions which in turn impacts on end user confidence in the forecasts.

Forecast reliability is assessed using the Probability Integral Transform (PIT) plot (Dawid 1984; De Gooijer and Zerom 2000; Gneiting et al. 2007; Laio and Tamea 2007; Thyer et al. 2009; Wang et al. 2009). The diagram plots the cumulative distribution of the PIT computed from observations transformed using the corresponding forecast *CDF* against a standard uniform variate. The proximity of the plotted curve to the diagonal line in the plot indicates the level of reliability of the forecasts (Fig. 3a).

2.2.2 Forecast Accuracy

Forecast accuracy is usually defined in a relative way as “the relative accuracy of a set of forecasts, with respect to some set of standard control or reference forecasts” (Wilks 1995 pp. 236–237). In many cases, the distribution of data from historical records has been used as the reference forecast. The generic form of a skill score is represented through Eq. 1,

$$\text{Skill Score} = \frac{\text{Score}_{\text{fcast}} - \text{Score}_{\text{ref}}}{\text{Score}_{\text{perf}} - \text{Score}_{\text{ref}}} \times 100 \ (\%) \quad (1)$$

where, $\text{Score}_{\text{fcast}}$ quantifies the error measure in streamflow forecasts, $\text{Score}_{\text{ref}}$ is the same error measure applied to the reference forecasts, and $\text{Score}_{\text{perf}}$ is the measure

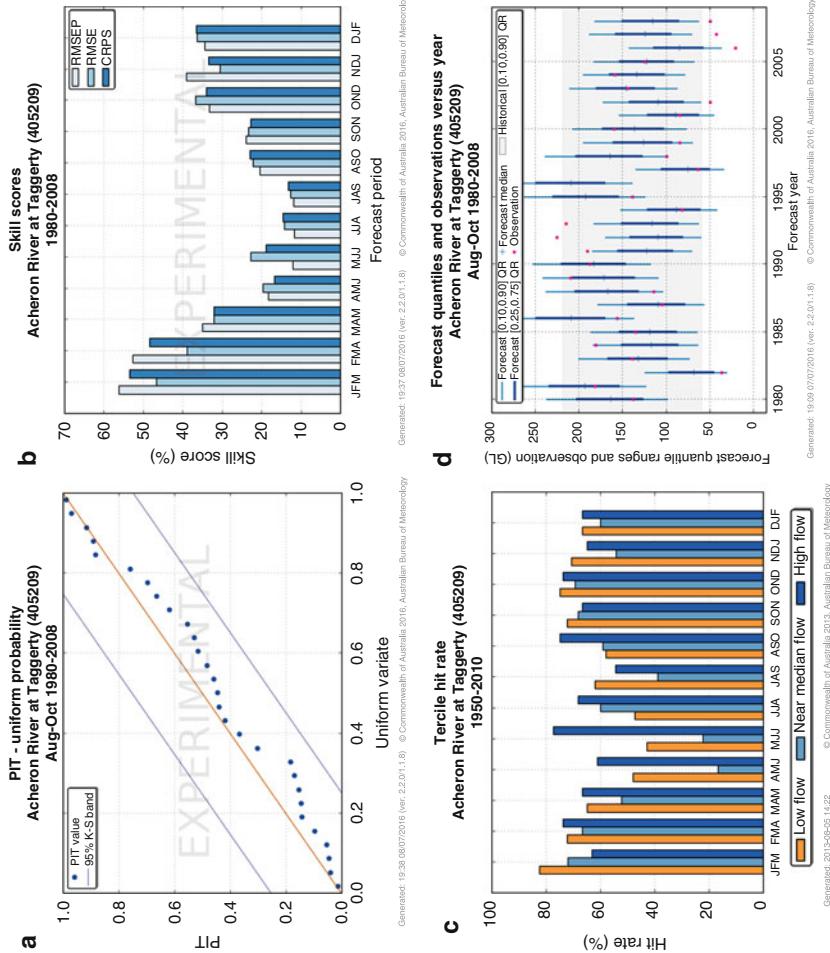


Fig. 3 (a) Forecast reliability represented through the Probability Integral Transform plot (PIT) or the predictive Q-Q plot. (b) Skill scores CRPS, RMSE, and RMSEP derived from forecast verification wrt the historical reference period (1980–2008). (c) Tercile hit rate for the low flow, near median flow, and high flow over the cross-validation period (d) Comparison of the interquantile range of the streamflow forecasts and observed streamflow over the cross-validation period 1950–2010

Table 1 Forecast rating thresholds for different metrics used to quantify accuracy and precision rating of the seasonal streamflow forecasts

Forecast rating	0	1	2	3	4	5
Reliability rating ^a (p-value)	$\leq 1.0\%$	< 2.5%	< 5.0%	< 7.5%	< 10.0%	$\geq 10.0\%$
Accuracy rating (CRPS)	≤ 0	< 10	< 20	< 30	< 40	≥ 40
Accuracy rating (RMSE)	≤ 0	< 10	< 20	< 30	< 40	≥ 40
Accuracy rating (RMSEP)	≤ 0	< 10	< 20	< 30	< 40	≥ 40
Accuracy rating – Low flow tercile	≤ 0.33	< 0.40	< 0.50	< 0.70	< 0.90	≥ 0.90
Accuracy rating – High flow tercile	≤ 0.33	< 0.40	< 0.50	< 0.70	< 0.90	≥ 0.90
Precision rating – IQR (10%, 90%)	≥ 1.0	> 0.95	> 0.85	> 0.75	> 0.6	≤ 0.60

^aReliability rating is derived from “p-value” obtained from the Kolmogorov-Smirnov test at significance thresholds varying from 1% to 10%

applied to perfect forecasts (any error measure might be used, e.g., CRPS; RMSE; RMSEP). The lower bound of the skill score depends on the selected measure, but the upper bound of the skill score is 100 (%) for any error measure. It is noteworthy that the actual value of a skill score can be different even for the same data, depending on the reference forecasts selected for the calculation. Note that observed streamflow for the period 1970–2010 (or starting year if observations commenced post 1970) is used as the reference period in this study and referred to as “climatology.”

The accuracy of the streamflow forecasts is derived using the following metrics – three skill scores based on the error measures, CRPS, RMSE, and RMSEP; tercile hit rate; and forecast precision (Fig. 3b, c; Table 1; see www.bom.gov.au/water/ssf; Laio and Tame 2007; Wang and Robertson 2011; Tuteja et al. 2011; Alfieri et al. 2014):

- *CRPS* – Continuous Rank Probability Score described by the area between the forecast distribution and a step function of the observation.
- *RMSE* – Root Mean Squared Error representing forecast errors directly in the measurement space.
- *RMSEP* – Root Mean Squared Error of forecast is the expected value of the distance between the median of streamflow forecasts and corresponding observations in probability space.
- *Tercile hit rate* – ratio of the binary outcome (1 or 0) depending on observation falling within the forecast tercile with largest probability and total number of observations (Fig. 1b and 3c).
- *Forecast precision* – determined as the ratio of the Inter Quantile Range (IQR) (10%, 90%) of the forecast distribution and IQR (10%, 90%) of the observations across the cross-validation period (IQR-80%; Fig. 3d).

Given that RMSE is based upon errors in the measurement space, it is sensitive to a small number of large errors relative to many small ones, and therefore it can potentially lead to conservative forecasts that are influenced by a few high flow events.

CRPS measures departure of the forecast distribution from observations, and like RMSE it is also sensitive to large errors. Note that for a deterministic forecast, the CRPS is reduced to MAE (Mean Absolute Error). CRPS is sensitive to ensemble size (Ferro et al. 2008), and therefore a large 5000-member ensemble size is used.

In the case of RMSEP, forecast errors are represented in the probability domain, and therefore, the influence of large errors in the original flow domain has a smaller impact on RMSEP compared to RMSE and CRPS. More frequent events have more influence on the errors in estimating RMSEP, which reduces the impact of outliers in the measurement space. However, for periods with a small variability in streamflow volumes, such as in the low flow season, a slight error in the measurement space can be amplified to an extremely large error in the probability space. In such a case, RMSEP can respond sensitively to even marginal errors in streamflow volumes.

Forecast precision for a given forecast location and month across all years is determined through the interquartile range. It is an important metric for determining optimal storage operations, e.g., elimination of the least plausible options in scenario analysis and designing environmental watering events for significant wetlands.

2.3 Aggregated Forecast Performance

In statistics, the Kolmogorov–Smirnov test (K–S test) is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare a sample with a reference probability distribution (one-sample K–S test) or to compare two samples (two-sample K–S test). The p-value from the K–S test is used to categorize the reliability rating (Table 1) which is derived from “p-value” obtained from the Kolmogorov–Smirnov test at significance thresholds varying from 1 to 10%. A “p-value” greater than 10% implies a high level of confidence in reliability of the forecasts, while a “p-value” of less than 1% implies a weak chance that forecast distributions are reliable.

Since each metric emphasizes different aspect of forecast accuracy, it is desirable to inspect all the performance metrics when comparing modeling results. The forecast accuracy rating for each of the metric – CRPS, RMSE, RMSEP, tercile hit rate, and forecast precision – is categorized across the range 0–5 (Table 1). Water managers in Australia are more interested in low flow and high flow conditions, and near median flow conditions are of least interest. Therefore, tercile hit rate accuracy of low flows and high flows are considered, and the use of near median flow hit rates is ignored in deriving overall composite indicator of the forecast performance.

Given high demand for water for consumptive and environmental use, a very conservative approach (often minimum observed flow) is used by water managers at the beginning of the annual water planning cycle, and the allocations are adaptively

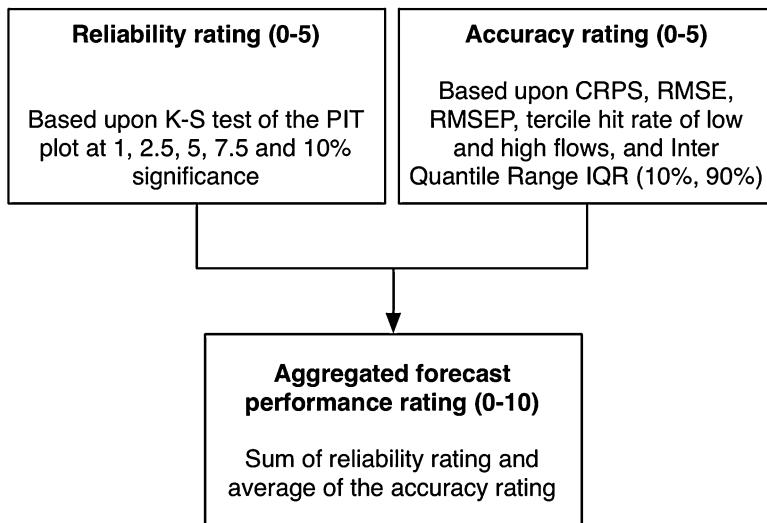


Fig. 4 Hierarchical approach to determine a single Aggregated Forecast Performance Index (AFPI) rating which is comprised of forecast reliability and accuracy of the streamflow forecasts. The rating is bounded between (0, 10) and is classified into 5 categories – low (0–2), moderate (2–4), medium (4–6), high (6–8), and very high (8–10)

increased over time during the filling season. For this reason, forecast precision rating based upon IQR (10%, 90%) is used by categorizing in the range 0–5 depending upon proportional narrowing of the uncertainty over climatology (Table 1).

The interrelationship between precision and reliability is important. A very precise forecast that turns out incorrect most of the time is undesirable. On the other hand, a very reliable forecast with large uncertainty too is unhelpful to water managers because a very small number of possibilities can only be eliminated in their scenario planning.

Finally, the aggregated streamflow forecast performance rating is determined using a hierarchical approach wherein forecast reliability or the baseline forecasts and accuracy representing incremental improvements over the baseline forecasts are all used to derive the Aggregated Forecast Performance Index (AFPI) (Fig. 4).

3 Results and Discussion

3.1 Forecast Reliability

The PIT uniform probability plots are used here to assess the overall reliability of the forecasts for individual locations. The PITs will be uniformly distributed if the forecasts are reliable. The p-value derived from the Kolmogorov-Smirnov (K-S) uniform test of the PITs can be used to indicate the reliability of the forecast distributions (Fig. 5). It is shown that the higher the p-value the better the

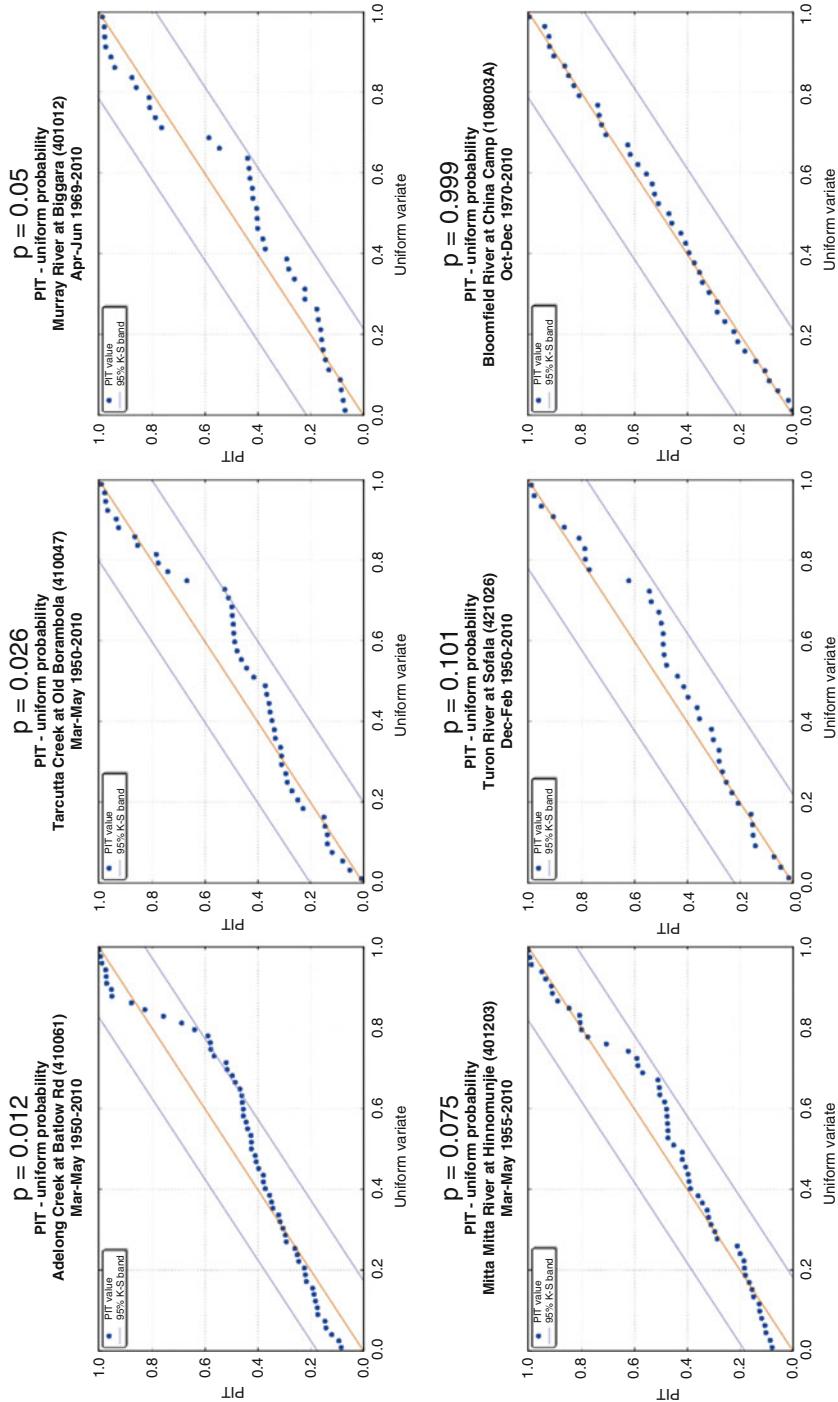


Fig. 5 PIT uniform probability plots with different p-value for 6 different forecast locations

reliability of the forecast distribution. The plots are suitable for small samples sizes, and they are also useful for diagnosing problems, such as bias and/or an unrealistic ensemble spread. However, the p-value alone cannot represent all aspects of the forecast errors. Following an investigation of the p-value across all sites and forecast periods along with a visual inspection of the PIT plots, it is concluded that seasonal streamflow forecasts derived from the BJP model are reliable (see also Sect. 3.3).

3.2 Forecast Accuracy

3.2.1 Skill Scores Based on CRPS, RMSE, and RMSEP

Streamflow forecast accuracy is assessed using the CRPS, RMSE, and RMSEP skill scores. The strengths and limitations of each metric were discussed in Sect. 2.2.2. Figure 6 shows the overall skills of the seasonal streamflow forecasts derived across the cross-validation period for the 147 service locations. On average, the forecasts are more skilful for the second half of the year which coincides with the filling of the storages, active growing season, and high irrigation demand. On average, the RMSEP skill score across all sites is the highest and the RMSE skill score is the lowest among the three metrics.

Since each skill score emphasizes different aspects of the forecast accuracy, it is desirable to inspect all the skill scores when assessing modeling results. It is evident from Fig. 2 that the seasonal streamflow forecasts are accurate and that they add improvements over climatology. Note that baseline forecasts from climatology itself

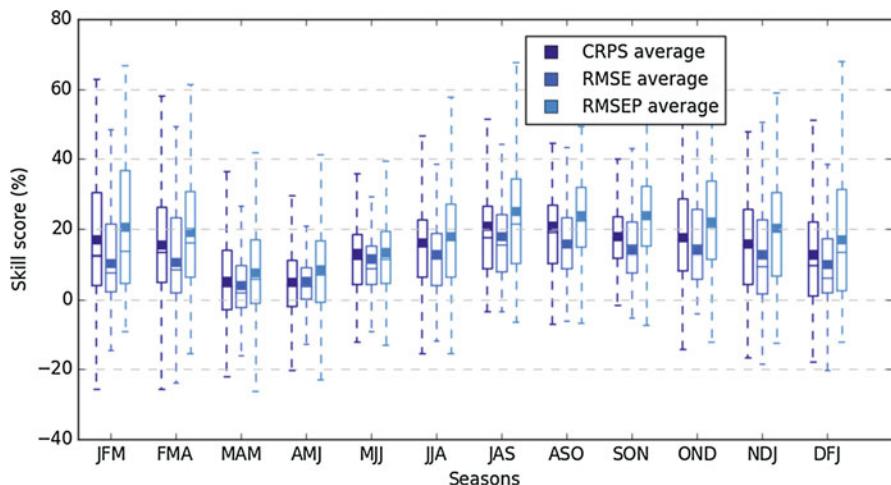


Fig. 6 CRPS, RMSE, and RMSEP skill scores of seasonal streamflow forecasts for the 147 service locations

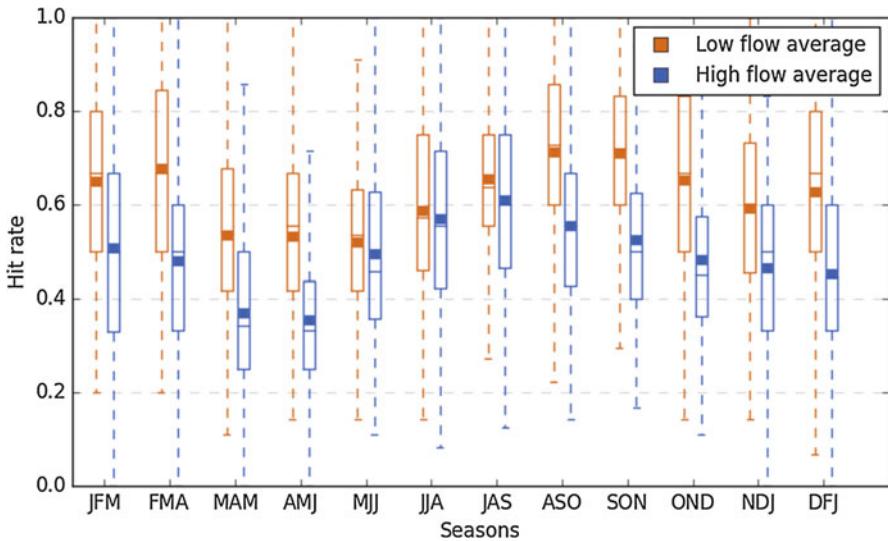


Fig. 7 Tercile hit rates for low and high flows of the seasonal streamflow forecasts generated for the 147 service locations

provide a reasonable basis given that observations at all the forecast locations extend beyond 30 years.

3.2.2 Tercile Hit Rates for Low and High Flows

Forecast accuracy in terms of the tercile hit rates for low and high flows for all 147 forecast locations across the cross-validation period is high (Fig. 7). On average, the forecasts for both high and low flows are better for the period extending from July to October. This is indeed an important outcome for water managers in that tercile forecasts are commonly used for water resource planning in Australia.

3.2.3 Forecast Precision

Forecast precision is assessed on the basis of the mean ratio of the interquartile range derived across the cross-validation period. Figure 8 shows the mean ratios of forecast interquartile range (10%, 90%), denoted as IQR-80%, and divided by climatology for the 147 service locations. On average, the forecasts have higher precision in the second half of the year. Across all sites and forecast periods, streamflow forecast uncertainty is narrower than climatology, and the reduction in uncertainty ranges, on average around 15%. Given that streamflow forecasts are reliable as well as more precise than climatology, they do offer insights for water managers to discount least plausible options in their scenario planning.

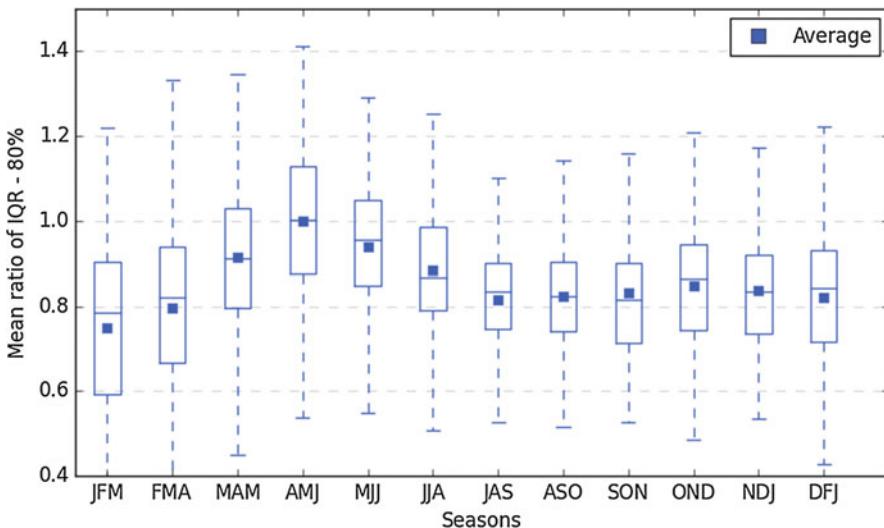


Fig. 8 Mean ratios of forecast interquartile range (10%, 90%) divided by climatology for the 147 service locations

3.3 Aggregated Forecast Performance

The p-value derived from the Kolmogorov-Smirnov (K-S) uniform test of the PITs is categorized across the range 0–5 as a measure to indicate the reliability of the forecast distributions (Fig. 7; Table 1). Figure 9 shows the reliability rating derived from the K-S p-value for the JFM, AMJ, JAS, and OND forecast periods and indicates that seasonal streamflow forecasts are very reliable for most service locations and forecast periods.

The forecast accuracy rating for each of the error measures – CRPS, RMSE, RMSEP skill scores and tercile hit rate – is categorized across the range 0–5 (Table 1). The forecast precision rating based upon the mean ratio of IQR-80% is used by categorizing in the range 0–5 (Table 1). Figure 10 shows the average of the precision rating and the five accuracy ratings for the JFM, AMJ, JAS, and OND forecast periods. The figure appropriately represents the incremental improvements over the baseline forecasts.

The final streamflow forecast performance index is aggregated from the forecast reliability rating and the average of the precision rating and the five accuracy ratings. Figure 11 shows the aggregated forecast performance index (AFPI) for the JFM, AMJ, JAS, and OND forecast periods. It demonstrates that the AFPI can adequately summarize the performance metrics of the seasonal streamflow forecasts. Therefore, it can be regarded as a simplified approach to communication of the probabilistic seasonal streamflow forecast performance across key water supply catchments in Australia.

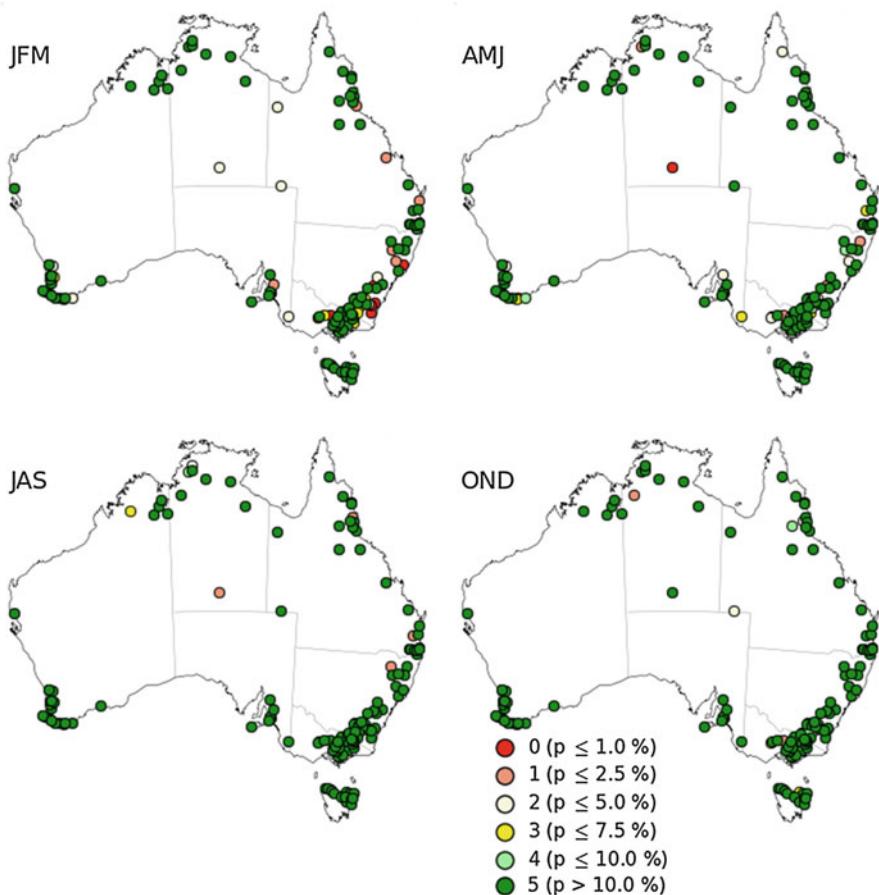


Fig. 9 Reliability rating based on the p-value of the K-S uniform test of forecast distributions

4 Conclusion

The Seasonal Streamflow Forecasting Service delivers monthly updates of 3 month ensemble-based forecasts at 147 locations across 75 river basins using the statistical *Bayesian Joint Probability* (BJP) model which is operational within the end-to-end seasonal forecasting system *WAFARi*. The current service coverage extends largely across eastern Australia (at September 2014), and work is underway to extend the service to other jurisdictions.

The BJP model is calibrated and verified separately for each site and forecast period using a rigorous “leave 1-year out” cross-validation procedure for 1980–2008. The service delivers extensive forecast verification products at a granular level for each forecast location including metrics that describe reliability (PIT

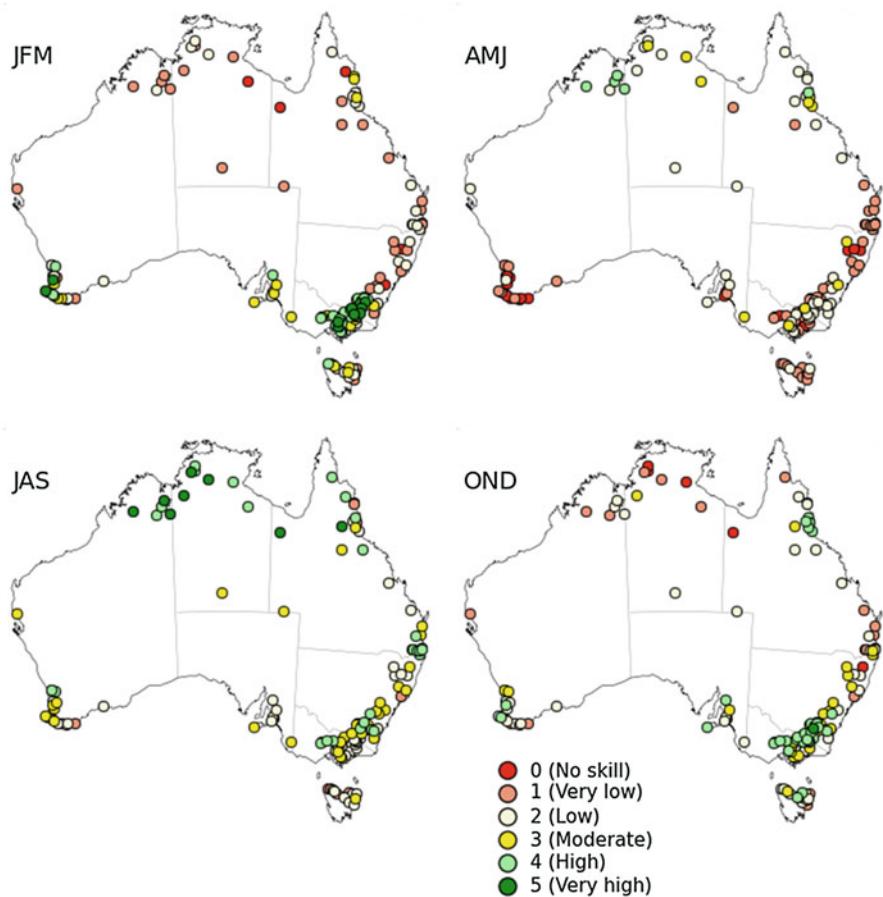


Fig. 10 Average of the six accuracy ratings – CRPS, RMSE, RMSEP, tercile hit rate of low and high flows, and interquartile range (10%, 90%)

plots), accuracy (CRPS, RMSE, RMSEP), tercile hit rates for low and high flow range, and precision represented by the interquantile range 10–90%).

Water managers in Australia now increasingly rely on these forecast products in their decision making, e.g., reservoir management, water allocation planning and delivery, and environmental flow management to support water markets and trading. Increased adoption of the water availability forecasts in Australia is directly dependent on improved user understanding of the forecast quality. A wide range of forecast verification products is available through the seasonal streamflow forecasting service of the Bureau of Meteorology. However, the products are innately complex and are not easily understandable as they require advanced competencies in statistics. Therefore, there is a need to summarize various forecast verification products with a simple and easily understandable composite forecast performance measure(s). In doing this, the

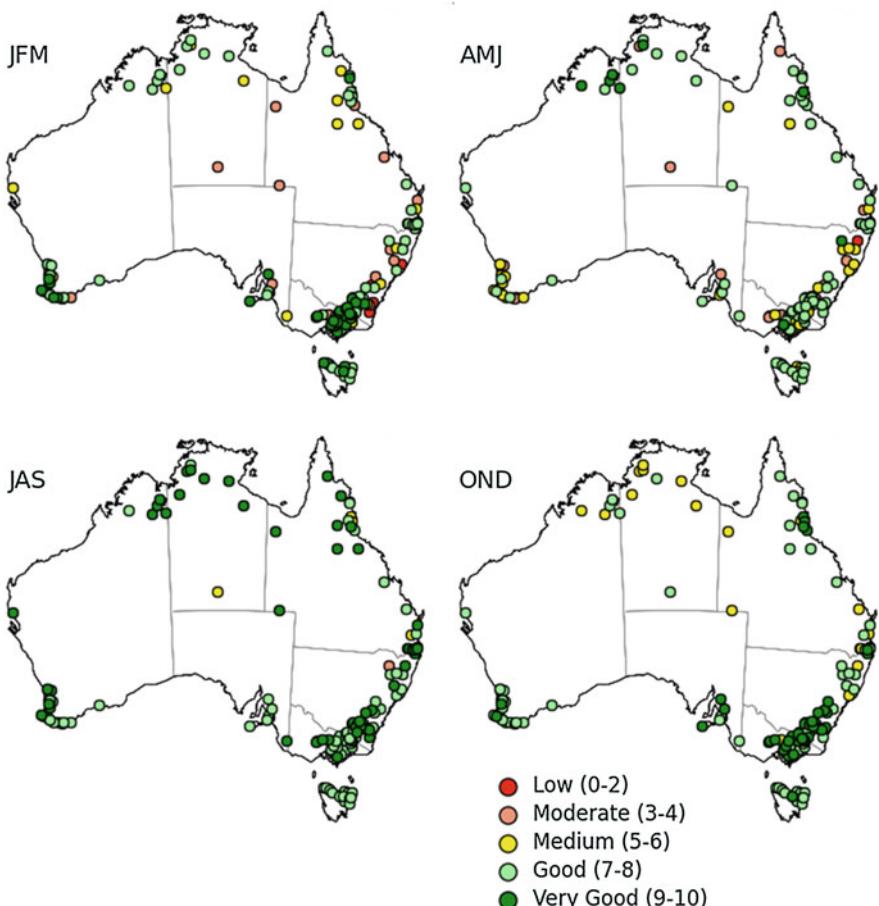


Fig. 11 Aggregated forecast performance index (AFPI) of seasonal streamflow forecasts. JFM and JAS are the high flow season in northern and southern Australia, respectively. AMJ and OND are transition periods between high flow and low flow regimes.

overall simplified forecast performance measure ought to make use of the multitude of performance measures derived from rigorous cross-validation methods discussed here, each of which have their own strengths and limitations. A simple forecast performance measure, called the “*Aggregated Forecast Performance Index (AFPI)*” was presented which captures key attributes such as forecast reliability and accuracy. It was demonstrated that Bureau’s seasonal streamflow forecasts are very reliable. They also offer improved accuracy by narrowing down the forecast uncertainty (up to 25%) with respect to reference climatology and hence offer a value proposition for water managers to improve their decision making. Finally, on the basis of a single easy-to-understand and well-informed aggregated measure, the skill of the Bureau’s seasonal streamflow forecasts has been communicated to its users effectively.

References

- L. Alfieri, F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, P. Salamon, Evaluation of ensemble streamflow predictions in Europe. *J. Hydrol.* **517**, 913–922 (2014). <https://doi.org/10.1016/j.jhydrol.2014.06.035>
- F. Chiew, J. Vaze, K.J. Hennessy, *Climate Data for Hydrologic Scenario Modelling Across the Murray-Darling Basin: A Report to the Australian Government from the CSIRO Murray-Darling Basin Sustainable Yields Project* (CSIRO, Canberra, 2008)
- A.P. Dawid, Present position and potential developments: some personal views: statistical theory: the prequential approach. *J. R. Stat. Soc. Ser. A* **147**, 278–292 (1984)
- J.G. De Gooijer, D. Zerom, Kernel-based multistep-ahead predictions of the US short-term interest rate. *J. Forecast.* **19**, 335–353 (2000)
- C.A.T. Ferro, D.S. Richardson, A.P. Weigel, On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol. Appl.* **15**, 19–24 (2008). <https://doi.org/10.1002/met.45>
- T. Gneiting, F. Balabdaoui, A.E. Raftery, Probabilistic forecasts, calibration and sharpness. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **69**, 243–268 (2007)
- F. Laio, S. Tamea, Verification tools for probabilistic forecasts of continuous hydrological variables, *Hydrology and Earth System Sciences. Hydrol. Earth Syst. Sci.* **11**, 1267–1277 (2007). www.hydrol-earth-syst-sci.net/11/1267/2007/
- T.A. McMahon, B.L. Finlayson, A. Haines, R. Srikanthan, Runoff variability: a global perspective. *IASH-AISH* **168**, 3–11 (1987)
- R.E. Morss, J.L. Demuth, J.K. Lazo, Communicating uncertainty in weather forecasts: a survey of the US public. *Weather Forecast.* **23**(5), 974–991 (2008)
- A.H. Murphy, What is a good forecast? An essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**(2), 281–293 (1993)
- D.E. Robertson, Q.J. Wang, A Bayesian approach to predictor selection for seasonal streamflow forecasting. *J. Hydrometeorol.* **13**, 155–171 (2012). <https://doi.org/10.1175/JHM-D-10-05009.1>
- D. Shin, A. Schepen, T. Peatey, S. Zhou, A. MacDonald, T. Chia, J. Perkins, N. Plummer, WAFARI: a new modelling system for seasonal streamflow forecasting service of the Bureau of Meteorology, Australia. *MODSIM2011*. Perth (2011).
- M. Thyre, B. Renard, D. Kavetski, G. Kuczera, S.W. Franks, S. Srikanthan, Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. *Water Resour. Res.* **45**, 22 (2009)
- N.K. Tuteja, D. Shin, R. Laugesen, U. Khan, Q. Shao, E. Wang, M. Li, H. Zheng, G. Kuczera, D. Kavetski, G. Evin, M. Thyre, A. MacDonald, T. Chia, B. Le, Experimental evaluation of the dynamic seasonal streamflow forecasting approach, Technical Report, *Bureau of Meteorology*, Melbourne (2011). http://www.bom.gov.au/water/about/publications/document/dynamic_seasonal_streamflow_forecasting.pdf
- R.A. Vertessy, Water information services for Australians. *Aust. J. Water Resour.* **16**(2), 91–106 (2013). <https://doi.org/10.7158/W13-MO01.2013.16.2>
- Q.J. Wang, D.E. Robertson, Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.* **47**, W02546 (2011)
- Q.J. Wang, D.E. Robertson, F.H.S. Chiew, A Bayesian joint probability modeling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.* **45**, W05407 (2009)
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences – An Introduction* (Academic, San Diego, 1995)
- T. Wilson, P. Feikema, J. Ridout, 2013 User feedback on the seasonal streamflow forecasts service, Bureau of Meteorology (2014)

Part X

Ensemble Forecast Application and Showcases



Introduction to Ensemble Forecast Applications and Showcases

Massimiliano Zappa, S. J. van Andel, and Hannah L. Cloke

Abstract

Hydrometeorological ensemble forecasting has gradually entered the offices of operational water managers changing forecasting practice across a wide range of applications. This section illustrates the range of applications available today, with the following showcases: Where in the world have hydrological ensemble prediction systems (HEPS) found practical application and are increasingly yielding added value in hazard mitigation? Can I anticipate flash floods? Can I anticipate and cope with floods? Do I really need to draw down my reservoir and make room for a flood that might not occur? Can I optimize the production of hydropower reservoir operations? Is there any chance that my municipality might be affected by critical water shortages in the coming weeks? What do we learn from the operation of a continental-scale prediction system addressing users in different countries? How does communication between forecast centers and end users work in real-life operations? These showcases illustrate some of the numerous HEPS that have been implemented around the world. We hope that the success stories and pitfalls discussed in these examples will inspire and support successful development of further new applications.

M. Zappa

Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

e-mail: massimiliano.zappa@wsl.ch; zappa@wsl.ch

S. J. van Andel

UNESCO-IHE Institute for Water Education, Delft, The Netherlands

e-mail: s.vanandel@un-ihe.org

H. L. Cloke (✉)

Department of Meteorology, Reading University, Reading, UK

e-mail: h.l.cloke@reading.ac.uk

Keywords

Ensemble showcase · Water management · Hazard mitigation · Flash floods · Hydropower · Forecast communication

Hydrometeorological ensemble forecasting has gradually entered the offices of operational water managers changing forecasting practice across a wide range of applications. This section illustrates the range of applications available today, with the following showcases:

Where in the world have hydrological ensemble prediction systems (HEPS) found practical application and are increasingly yielding added value in hazard mitigation?

Pappenberger et al. provide a worldwide record of currently deployed HEPS targeted at providing support to prevent and mitigate floods, droughts, and other hydrology-related hazards. This record is however a snapshot, with new systems coming online frequently, as HEPS continues to grow in popularity and state-of-the-art research transforms into useful applications for agencies responsible for warning and intervention.

Can I anticipate flash floods?

Wittwer et al. and Alfieri et al. introduce four different ensemble approaches for early prediction of flash floods. These contributions give an insightful overview on how the problem of prediction of flash floods in gauged and ungauged areas has been addressed in France, Catalonia, Southern Switzerland, and up to the European scale (see also Alfieri and Thielen 2015).

Can I anticipate and cope with floods?

HEPS are also needed as input for forecasting other kinds of hazards. Mueller et al. provide an example of such a system which propagates hydrological ensembles into a hydraulic model. The consideration of the river morphology and the hydraulic structures gives an example of how added value can be obtained from HEPS. The tools used by Mueller et al. also demonstrate how the combination of HEPS and protection measures can potentially reduce damages yielded by flood inundation.

Do I really need to draw down my reservoir and make room for a flood that might not occur?

Reservoirs are useful structural measures in order to guarantee flood protection of the downstream areas of a catchment. Since 2008, the city of Zürich has relied on such a strategy to cope with critical flood situations. An operational HEPS delivers

useful information for deliberating on a controlled drawdown of an artificial lake (Addor et al. 2011). This issue is also elaborated in the subchapter by Wang et al. in case of the Han River in China. Boucher and Ramos also address this topic, pointing at the need for short-term and medium-range HEPS in order to manage reservoirs accordingly in case of flood situations.

Can I optimize the production of hydropower reservoir operations?

The current situation of energy demand, production, and long-term targets cause several new challenges for the hydropower industry (Björnsen-Gurung et al. 2016). New renewable energy such as wind and solar energy mixed up the market and management of energy in Europe and in several other regions of the world. Boucher and Ramos give an overview on the services that hydropower reservoirs have to provide within the energy sector and how HEPS can contribute to optimal operations. The contribution by Tucci et al. focuses on a similar topic and provides useful experience on the long-term use of HEPS in reservoir management. While the science of using HEPS started in the early 2000s with the hindcast of single (extreme) events, it is nowadays more and more necessary to evaluate HEPS in a fully (quasi-)operational framework in order to also identify the ratio of false alarms.

Is there any chance that my municipality might be affected by critical water shortages in the coming weeks?

Using probabilistic decision-making tools can help users of very complex systems to effectively take account of uncertainties. One such example is that of water supply and considerations such as the threat of critical water shortage. Porter et al. provide some insight into a new Operations Support Tool for water supply operations decisions for the New York City Department of Environmental Protection. The tool is designed to help take decisions in a system with very large complexities. It is driven with ensemble hydrological forecasts allowing probabilistic decision-making.

How much goods can I load on my boat ready to navigate from Amsterdam to Basel?

Application of HEPS is increasingly useful for questions that go well beyond the realm of hydrology, natural hazard mitigation, and hydropower production. There is one specific economic sector that might be seen as a key potential customer of hydrological ensemble predictions, that being the navigation industry, and that manages the transportation of goods along rivers. One of the most frequent measures used to evaluate the skill of HEPS is the cost-loss ratio (Roulin 2007). A decision on volume of goods loaded onto a ship could be linked to the probability of having favorable conditions on the river during the whole duration of the trip. Meißner et al. (2017, accepted) present an example on how HEPS are applied in the shipment industry along major European rivers.

What do we learn from the operation of a continental-scale prediction system addressing users in different countries?

Operating a continental-scale HEPS system for the early prediction of floods and other hydrological events is particularly challenging in terms of data management and communication. The European Flood Awareness System (EFAS) (previously known as the European Flood Alert System) (Thielen et al. 2009) was one of the first systems using HEPS. After more than a decade of operations, Thielen et al. present a summary of the experience gained since the operational deployment of EFAS in 2003. One of the recent targets for HEPS research and applications is that of implementing HEPS at the seasonal timescale, as demonstrated by the collection of papers within the HESS special issue on the topic (https://www.hydrol-earth-syst-sci.net/special_issue824.html). In this volume, Wood et al. demonstrate seasonal drought forecasts for the United States and present a verification over several years of forecasts.

How does communication between forecast centers and end users work in real-life operations?

HEPS predictions can be presented in several different ways (Bruen et al. 2010), and visualizations of these predictions are usually quite different to what users are used to. So, in the end, the success of any forecasting chain also depends on effective communication with end users. While in some cases demonstration projects might trigger interest in HEPS (Zappa et al. 2008), only tailored training and multi-way dialogue can ease the task of users taking decisions under considerations of uncertainties (Wetterhall et al. 2013). Wittwer et al. and Thielen et al. give some interesting opinions central to this issue of operational HEPS systems.

These showcases illustrate some of the numerous HEPS that have been implemented around the world. We hope that the success stories and pitfalls discussed in these examples will inspire and support successful development of further new applications.

References

- N. Addor, S. Jaun, F. Fundel, M. Zappa, An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* **15**, 2327–2347 (2011). <https://doi.org/10.5194/hess-15-2327-2011>
- L. Alfieri, J. Thielen, A European precipitation index for extreme rain-storm and flash flood early warning. *Meteorol. Appl.* **22**(1), 3–13 (2015). <https://doi.org/10.1002/met.1328>
- A. Björnsen Gurung, A. Borsdorf, L. Fürerer, F. Kienast, P. Matt, C. Scheidegger, L. Schmocke, M. Zappa, K. Volkart, Rethinking pumped storage hydropower in the European Alps. *Mt. Res. Dev.* **36**(2), 222–232 (2016). <https://doi.org/10.1659/MRD-JOURNAL-D-15-00069.1>
- M. Bruen, P. Krahe, M. Zappa, J. Olsson, B. Vehviläinen, K. Kok, K. Daamen, Visualizing flood forecasting uncertainty: some current European EPS platforms – COST731 working group 3. *Atmos. Sci. Lett.* **11**, 92–99 (2010). <https://doi.org/10.1002/asl.258>

- D. Meißner, B. Klein, M. Ionita, Development of a monthly to seasonal forecast framework tailored to inland waterway transport in Central Europe. *Hydrol. Earth Syst. Sci.* **21**, 6401–6423 (2017). <https://doi.org/10.5194/hess-21-6401-2017>
- E. Roulin, Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrol. Earth Syst. Sci.* **11**, 725–737 (2007)
- J. Thielen, J. Bartholmes, M.-H. Ramos, A. de Roo, The European flood alert system – part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125–140 (2009). <https://doi.org/10.5194/hess-13-125-2009>
- F. Wetterhall, F. Pappenberger, L. Alfieri, H.L. Cloke, J. Thielen-del Pozo, S. Balabanova, J. Daňhelka, A. Vogelbacher, P. Salamon, I. Carrasco, A.J. Cabrera-Tordera, M. Corzo-Toscano, M. García-Padilla, R.J. García-Sánchez, C. Ardilouze, S. Jurela, B. Terek, A. Csik, J. Casey, G. Stankūnavičius, V. Ceres, E. Sprokkereef, J. Stam, E. Anghel, D. Vladikovic, C. Alionte Eklund, N. Hjerdt, H. Djerv, F. Holmberg, J. Nilsson, K. Nyström, M. Sušnik, M. Hazlinger, M. Holubecka, HESS opinions “forecaster priorities for improving probabilistic flood forecasts”. *Hydrol. Earth Syst. Sci.* **17**, 4389–4399 (2013). <https://doi.org/10.5194/hess-17-4389-2013>
- M. Zappa, M.W. Rotach, M. Arpagaus, M. Dominger, C. Hegg, A. Montani, R. Ranzi, F. Ament, U. Germann, G. Grossi, S. Jaun, A. Rossa, S. Vogt, A. Walser, J. Werhan, C. Wunram, MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems. *Atmos. Sci. Lett.* (2008). <https://doi.org/10.1002/asl.183>. 8 pp



Hydrological Ensemble Prediction Systems Around the Globe

Florian Pappenberger, Thomas C. Pagano, J. D. Brown, Lorenzo Alfieri, D. A. Lavers, L. Berthet, F. Bressand, Hannah L. Cloke, M. Cranston, J. Danhelka, J. Demargne, N. Demuth, C. de Saint-Aubin, P. M. Feikema, M. A. Fresch, R. Garçon, A. Gelfan, Y. He, Y. -Z. Hu, B. Janet, N. Jurdy, P. Javelle, L. Kuchment, Y. Laborda, E. Langsholt, M. Le Lay, Z. J. Li, F. Mannessiez, A. Marchandise, R. Marty, D. Meißner, D. Manful, D. Organde, V. Pourret, Silke Rademacher, Maria-Helena Ramos, D. Reinbold, S. Tibaldi, P. Silvano, Peter Salamon, D. Shin, C. Sorbet, Eric Sprokkereef, V. Thiemig, Narendra Kumar Tuteja, S. J. van Andel, Jan S. Verkade, B. Vehviläinen, A. Vogelbacher, Fredrik Wetterhall, Massimiliano Zappa, R. E. Van der Zwan, and Jutta Thielen-del Pozo

Contents

1	Introduction	1189
2	Systems' Descriptions	1207
2.1	Target Variables	1207
2.2	Systems' Objectives	1208
2.3	Forecast Horizon	1210
2.4	Hydrological Models and Forcings	1210
2.5	Uncertainties Considered	1212
2.6	Users	1214
3	Conclusions: Summary and Future Outlook	1215
	References	1216

F. Pappenberger (✉)

European Centre for Medium-Range Weather Forecasts, ECMWF, Reading, UK
e-mail: florian.pappenberger@ecmwf.int

T. C. Pagano · P. M. Feikema · D. Shin
Bureau of Meteorology, Melbourne, VIC, Australia

J. D. Brown
Hydrologic Solutions Limited, Southampton, UK

L. Alfieri
Directorate for Space, Security and Migration, European Commission – Joint Research Centre, Ispra, VA, Italy

Abstract

A large number of hydrological forecasting systems exist across the globe. Recent advances have pushed the limits of predictability of discharge and other hydrological variables from a few hours to several days or even months. In this chapter, we aim to give an overview of Hydrological Ensemble Prediction Systems across the globe. It provides brief descriptions of existing or preoperational systems as background, and discusses the challenges ahead. This overview shows that there is at least one system per continent, though their geographic domain varies considerably among very small catchments, countries national and interregional basins, transnational basins, continents, or even the entire globe. It highlights common challenges and differences.

D. A. Lavers

European Centre for Medium Range Weather Forecasts, Reading, UK

L. Berthet

Loire river Flood Forecasting Centre, Orléans, Italy

F. Bressand · Y. Laborda · F. Mannessiez

Service de Prévision des Crues Grand Delta, Nîmes, France

H. L. Cloke

Department of Meteorology, Reading University, Reading, UK

Department of Environmental Sciences and Geography, Reading University, Reading, UK

M. Cranston

RAB Consultants/University of Dundee, Stirling, Italy

J. Danhelka

Czech Hydrometeorological Institute, Prague, Czech Republic

J. Demargne · D. Organde

HYDRIS Hydrologie, Saint Mathieu de Tréviers, France

N. Demuth

Landesamt für Umwelt, Rhineland Palatinate, Mainz, Germany

C. de Saint-Aubin · B. Janet

Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI), Toulouse, France

M. A. Fresch

Office of Water Prediction, U.S. National Weather Service, Silver Spring, MD, USA

R. Garçon · M. Le Lay

EDF DTG, Grenoble, France

A. Gelfan · L. Kuchment

Water Problems Institute of Russian Academy of Sciences (WPI RAS), Moscow, Russia

Y. He · D. Manful

Tyndall Centre for Climate Change Research, School of Environmental Sciences, University of East Anglia, Norwich, UK

Keywords

African Flood Forecasting System (AFFS) · AquaLog/Hydrog LAEF · Community Hydrologic Prediction System (CHPS) · Ensemble Prediction System · European Flood Awareness System (EFAS) · Flood Early Warning System for the Po River (FEWSPo) · Forecast centres · Forecast horizons · Global Flood Awareness System · GloFAS · HEPS · HUGO · Hydrologic Ensemble Forecast Service (HEFS) · Hydrological ensemble prediction systems (HEPS) · Hydrological models and forcings · LARSIM forecast systems · MMEFS · PREDICTOR · Pre-operational HEPS systems · RWsOS Rivers · Thorpex Interactive Grand Global Ensemble (TIGGE) · Water level forecast

1 Introduction

There has been a surge in the scientific development and implementation of operational hydrological ensemble prediction systems (HEPS, see Fig. 1 for a typical forecast). Recent advances have pushed the limits of predictability of discharge and

Y.-Z. Hu · Z. J. Li · D. Meißner · S. Rademacher
German Federal Institute of Hydrology (BfG), Koblenz, Germany

N. Jurdy
Service de Prévision des Crues Meuse-Moselle, Metz, France

P. Javelle
Irstea, OHAX Hydrology Unit, Aix-en-Provence, France

E. Langsholt
Ministry of Petroleum and Energy, Norwegian Water Resources and Energy Directorate, Hydrology Department (NVE), Oslo, Norway

A. Marchandise
Service de Prévision des Crues Méditerranée Ouest, Carcassonne, France

R. Marty · D. Reinbold
Loire-Cher-Indre Flood Forecasting Centre, Orléans, France

V. Pourret · C. Sorbet
Météo-France, Toulouse, France

M.-H. Ramos
IRSTEA, Antony, France

S. Tibaldi
ARPA Emilia Romagna, Bologna, Italy

P. Silvano
ARPA Emilia Romagna, Parma, Italy

P. Salamon · V. Thiemig
European Commission, Joint Research Centre, Ispra, Italy

E. Sprokereaef
Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, The Netherlands

other hydrological variables from a few hours to several days or even months (Croke and Pappenberger 2009; Thielen et al. 2009; Cuo et al. 2011). For example, drought forecasting systems are now often skilful for some months ahead. Flood events, however, are usually predictable on much shorter time scales and tend to rely on medium (3–15 days) or short (up to 3 days) range forecasts. Other aspects, such as minimum and maximum water depths for shipping, for instance, have seasonal predictability, but with factors (e.g., out of bank flow) affecting them also on the short to medium range. It is therefore impossible to make a generalized statement on predictability for all operational and preoperational HEPS applications across the globe, given the uniqueness of hydrological catchments and the variety of objectives in forecasting (Thielen et al. 2009). Also the quality of the inputs to hydrological models is important; e.g., droughts operate over larger spatial domains which likely encompass many weather forecast model (or climate model) grid cells, whereas floods can occur in much smaller locations and on a spatial scale which is difficult to predict using even the highest resolution forecast models, because the domains of interest are only a few grid cells large.

However, there is no scientific doubt that the use of ensemble forecasting (i.e., multiple forecasts) enhances the available forecast information (Alfieri et al. 2012). The uptake of ensembles in operational systems nevertheless is often made difficult by cultural, legal, monetary, and technical challenges (Pagano et al. 2014). Therefore, there are still relatively few operational HEPS in practice (Croke et al. 2009).

N. K. Tuteja

Bureau of Meteorology, Canberra, ACT, Australia

S. J. van Andel

UNESCO-IHE Institute for Water Education, Delft, The Netherlands

J. S. Verkade

Deltareas, Delft, The Netherlands

Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands,
River Forecasting Service, Lelystad, The Netherlands

Delft University of Technology, Delft, The Netherlands

B. Vehviläinen

Finish Environment Institute, Helsinki, Finland

A. Vogelbacher

Bayerisches Landesamt für Umwelt, Augsburg, Germany

F. Wetterhall

Forecast Department, European Centre for Medium-Range Weather Forecasts, Reading, UK

M. Zappa

Swiss Federal Institute for Forest, Snow and Landscape Research WSL, Zurich, Switzerland

R. E. Van der Zwan

Principal Water Board of Rijnland, Leiden, The Netherlands

J. Thielen-del Pozo (✉)

European Commission, Joint Research Centre, Ispra, Italy

e-mail: jutta.thielen@jrc.ec.europa.eu

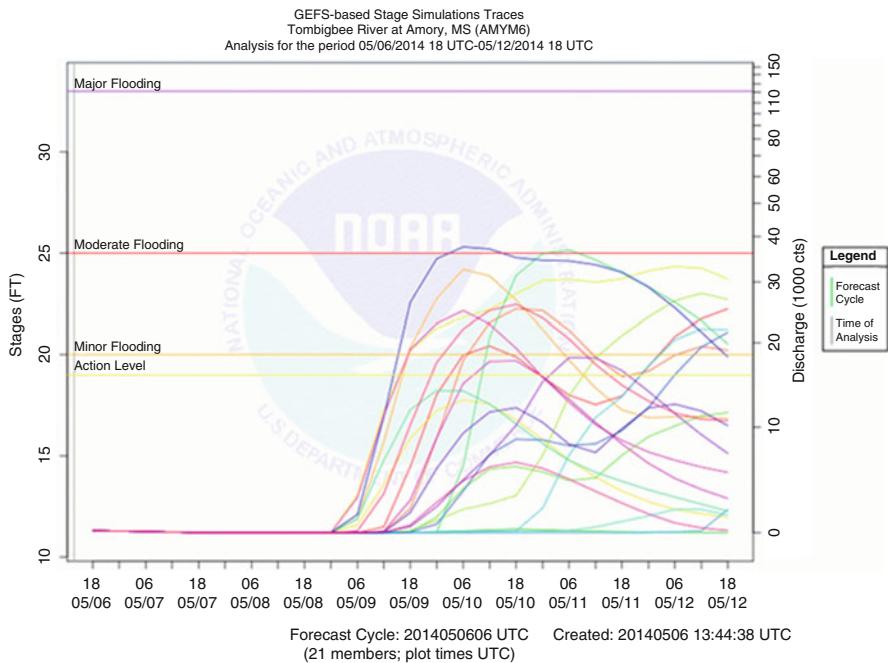


Fig. 1 A typical ensemble flood forecast (Source: MMEFS system at <http://www.weather.gov/>)

Even when HEPS have been made operational, their relatively recent implementation means that any forecast evaluation can only be undertaken on a short record, which precludes any significant assessment on predicting extreme values (Pappenberger et al. 2011); this results in forecast evaluation being more frequently done on case studies (Younis et al. 2008; He et al. 2009). Given the nascent nature of these operational systems, and the different contexts and domains in which they are applied, comparisons of operating standards are rare and community standards of practice are in their infancy. Nonetheless, learning from other systems' experience is a wise strategy and may foster the implementation of a new system or the upgrade of an existing one.

In this chapter, we aim to give an overview of Hydrological Ensemble Prediction Systems across the globe. This chapter provides brief descriptions of existing or preoperational systems as background, and discusses the challenges ahead. The various operational HEPS considered in this chapter are listed in Table 1 (S1 to S24). This overview shows that there is at least one system per continent, though their geographic domain varies considerably between extremely small catchments (e.g., 10 km² for S13), countries (e.g., Scotland for S8 or Finland for S20), national and interregional basins (e.g., S4), transnational basins (e.g., the Rhine in S16 or Moselle in S19), continents (e.g., Africa in S15), or even the entire globe (S14, see Fig. 2 for river reaches covered by the global system GloFAS).

Table 1 Forecast centers with operational or preoperational HEPS systems

ID	System acronym	Forecast center/System name:	System overview	References
S1	EFAS	Forecast center/System name: European Flood Awareness System Status: Operational Provider: European Commission -Copernicus program Domain: Europe	<p>The European Flood Awareness System (EFAS) has been developed at the Joint Research Center of the European Commission (EC-JRC) since 2002, in close collaboration with national hydrological and meteorological services and other European research institutes. The system was designed to give a European overview of ongoing floods and to forecast floods with the aim of early warning for national and transnational river basins at imminent risk of extreme runoff conditions. The uncertainty of weather forecasts is accounted for using a multimodel approach, i.e., predictions come from different atmospheric circulation models, including deterministic weather predictions and the ensemble prediction system of the European Centre for Medium-Range Weather Forecasts. Weather predictions are the input to the Lisflood hydrological model. Model outputs are daily forecasts of discharge up to 15 days in advance. The system became fully operational in 2012</p>	Thielen et al. 2009; van der Krijff et al. 2010; Pappenberger et al. 2011; Afferet et al. 2014; van der Krijff et al. 2010; Burek et al. 2013; Emerton et al. 2015; www.efas.eu
S2	HEFS	Forecast center/System name: Hydrologic Ensemble Forecasting Service Status: Operational Provider: US National Weather Service Domain: USA	<p>The Hydrologic Ensemble Forecast Service (HEFS) was developed by the Office of Hydrologic Development of the US National Weather Service (NWS). The HEFS models the total uncertainty in future streamflow and corrects for biases in the meteorological forcing and streamflow forecasts. Inputs are processed by a preprocessor and include temperature and precipitation forecasts from the Global Ensemble Forecast System (GEFS) and the Climate Forecast System (CFSv2 from NCEP) among others. The space-time covariability of precipitation and temperature is modelled with the Schaake Shuffle method. The hydrologic uncertainties and biases are modelled statistically. The HEFS is implemented within the Community Hydrologic Prediction System (CHPS), which allows for operational forecasting and hindcasting. CHPS is an open service-oriented architecture built on the Delft-FEWS framework. Verification is conducted with the Ensemble Verification System. The HEFS is being implemented in several phases; having been tested at several NWS River Forecast Centers (RFCs), the system is being rolled out across all RFCs for operational use in 2015–2016</p>	Demargne et al. 2014a, b; Schaake et al. 2007; Wu et al. 2011; Clark et al. 2004; Roe et al. 2010; Werner et al. 2013; Brown et al. 2010; Emerton et al. 2015

S3	PER	Forecast center/System name: Pilot EPS Rijnland	The catchment of the water system of the Rijnland Water Board is a plain region about 30 by 35 km in the Western part of the Netherlands. Except for the dune area along the North Sea coast, most of the catchment consists of reclaimed areas, several metres below sea level. Drainage is done through a system of drains, ditches, and bigger canals by means of pumping stations to discharge the excess water to the river Rhine or North Sea. The main canals are interconnected and therefore also function as a storage reservoir. Therefore, a medium-range 1–10 day ensemble forecasting system has been set-up to predict water levels in the reservoir. The AQUARIUS modelling system uses a spatially lumped hydrologic model and the results are postprocessed. The alert levels and decision and control rules for anticipatory pumping have been optimized, minimizing flood damage costs while keeping adverse damage costs of false alarms at acceptable level, based on a 7.5 year hindcasting exercise. The forecasting system has been operational since January 2011 for testing by Rijnland operators alongside their formal operational routines. The system is currently hosted at UNESCO-IHE and accessible to the Water Board through a website	Van Andel et al. 2014
S4	FEWSOO/ER	Forecast center/System name: Flood Early Warning System for the Po River and the Emilia Romagna Region	The Flood Early Warning System for the Po River (FEWSPo) has been developed at ARPA Emilia Romagna since 2005, in close collaboration with other Italian agencies. The FEWSPo system is an operational tool for flood and drought planning, forecasting and emergency management. Flood forecasts, up to 5 days in advance, are given through a multimodel ensemble approach combining several meteorological inputs (COSMO12, COSMO17, COSMO-LEPS, observed precipitation and temperature) with hydrological-hydraulic models (HEC-HMS/RAS, MIKE11 NAM/HD, Topkapı/Sobek) and observed discharges/water levels. Drought forecasts are also given, up to 3 months in advance, by combining meteorological inputs (Deterministic ECMWF GCM 0–6 days, ECMWF 7–14days VARREPS, ECMWF monthly forecast, seasonal downscaled forecast, observed precipitations and temperature)	(continued)

Table 1 (continued)

ID	System acronym	Forecast center/System name:	System overview	References
S5	WPI RAS	Forecast center/System name: Water Problems Institute of Russian Academy of Sciences Status: Operational Provider: Water Problems Institute of Russian Academy of Sciences Domain: Catchments of the Vyatka River, Sosna River, and Seim River (European part of Russia)	A forecast system for long-term (with a lead-time of 2–4 months) ensemble forecasting of snowmelt flood hydrographs has been developed at the WPI RAS. The system is based on the use of a physically based distributed model of runoff generation. The model is used to simulate initial river basin states (e.g., soil moisture and depth of frozen soil) and predict runoff hydrographs. Uncertainty of weather scenarios is accounted for using a stochastic weather generator, and an ensemble of weather sequences generates an ensemble forecast of daily streamflow discharges	Kuchment and Gelfan 2007, 2009; Kuchment et al. 1986; Kuchment and Gelfan 2011
S6	SSF	Forecast center/System name: Seasonal Streamflow Forecast Status: Operational Provider: Bureau of Meteorology Domain: Australia	Every month, the service delivers 3 months ahead probabilistic forecasts of total streamflow volume at gauging locations or total inflows into major water storages (http://www.bom.gov.au/water/ssf). The current service uses a statistical model relating the observed streamflow in the previous month, along with several climate indicators to future streamflow. A modelling system known as the Water Availability Forecasts of Australian Rivers (WAFAri) is used by the Bureau of Meteorology to allow an entire workflow for the streamflow forecasts, including data management to a central database, forecast generation, and web publication. Forecasts based on the dynamical approach currently exist in an experimental phase. The dynamical rainfall-runoff model is driven by seasonal rainfall forecasts from a global climate model, downscaled to provide catchment scale rainfall	Shin et al. 2011; Wang et al. 2009a; Wang and Robertson 2011

S7	SPC-LCI	Forecast center/System name: Loire and Allier Forecasting System Status: Operational Provider: Service de Prévision des Crues Loire-Cher-Indre and Service de Prévision des Crues Allier Domain: The Loire and Allier catchment	In order to issue forecasts (and warnings), the hydrological prediction systems developed in the regional services in France may include deterministic meteorological predictions as flood guidance (daily expert QPFs), simple correlation models or charts linking water level observed before the event, the (observed and/or forecasted) accumulated precipitation during the event are also used to issue warnings operationally where hydrological and/or hydraulics models (GRP, GR4H, SCS) are unable to correctly represent the evolution of water levels and/or discharges. The regional services covering the Allier (SPC Allier) and the Loire (SPCLCI) catchments use, as input, probabilistic QPFs provided by an alternative system based on analogue sorting approach (RainFAST). Hydrological models are set up for small to medium upper basins (up to 2,000 km ²). Several meteorological scenarios may be used as input. However, the hydrological forecasting expert selects only one scenario to propagate through the hydraulic models (1D models: HYDRA, MIKE and/or the statistical models. In order to take into account the downward flood propagation, simple reservoir operation models are used to anticipate reservoir's outflow to the main channel and tributaries	SPC LCI 2013; SPC-Allier 2013; Marty et al. 2012; DHI 2007; Maden and Skotner 2005
S8	FEWS Scot	Forecast center/System name: FEWS Scotland Status: Operational Provider: Scottish Flood Forecasting Service Domain: Scotland	Scotland's current flood forecasting system, 'FEWS Scotland', employs the Delft-FEWS platform. This brings together meteorological (radar and NWP forecasts) and hydrological (rain gauges and river levels) information to drive hydrological and hydrodynamic models. One of the key developments in 2011 was the introduction of a countrywide distributed hydrological model Grid-to-Grid (G2G). G2G principally runs in deterministic mode and employs radar and rain gauge estimates of rainfall together with weather model predictions to produce forecast river flows, as gridded time-series at a resolution of 1 km and for up to 5 days ahead. G2G is also run operationally using 24 ensemble predictions of rainfall from MOGREPS to provide probabilistic flood forecasts	Werner et al. 2004; Cranston et al. 2011; Maxey et al. 2012

(continued)

Table 1 (continued)

ID	System acronym	Forecast center/System name:	System overview	References
S9	IFKIS	name: IFKIS-Sihl/IFKIS-Ticino Status: Operational Provider: WSL	<p>The real-time system used at WSL was originally implemented in several basins in spring 2007 to participate to the MAP-DPHASE project. The PREYAH hydrological model is initialized with weather radar QPE (Quantitative Precipitation Estimates) and forced by various NWP. The system was designed to demonstrate the added value of HEPS in basins with areas smaller than 2,000 km², as well as the advantage of HEPS over deterministic forecasts</p> <p>The Sihl River modelling system is designed to avoid severe flooding in Zürich and uses novel ways to display persistence maps and time-lagged forecasts. The Ticino River system connects weather radar ensembles to operational hydrological models and covers basins ranging from 40 to 6,600 km²</p>	Zappa et al. 2008; Viviroli et al. 2009; Germann et al. 2009; Zappa et al. 2008; Addor et al. 2011
S10	WAVOS	Forecast center/System name: WAVOS, FEWS (combination of two forecast systems): WAVOS focuses on hydrodynamics, FEWS focuses on hydrology. Status: Operational Provider: German Federal Institute of Hydrology (BfG) Domain: Germany, German Federal Waterways	<p>The core of BfG's forecasting system WAVOS for the German Federal Waterways is a 1D hydrodynamic model calculating water levels at gauges relevant for navigation. WAVOS acts as an operational environment handling the data feed to the models, scheduling model runs as well as exporting, processing, and disseminating model results. WAVOS is operated by 7 forecasting centers (including BfG itself) for the 5 German waterways Rhine, Elbe, Danube, Main, and Odra. As future water-levels in the major rivers highly depend on discharges generated in the corresponding river basin, WAVOS is able to import discharge forecasts from external sources. These forecasts are generated by other forecasting centers or by BfG's hydrological forecasting models running in the FEWS environment. Within BfG's forecasting system, FEWS acts as the interface to the meteorological input. Lead-times range between 1 and 10 days, as a function of the gauging stations and the users' needs. The forecasts are published via the Internet on the German waterway information service</p>	Meißner and Rademacher 2010; Meißner et al. 2012; Werner et al. 2013; Renner et al. 2009; https://www.elwiss.de

S11	NEWS	Forecast center/System name: Novel Flood Warning and Risk Assessment System Status: Experimental Provider: NEWS Domain: Upper Huai, China	This system uses ensemble precipitation forecasts from the Thorpx Interactive Grand Global Ensemble (TIGGE) archive for flood warning on the large-scale Upper Huai catchment ($30,672 \text{ km}^2$) located in east-central China. It was jointly developed by scientists at King's College London, ECMWF, Hohai University in China, and the user organization, Anhui Province Bureau of Hydrology. The Xinanjiang model is implemented to produce ensemble of river discharge forecasts. It demonstrated to be successful in the hindcast of the 2008 summer flood events in the catchment. Further development of the system is ongoing with the Anhui Province Bureau of Hydrology and Huai River Basin Meteorological Centre	He et al. 2009
S12	PREDICTOR	Forecast center/System name: PREDICTOR Status: Operational Provider: EDF Domain: several catchments in France	The development of PREDICTOR, the EDF tool to help in producing probabilistic streamflow forecasts, started in 2008. The tool was fully developed in-house by EDF. Operational daily forecasts started at the end of 2010. PREDICTOR uses different meteorological inputs, both deterministic and probabilistic (ARPEGE, AROME, High. Res. ECMWF, EPS-ECMWF, analogue-based forecasts) to force a conceptual hydrological model (MORDOR). PREDICTOR is unusual in its consideration of both meteorological and hydrological uncertainties by pre/postprocessing, while integrating forecaster expertise (forecasters can modify forecasted distributions of mean daily rainfall, mean daily air temperature and streamflow)	Desaint et al. 2009; https://hepex.irstea.fr/operational-use-of-ensemble-hydrometeorological-forecasts-at-edf-french-producer-of-energy
S13	RWSOS Rivers	Forecast center/System name: RWSOS Rivers Status: Operational Provider: Rijkswaterstaat Domain: Rhine and Meuse catchments	RWSOS Rivers is used by the Rijkswaterstaat Water Management Center of the Netherlands (WMCN) and Rijkswaterstaat Regional Directories to produce flood forecasts for locations along the Rhine and Meuse Rivers within the Netherlands. The system infrastructure comprises an implementation of the Delft-FEWS system that includes streamflow generation and streamflow routing models that cover the entire Rhine and Meuse catchments. Relevant meteorological and hydrological data, both observations and forecasts, are managed by the Delft-FEWS system. FEWS also includes forecaster workflows for running models and disseminating results. The system includes	Werner et al. 2013; Lindström et al. 1997

(continued)

Table 1 (continued)

ID	System acronym	System overview	References
S14	GloFAS	<p>Forecast center/system name: Global Flood Awareness System</p> <p>Status: Preoperational</p> <p>Provider: JRC/ECMWF</p> <p>Domain: Global</p> <p>HBV90-based hydrologic models and SOBEK-based hydrodynamic models as well as some additional routines for data assimilation and postprocessing of forecasts. The forecasting system is used to forecast streamflow rates at Lobith (Rhine) and St Pieter (Meuse) by WMCN; regional directorates use the system to then predict water levels along these rivers</p> <p>GloFAS is a combined effort by the Joint Research Centre (JRC) of the European Commission and the European Centre for Medium-Range Weather Forecast (ECMWF). GloFAS is based on a similar technology to that of the European Flood Awareness System (see above, S1, EFAS), except that it has a global domain. It uses the ECMWF Ensemble weather forecasts and ERA-Interim reanalysis as main inputs. Every day, 30-day streamflow forecasts are computed and compared to warning thresholds to detect upcoming floods. Results are disseminated on a dedicated website (www.globalfloods.eu) via warning maps based on three threshold levels (medium, high, and severe), with additional information displayed at selected dynamic reporting points</p> <p>The system has already demonstrated its potential in predicting recent catastrophic floods. Alfieri et al. (2013) showed that current ensemble streamflow predictions can enable skilful detection of hazardous events with forecast horizons as long as 1 month in large river basins</p>	<p>Dee et al. 2011; Alfieri et al. 2013; Emerton et al. 2015</p>

S15	AFFS	Forecast center/System name: African Flood Forecasting System)	The African Flood Forecasting System (AFFS) aims at producing daily probabilistic flood forecast information for the whole of Africa, i.e., for all medium- to large-size African river basins, with a maximum lead time of 15 days. The key components are the hydrological model LISFLOOD, the African GIS database, the meteorological ensemble predictions of the ECMWF (ECMWF-ENS), and the already well-established flood forecast methodology used within the European Flood Awareness System (see above, S1, EFAS). LISFLOOD was set up on the pan-African scale with a spatial resolution of 0.1° and optimized for African conditions (including modifications of the hydrological processes modelled, and parameter optimization). Results are visualized in maps showing the exceedance of critical hydrological thresholds, as well as ensemble quantile plots at key locations	Burek et al. 2013; Thiemig et al. 2014
		Status: Experimental		
		Provider: JRC	First evaluations of the forecasting performance of AFFS demonstrated that AFFS is capable of detecting 69 % of the reported flood events correctly. The system showed a particular strength in predicting riverine flood events of long duration (>1 week) and large affected areas ($>10,000 \text{ km}^2$). Additionally, AFFS proved capable of predicting flooding in ungauged river basins, which represents a major added value for regions in Africa where data networks are sparse and data for calibration and monitoring are often not available	

(continued)

Table 1 (continued)

ID	System acronym	Forecast center/System name:	System overview	References
S16	AIGA	AIGA-Ensemble Status: Operational for the deterministic system Testing-phase for the ensemble system Provider: Irstea/Météo-France Domain: Southern France, to be extended to mainland France First ensemble tests for the Meuse (9000 km ²) and Moselle river basins (11,500 km ²)	To support flash flood warnings from small-to-medium ungauged catchments the operational single-valued warning system called AIGA includes, a simplified distributed hydrologic model run every 15 min at a 1-km ² resolution using the Météo-France radar-gauge rainfall grids as forcing inputs. AIGA produces peak discharge estimates along the river network, which are then compared to regionalized flood frequency estimates. The output product is a map of the river network with a color chart indicating the range of the AIGA-estimated return period of ongoing events. The single-valued AIGA system is operational in the Southern part of France since 2005. High-resolution gridded precipitation forecasts will include Météo-France's Application of Research to Operations of Mesoscale ensembles at a 2.5-km resolution (AROME) and operational precipitation ensembles from the multimodel COSMO-DE-EPS convection-permitting model run at the Deutscher Wetterdienst (DWD) (2.8 km resolution)	Javelle et al. 2010, 2014; Arnaud and Lavabre 2002; Gebhardt et al. 2011; Seity et al. 2011

S17	CHROME	Forecast center/System name: CHROME	The CHROME development project aims at building an operational hydrometeorological ensemble forecasting chain to produce Flood Vigilance Warnings and quantitative flow forecasts with meaningful anticipation over small size gauged catchments in France. By ingesting the AROME hourly gridded quantitative precipitation forecasts (QPF), the CHROME forecasting system produces hourly discharges with a forecast range up to 30 h. The specificity of the system is that it takes into account both uncertainties from precipitation forcing and hydrological modelling. To account for the hydrological uncertainty, a multithydrological model approach is carried out through the implementation of four different distributed rainfall-runoff models. To deal with the meteorological uncertainty, an AROME QPF perturbation method is used to produce 50 gridded rainfall scenarios. These precipitation ensembles are then ingested by the hydrological models providing 50 30-h forecast discharges. In addition, a time-lagged ensemble forecast formed by merging the results of the successive AROME operational runs (4 runs a day, every 6 h) supplement the system's outputs by considering the consistency of the meteorological forcings	Seity et al. 2010; Vincendon et al. 2011
		Status: Preoperational		
		Provider: SCHAPI and Météo-France		
		Domain: Gardon d'Anduze River (545 km ²), Ardeche River (2000 km ²), Céze River at Bagnols-sur-Cèze (1600 km ²)		

(continued)

Table 1 (continued)

ID	System acronym	Forecast center/System name:	System overview	References
S18	MMEFS	Forecast center/System name: Meteorological Model-based Ensemble Forecast System Status: Operational Provider: NOAA/NWS Domain: Most of the eastern US: four NWS River Forecast Center domains: Northeast, Southeast, Mid-Atlantic, Ohio River	MMEFS became operational in 2012 after several years of experimental real-time operations, and now provides 1–7 day lead ensemble forecasts at all forecast locations of the participating RFCs. In contrast to OHD-led HEFS efforts, MMEFS development was led by several RFCs. The MMEFS uses temperature and precipitation ensemble forecast outputs from the NOAA National Centers for Environmental Prediction (NCEP) Global Ensemble Forecast System (GEFS), the Short Range Ensemble Forecasts (SREF), and the North American Ensemble Forecast System (NAEFS). The raw model gridded forecasts are interpolated to catchment model centroids for direct input to the NWS Community Hydrologic Prediction System (CHPS), and output graphics are produced using the free R statistical software. The system runs automatically, without forecaster intervention, with subdaily updates.	http://www.erh.noaa.gov/mmeis/

S19	LARSIM Moselle and Rhineland- Palatinate	Forecast center/System name: LARSIM Moselle and Rhineland-Palatinate Status: Operational for the deterministic system Preoperational for the ensemble system Provider: Landesamt für Umwelt, Wasserwirtschaft und Gewerbeaufsicht Rhineland- Palatinate (Germany) Domain: Moselle Basin (France, Luxembourg, Germany) and federal state of Rhineland- Palatinate (Germany)	LARSIM forecast systems for the Moselle Basin and for the federal state of Rhineland-Palatinate have been operational for flood forecasting since 2008. The hydrological model is forced by the numerical weather forecasts of the German Weather Service with a lead time of up to 1 week. The hydrological model for the Moselle basin can additionally be driven by forecasts delivered by Météo-France (Arième and Arpège). The application of a 20 member ensemble precipitation forecast with a lead time of 27 h (COSMO-EPS, German Weather Service) is implemented and runs daily in preoperational mode. These forecasts are fed into the water balance model LARSIM. The LARSIM model for the Moselle basin (28,300 km ²) is a raster model with a spatial resolution of 1 × 1 km grid size, while the model for the remaining area of Rhineland-Palatinate (25,070 km ²) is covered by a model partitioned into subbasins. Both models are operated at an hourly temporal resolution
-----	---	---	---

(continued)

Table 1 (continued)

ID	System acronym	Forecast center/System name:	System overview	References
S20	WSFS	Watershed Simulation and Forecasting System WSFS	<p>The WSFS (Watershed Simulation and Forecasting System) is widely used in Finland for real time hydrological simulation and forecasting. The distribution of the model is based on a watershed division with 60–100 km² subbasins. In the research version, a 1 × 1 km² grid is used. Remote sensing data used in the system are satellite data of snow cover extent and water equivalent and precipitation from weather radars. Assimilation methods for the use of satellite snow data in WSFS have been developed for operational use. Assimilation methods of flood area and soil moisture (SMOS) data from satellites are under development. Weather radar has been in operational use since 1998 and is still under development to increase the accuracy of areal precipitation estimates. The automatic model updating system developed is an important part of the WSFS. Model state updating is done against water level, discharge, snow line, and satellite snow observations. In operation, WSFS automatically collects meteorological and hydrological data, runs hydrological forecasts, and distributes forecasts onto the Internet www.environment.fi/waterforecast.</p>	Vehviläinen 1994

S21	HWN	Forecast center/System name: Hydrological warning system for Norway Status: Operational Provider: Norwegian Water Resources and Energy Directorate, Hydrology Department Domain: Continental Norway	<p>This system has been operational since 2000. Conceptual rainfall-runoff models (the HBV-model) for 147 catchments, from 6 to 15,500 km², provide daily output data that are postprocessed to quantify uncertainty. The uncertainty quantification system was developed in collaboration with the Norwegian Computing Centre, and accounts for both uncertainty due to the weather forecast and due to the model. Forecasts with quantified uncertainty are updated four times a day, i.e., when new meteorological forecasts are available.</p> <p>Meteorological forecasts are provided by AROME for the first 2 days and ECMWF for the following 7 days. An alternative daily model system of DDD-models contributes to a small model ensemble.</p> <p>Recently, the daily system was supplemented by a 3-hourly DDD-model system that improves small-scale forecasting. So far, no quantification of uncertainty is connected to the DDD models. The results from the catchment model simulations are the basis for flood forecasts that are published on www.varsom.no, up to 3 days in advance</p>	<p>Sæthun 1996; Follestad and Hest 1998; Langstrud et al. 1998a, b; Skauen and Onof 2013</p>
S22	HUGO	Forecast center/System name: HUGO Status: Operational Provider: Bayerisches Landesamt für Umwelt Domain: Bavaria	<p>The forecast system HUGO represents an integrative interface for operating various hydrological models throughout Bavaria. The objectives of HUGO are to provide flood forecasting and early warnings for regional public authorities (water management offices, civil protection); forecast discharge and river levels are then passed on to the flood warning system of Bavaria. It links ~30 different regional model configurations, covering large parts of the Upper Danube Basin and the Main Basin. Depending on the meteorological and hydrological situation, model runs are executed between once a day and up to an hour or less of lead time (either automated setup or manual simulation by experienced specialists). Regional and local public authorities are informed about the predictions and may access the simulated results for issuing regionalised warnings. In order to provide a high reliability of the flood predictions, ensemble weather forecasts can be fed into the hydrological models</p>	Holle 2009

(continued)

Table 1 (continued)

ID	System acronym	Forecast center/system name:	System overview	References
S23	SESP	Forecast center/system name: AquaLog/Hydrog SESP (short-term ensemble prediction) Status: Operational Provider: Czech Hydrometeorological Institute Domain: Czech Republic	The AquaLog/Hydrog LAEF forecasting system has been developed at the Czech Hydrometeorological Institute. The system was designed to complement the deterministic AquaLog/Hydrog runoff forecast with some indication of probability of exceedance of flood risk thresholds on Czech rivers. The QPF uncertainty has been identified as a dominant source of uncertainty in the Czech Republic, and therefore the system is limited to the use of input QPF ensemble from NW P ALADIN-LAEF operated by ZAMG. Model outputs are daily forecasts of discharge up to 48 days in advance	Wang et al. 2009
S24	MESP	Forecast center/system name: AquaLog MESP (monthly ensembled streamflow prediction) Status: Operational Provider: Czech Hydrometeorological Institute Domain: Czech Republic	The AquaLog MESP forecasting system has been developed at the Czech Hydrometeorological Institute. The system is based on the generated ensemble of 1,000 years of precipitation and temperature, from the historical observation and expected above/below/normal monthly climate outlook. 45 ensemble members are selected and distributed in space and time in order to input hydrological simulation. Initial hydrological conditions are based on short-term deterministic runs of AquaLog system outputs and are daily discharge forecasts up to 30 days in advance	

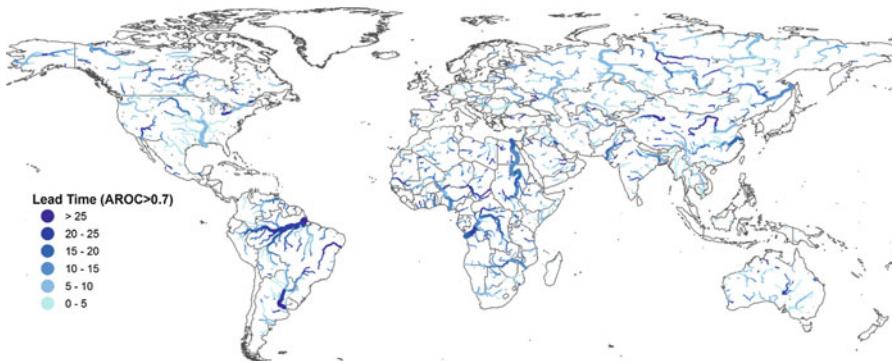


Fig. 2 Performance of the Global Flood Awareness System in the main world rivers for the period 2009–2010. Shades of *blue* indicate the maximum lead time of skilful ensemble streamflow predictions in early flood detection, i.e., ROC area > 0.7 (Adapted from Alfieri et al. 2013)

Each system is optimized for a specific range of basin sizes, climate conditions, and hydrological processes, and may use varying levels of local data. This indicates an incredible wealth of experience in conducting research, designing and implementing operational systems, and operating them on a real-time basis, 24 h a day, 7 days a week. Most systems described are operational, although some are in a preoperational state (defined as a system which is run regularly and has some degree of monitoring, but is not fully used operationally, as stated by the system's provider). On average, 2 years were needed for the operational implementation of a system, with a preceding period of 5–10 years of research. This time-lapse may seem to be extremely short, but it recognizes that many of the systems presented in Table 1 were built on previous operational efforts that were already directing them to become a HEPS. Note that only half of the systems in Table 1 were operational before 2012 (the earliest ones started in 2006), illustrating that operational hydrological ensemble forecasting is a rapidly growing, but still nascent field. In the following section, the main features of the HEPS presented in Table 1 are discussed e.g., target variables, objectives, visualization. This chapter concludes with an outlook on the priorities for further operationalizing HEPS around the globe.

2 Systems' Descriptions

2.1 Target Variables

Table 1 shows that HEPS usually forecast a large number of variables including soil moisture, snow cover, and snow water equivalent, with all of the systems described in this chapter predicting discharge or stage as part of their set-up. Sometimes these are expressed in real-world units (e.g., river depth in metres) or as return periods (e.g., chance of experiencing a 10-year flood). Discharges or river water levels may not always be the target variable as exemplified by S3, which predicts inflows to and

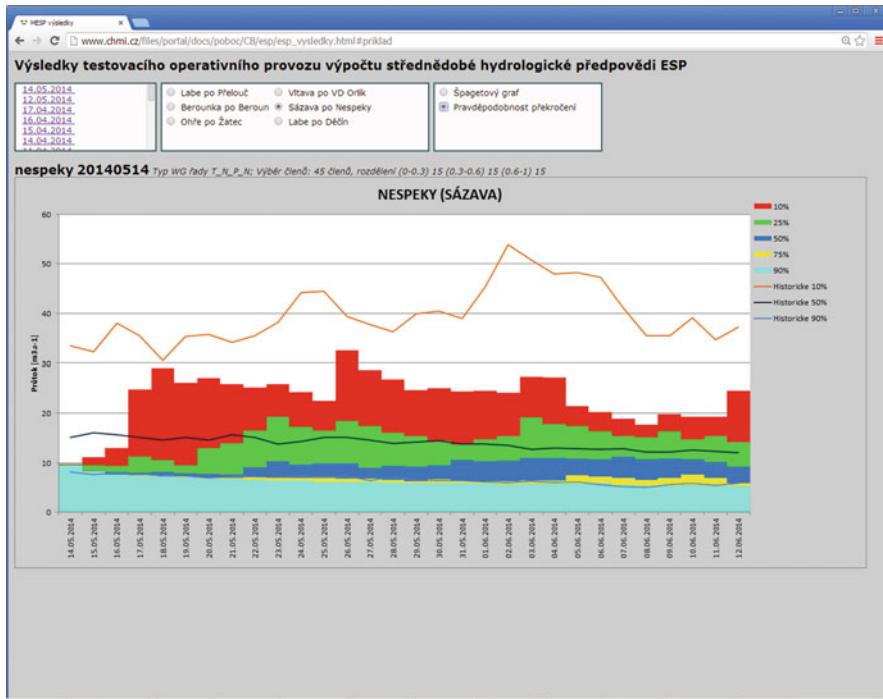


Fig. 3 Example of output of forecast system S23 providing monthly forecasts for reservoir operators

water levels in reservoir systems (see Fig. 4). All of the forecasting systems in Table 1 which have a civil protection focus tend to set the discharge/water level forecasts in the context of risk and use the return period to convey the disaster magnitude (e.g., in S1, S8, and S16). S1 also takes the explicit embedding in a risk framework further and predicts affected population and potential monetary damage on a district level.

2.2 Systems' Objectives

Some systems focus solely on floods (e.g., S1, S2, S3, S4, S5, S7, S11, S13, S15, S16, S17, S19, S20, S21, S22, S23), which requires stream flow predictions. These can furthermore be used for water resource forecasting (e.g., in S2, S4, S6, S8, S23, and S9). This may include low flows or other applications such as navigation (S10) or forecasting systems specifically targeted to reservoir operators (S23; see Fig. 3). It is not uncommon for operational forecasts to have a multiobjective strategy, thus

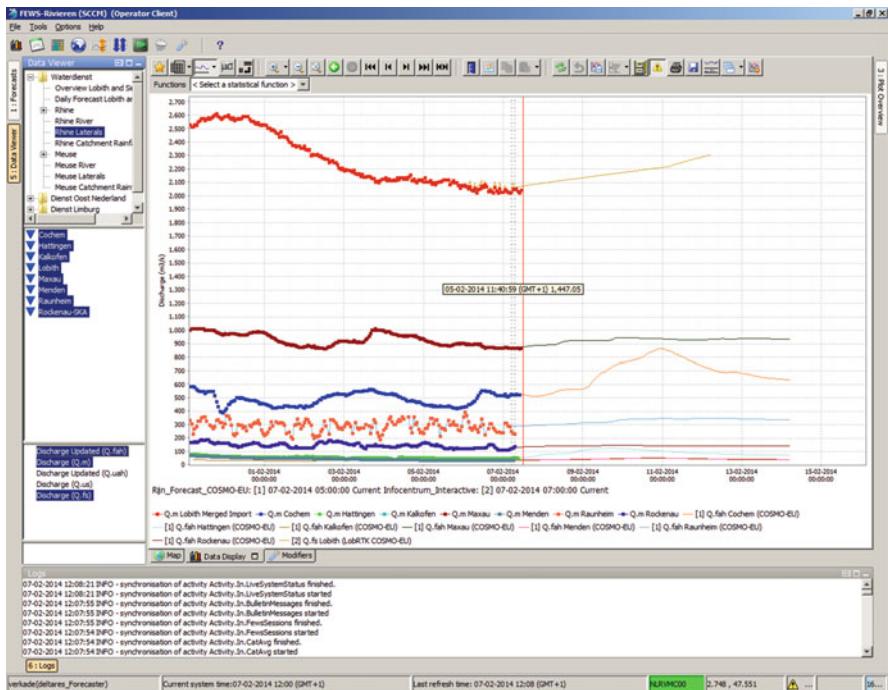


Fig. 4 Screenshot of the FEWS forecasting platform showing single-value forecast results from the RWsOS Rivers system (S13)

having different products or systems for particular applications (in most cases, it is fairly straight forward to use outputs in the form of discharge or water level for other applications). For example, S13 has a distinct multisystem strategy, which includes a fluvial flood forecasting system with other parallel systems in use for low flow forecasting (RWsOS Water Management), for flood forecasting for the “great lakes” IJsselmeer and Markermeer (Name of forecast system: RWsOS Lakes), and for the coastal zone (Name of forecast system RWsOS Noordzee). All of S13s systems are based on the same underlying technology (FEWS, see Fig. 4). Other systems provide the flexibility to develop specific products for particular applications (e.g., S2), as well as providing end-users with the “raw” ensemble forecasts that can be integrated into their own systems (e.g., for decision support).

In contrast, S12 (PREDICTOR) has embraced a multipurpose single system strategy. The reasons for hydrometeorological forecasting by S12 are listed by Desaint et al. (2009) or Le Lay et al. (2013) as follows: (i) monitoring hydrological risk (90 flow thresholds are monitored at more than 50 points in 31 watercourses); (ii) short-term (1–14 days) flow forecasting on 115 points at more than 50 watercourses; (iii) monitoring the risk of extreme events, storms in the French Southwest, strong winds and heavy snow throughout France; (iv) low-flow forecasting

Table 2 Forecast horizons

Horizon	Systems (see Table 1 for descriptions)
Sub daily	S2, S4, S7, S8, S9, S13, S16, S17, S19, S20, S22, S23
1–5 days	S1, S2, S3, S4, S7, S8, S9, S10, S11, S12, S13, S14, S15, S16, S17, S19, S20, S21, S22, S23
5–30 days	S1, S2, S3, S10, S11, S12, S14, S15, S19, S20, S21, S24
Seasonal	S2, S6, S10, S14, S20, S21
Annual	S2, S20

(lead times of weeks to months) on a few large French river basins; (v) the prediction of long-term (months ahead) inflows to reservoirs located at upstream valley areas; (vi) short-term (1–10 days) water temperature forecasting in six major rivers; and, more recently, (vii) sediment transport prediction. The full operational system of S12comprises two forecasting centers located in the cities of Grenoble and Toulouse.

There is no correct or incorrect approach to constructing a multiobjective forecasting system, as it depends on various circumstances including the role of the organization operating the HEPS system.

2.3 Forecast Horizon

The operational systems in Table 1 cover a range of different forecast horizons (Table 2), guided mainly by the intended application. The forecast horizon from 1 to 5 days is clearly the one which is covered most, largely because it represents the overlap between short-range and medium-range forecasting systems. There are only two systems (S2 and S20) which have an annual forecast horizon; these are also the only systems which predict across all forecast horizons from hourly, through the medium range to the month and seasonal time range.

2.4 Hydrological Models and Forcings

At the core of most operational forecasting systems is a hydrological model, which has to be forced by boundary conditions to create a forecast. The exception is S6 where the approach does not use a rainfall-runoff model (but instead uses a statistical Bayesian Joint Probability approach; Wang et al. 2009a; Wang and Robertson 2011). Otherwise, there is a large variety of different hydrological models used, of which about half are spatially lumped and half are spatially distributed. Models include: Lisflood (van der Knijff et al. 2010), HTESSEL (Balsamo et al. 2009), HBV (Bergström 1995; Lindström et al. 1997), MORDOR (Garçon 1999), Sacramento

Soil Moisture Accounting (Burnash 1995), G2G (Cranston et al. 2011), GRP (Berthet et al. 2009), LARSIM (Ludwig and Bremicker 2006; Bremicker et al. 2013), and DDD (Beldring et al. 2003). Sometimes the hydrological models do not include any substantial river routing and need to be coupled to hydrodynamic models (e.g., S10 and S13 use SOBEK and S14 uses Lisflood-FP; see e.g., Neal et al. 2011).

Most centers use only one hydrological model, with the exception of S4, S7, S14, S16, and S17 which employ a combination of models HEC-HMS and HEC-RAS – MIKE11-NAM, Mike11-HD, Topkapi/Sobek in the case of S4 (Casici et al. 2011); HTESSEL/LISFLOOD in the case of S14; and GR4J (Perrin et al. 2003)/GR-SD (Javelle et al. 2010, 2014) in the case of S7 and S16. S17 employs four hydrological models: MARINE (Roux et al. 2011), SCS-LR (Coustau et al. 2012), ALHTAIR (Bressand 2002), and ISBATOP (Vincendon et al. 2010). The scientific debate on the tangible advantages and disadvantages of multimodel approaches in hydrology is still ongoing (Kauffeldt et al. 2014; Anctil et al. 2014; Pappenberger 2014).

The variety of models reflects the large number of hydrological models which are suitable for operational settings (for a review in Europe, see Kauffeldt et al. 2014), as well as the need for detailed representation of physical processes in certain catchments. Therefore, large continental forecasting systems such as S2 employ different models in different regions. Although S2 primarily uses the Sacramento Soil Moisture Accounting Model (SAC-SMA; Burnash 1995), together with the Snow Accumulation and Ablation Model (Snow-17; Anderson 1973), some subregions such as the Middle-Atlantic River Forecasting Center use empirical models, based on the Antecedent Precipitation Index, but adapted for continuous simulations (Sittner et al. 1969).

Interestingly, very few forecasting systems rely on forcing (e.g., weather) forecasts from just one source. At a minimum, all operational systems produce uncertainties either through a statistical methodology or by using a single ensemble system. For example, S6 uses total streamflow from the previous month, and S3 and S14 use ECMWF ensemble weather forecasts only. All other systems use or actively experiment with multiforcing systems, which means using forecasts from different meteorological modelling systems or centers. S11 is extreme in the sense that it uses 9 TIGGE ensembles totalling over 300 individual weather forecasts. S10 is a typical configuration using: (1) measured water-levels and discharges, (2) measured temperature and precipitation, (3) COSMO-EU + GME weather forecasts (German Weather Service), (4) COSMO-LEPS weather forecasts (ARPA-SIM), and (5) ECMWF high-resolution and ensemble weather forecasts. In different regions such set-ups are influenced by the locally available models and their higher quality. For example, S8 is a variant of S10 and uses: UKV deterministic (1.5 km to 36 h), Euro4 (4 km to 120 h), MOGREPS-UK (upscaled 2.2 km to 36 h), and MOGREPS-Global (downscaled 33 km for following 18 h) (for details see references for S5). The latter already demonstrates a range of forcings with different lead times, similar to S2, which is based on single-valued operational precipitation and temperature forecasts from the

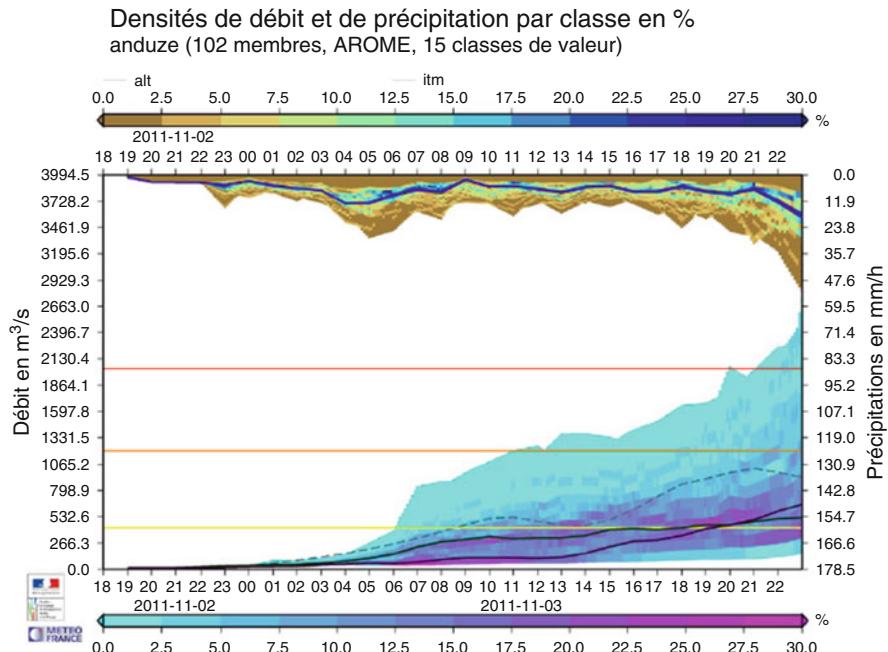


Fig. 5 Density of spaghetti graph of the lagged ensemble discharge and precipitation forecasts on the Gardon River at Anduze, SCS-LR and ISBATOP models at 18H UTC on the 02/11/2011. Observation is represented by the *dashed line*. The *thick lines* are the different forecasts given by SCS-LR (green) and ISBATOP (purple) forced by the operational AROME QPF (From S17)

NWS RFCs (~1–5 days), the Global Ensemble Forecast System (1–16 days), the Climate Forecast System (16 days–9 months) and climatological forcing (9 months–1 year, and all forecast horizons as a baseline). There is a large variety of other combinations used. S4, for example, complements its existing forcings (Cosmo N2-RUC, COSMO I2 COSMO I7, COSMO LEPS) with statistical analogs (as does S12). Furthermore, S4 uses a downscaling approach (weather generator) to create statistical ensembles from low-resolution seasonal forecasts for drought prediction (Agnetti et al. 2011). If such techniques are combined with a lagged ensemble system, then it is possible for 100–300 ensemble members to be generated to quantify forecast uncertainty due to the atmospheric forcing (e.g., S17, see Fig. 5).

2.5 Uncertainties Considered

Uncertainties in the systems arise because of the meteorological forcing, hydrological models, and the decision systems involved. Overall, the multimodelling

philosophy seems to be limited to using forcings from multiple weather models, rather than the use of many hydrological models or parameters, with the exception of S21. In addition, the dominant source of uncertainty in many of the forecasting systems presented here is meteorological uncertainty, although this is not universally true (see e.g., Fundel and Zappa 2011; Fundel et al. 2013) as is recognized in S20 which also integrates uncertainty in snow and soil moisture initial conditions.

As shown in the previous section, many operational systems use ensembles to quantify meteorological forcing uncertainty. Many statistical methods, such as S6, have uncertainties as part of their system design. Other methods, more physically based model cascades need to employ techniques which allow them to incorporate all of the uncertainties. The relative magnitudes of the various sources of uncertainty are catchment- and application-specific and require a detailed analysis (e.g., S9). Therefore, the type of uncertainty quantification employed in operational forecasting systems varies widely. However, it is not uncommon that several sources of uncertainty and bias are considered in a lumped way through pre- or postprocessing. For example, in S2 the total and predictive uncertainty in the future streamflow are separated into contributions from the meteorological forcing and from the hydrologic modelling. The Meteorological Ensemble Forecast Processor (MEFP) accounts for the uncertainty in precipitation and temperature forcing from raw model forecasts, while correcting for biases (Schaake et al. 2007; Wu et al. 2011). Raw streamflow forecasts are generated from the

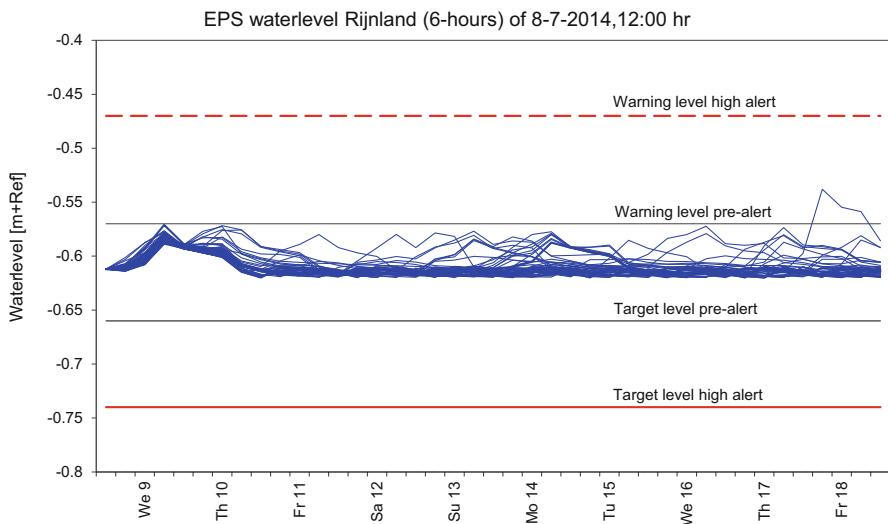


Fig. 6 Ensemble water level forecast from the operational Pilot EPS Rijnland for flood early warning and reservoir control. The *horizontal lines* indicate what would be the optimized warning levels and corresponding target levels based on a 7.5 year re-forecast exercise (Van Andel et al. 2014)

bias-corrected meteorological forecasts. The residual (hydrologic) uncertainties and biases are modelled with the Ensemble Post-processor (Seo et al. 2006). Other hydrological postprocessors used by operational systems include: VAREX (S1, Bogner and Pappenberger 2011), quantile regression (S7 and S13, Weerts et al. 2011), BMA or EMOS techniques (S10 and S1), Model Conditional Processor (S4), and ensemble dressing techniques (S12, Zalachori et al. 2012; Chardon et al. 2014), amongst many others.

The above methods of postprocessing have mainly been based on discharge. However, some systems also approach this issue from a decision-making level. For example, for S3, optimal alert levels (Fig. 6) and control strategy (corresponding target levels) have been determined (to minimize flood damage costs while keeping adverse damage costs of false alarms at an acceptable level) based on a 7.5 year reforecast exercise (Van Andel et al. 2014).

2.6 Users

The users' requirements and needs shape the character of all operational hydrological forecasting systems. The current users include hydrological forecasters, civil protection, water managers and users, irrigators, urban and rural water supply authorities, environmental managers, humanitarian aid organizations and

www.environment.fi

Hydrological Forecasts and Maps | Finnish Environment Institute (SYKE)

Hydrological forecasts: Kemijoki watershed - Ounasjoki Kittilä

Term translation

Hydrological forecasts > Kemijoki > Total catchment area / Subcatchment area
Vedenkorkeus Virtaamaa Lämpötilat Sade Haildentila Lumi Varastot Valuma Summat

Tulvakartta, havainnoaseman sijainti ja yksinkertainen ennustekuva

Suurennenna kartta Pienennenna kartta

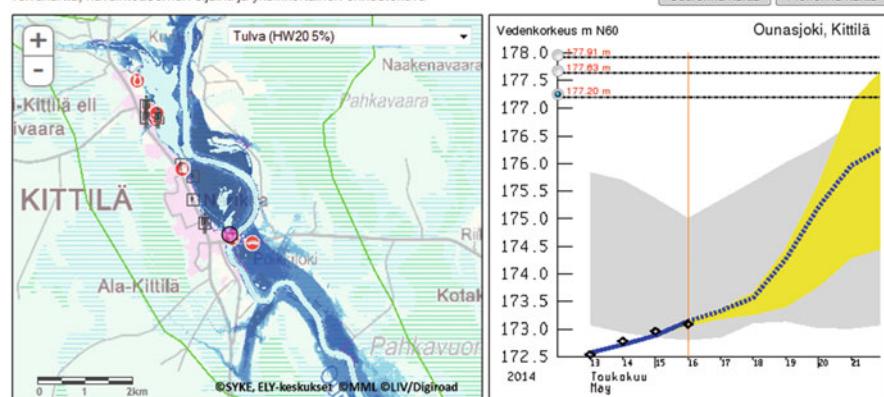


Fig. 7 An example of water level forecast at Kittilä by S20 and corresponding flood area map with 20 years flood (Source: environment.fi)

hydroelectricity generators at local, national, or transnational levels. S20, for example, lists flood mitigation authorities, rescue services, lake regulation authorities, media, county authorities, water power plant operators as users who access their output via a web interface (see Fig. 7).

Some systems are anticipated to expand in terms of geographic area. For example, S2 is currently used operationally at five RFCs, namely the Arkansas-Red Basin RFC (ABRFC), the Colorado Basin RFC (CBRFC), the California-Nevada RFC (CNRFC), the Middle Atlantic RFC (MARFC), and the North East RFC (NERFC). In addition, the New York City Department of Environmental Protection (NYCDEP) is using the HEFS (S2) to improve the management of water supply to NYC. S2 is also currently being expanded to the other RFCs in the USA. S1 is undergoing similar geographic expansions (currently, most users are agencies in the European Union, but this will expand eastwards to include partners outside the European Union).

Other systems are expanding into new application areas where they have not been previously engaging e.g., flood management and planning (S4), reservoir management (S5), energy applications (S9, S10), and water quality (S10). Some systems are simply trying to engage a wider user base within their existing organizational structure (S2, S8, S13, S16). In particular in S4, a number of advanced new applications have been implemented in the field of water quality like the Delft 3D model of the Romagna coast (nested into AdriaROMS model), the oil spill model of the Po river (Sobek/Delwaq model), or the Sobek model for estimation and forecasting of salt intrusion in the Po river delta. S22 is particularly bold in its ambition and anticipates that *everyone* may use the system in the future.

3 Conclusions: Summary and Future Outlook

This chapter summarized hydrological ensemble prediction systems across the globe. The core message is that all current systems differ (1) because they are tailored towards end user needs, and (2) due to varying inputs and setup, particular for each system. In the near future, it is expected that HEPS will be a well-established part of operational forecasting chains. In this chapter, we have listed 24 systems and provided brief descriptions of the systems. Forecast objectives range from water resources to natural hazards and to areas such as river navigation. The systems provide forecasts from hours to months and from small catchments to continents. Uncertainties and biases are captured in different ways, sometimes using raw meteorological ensembles alone and, in other cases, including statistical pre- and post-processing to account for the total predictive uncertainty, thus correcting for biases in the meteorological and hydrologic forecasts.

The future challenges are manifold. Pagano et al. (2014), Wetterhall et al. (2013), and Emerton et al. (2015) have summarized many of the challenges faced by operational forecasting systems. For example, Wetterhall et al. (2013) performed a review of the challenges and priorities of EFAS (S1). They found that the most popular priorities for development were to include verification of past forecast

performance, a multimodel approach for hydrological modelling, extending forecast skill in the medium range (>3 days) and more focus on education and training on the interpretation of forecasts.

This is echoed by the requirements of other systems, which highlight that future challenges include availability of data in real time and historically (discharge, meteorological variables, economic damage, etc.), on-going funding of current systems and future upgrades, staff turnover, successful operational rollout (of the systems which are not operational yet or are expanding), continued hindcasting with operational meteorological models, training of the stakeholders (decision-makers), and development of graphical and other products that are useful and understandable to a nonspecialist audience (where applicable). Often systems' developers and users highlight the need to develop an *ensemble culture*, which shows uncertainties in hydrological forecasting in an easy-to-understand way to better support decisions. In particular, the social and economic value of probabilistic forecasts needs to be established and demonstrated better. Other important aspects are related to a system's interoperability, through the use of internationally recognized data standards (like, for example, WaterML2) and service types, thus supporting a services stack framework that shares catalogue data, metadata, and data with the international hydrological community, including ensemble prediction activities.

References

- N. Addor, S. Jaun, F. Fundel, M. Zappa, An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* **15**, 2327–2347 (2011). <https://doi.org/10.5194/hess-15-2327-2011>
- A. Agnelli, M. Del Longo, C. De Michele, V. Pavan, S. Pecora, R. Vezzoli, E. Zenoni, Seasonal to daily drought prediction in the Po catchment, Italy, in *Proceedings of the 13th Plinius Conference on Mediterranean Storms Savona, Italy* 7–9 Sept 2011
- L. Alfieri, P. Salamon, F. Pappenberger, F. Wetterhall, J. Thielen, Operational early warning systems for water-related hazards in Europe. *Environ. Sci Pol* **21**, 35–49 (2012)
- L. Alfieri, P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, F. Pappenberger, GloFAS – global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.* **17**, 1161–1175 (2013). <https://doi.org/10.5194/hess-17-1161-2013>
- L. Alfieri, F. Pappenberger, F. Wetterhall, T. Haiden, D. Richardson, P. Salamon, Evaluation of ensemble streamflow predictions in Europe, *J. Hydrol.*, (online version) (2014). <https://doi.org/10.1016/j.jhydrol.2014.06.035>
- F. Anctil, M.H. Ramos, F. Pappenberger, Multi-model approaches for river flow forecasting: blessing or burden? (2014) <http://hepex.irstea.fr/multi-model-approaches-for-river-flow-forecasting-blessing-or-burden/>. Last accessed 29 Apr 2014
- E.A. Anderson, National Weather Service River Forecast System-Snow Accumulation and Ablation Model, NOAA Technical Memorandum: NWS Hydro-17, U.S. National Weather Service (1973)
- P. Arnaud, J. Lavabre, Coupled rainfall model and discharge model for flood frequency estimation. *Water Resour. Res.* **38**–6 (2002). <https://doi.org/10.1029/2001WR000474>
- G. Balsamo, P. Viterbo, A. Beljaars, B. van den Hurk, M. Hirschi, A.K. Betts, K. Scipal, A revised hydrology for the ECMWF model: verification from field site to terrestrial water storage and impact in the Integrated Forecast System. *J. Hydrometeorol.* **10**, 623–643 (2009)
- S. Beldring, K. Engeland, L.A. Roald, N.R. Sælthun, A. Voksø, Estimation of parameters in a distributed precipitation-runoff model for Norway. *HESS* **7**(3), 304–316 (2003)

- S. Bergström, The HBV model, in *Computer Models of Watershed Hydrology*, ed. by V.P. Singh (Water Resources Publications, Highlands Ranch, 1995)
- L. Berthet, V. Andréassian, C. Perrin, P. Javelle, How crucial is it to account for the antecedent moisture conditions in flood forecasting? Comparison of event-based and continuous approaches on 178 catchments. *Hydrol. Earth Syst. Sci.* **13**(6), 819–831 (2009). <https://doi.org/10.5194/hess-13-819-2009>
- K. Bogner, F. Pappenberger, Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system *Water Resour. Res.* **47** (2011)
- M. Bremicker, G. Brahmer, N. Demuth, F.-K. Holle, I. Haag, Räumlich hoch aufgelöste LARSIM Wasserhaushaltmodelle für die Hochwasservorhersage und weitere Anwendungen. *KW Korrespondenz Wasserwirtschaft*. **6**(9), 509–514 (2013)
- F. Bressand, Le projet ALHTAIR du service d'annonce des crues. *La Houille Blanche* **2**, 64–68 (2002)
- J.D. Brown, J. Demargne, D.-J. Seo, Y. Liu, The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Softw.* **25**, 854–872 (2010)
- P. Burek, J. Van Der Knijff, A. De Roo, LISFLOOD, distributed water balance and flood simulation model – revised user manual 2013. *Inst. Environ. Sustain.* **2013**, 150
- R.J.C. Burnash, The NWS river forecast system – catchment modeling, in *Computer Models of Watershed Hydrology*, ed. by V.P. Singh (Water Resources Publications, Littleton, 1995), pp. 311–366
- L. Casicci, A. Montani, S. Pecora, F. Tonelli, M. Vergnani, Pre-operational use of a meteorological and hydrological/hydraulic ensemble approach on the Po River, in *Proceedings of The European Geosciences Union General Assembly*, Vienna, 3–8 Apr 2011
- J. Chardon, T. Mathevret, M. Le Lay, Comparison of two uncertainty dressing methods: SAD VS DAD. In prep. (2014)
- M. Clark, S. Gangopadhyay, L. Hay, B. Rajagopalan, R. Wilby, The Schaake shuffle: a method for reconstructing space–time variability in forecasted precipitation and temperature fields. *J. Hydrometeorol.* **5**, 243–262 (2004)
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**(3–4), 613–626 (2009)
- H.L. Cloke, J. Thielen, F. Pappenberger, S. Nobert, P. Salamon, R. Buizza, G. Bálint, C. Edlund, A. Koistinen, C. de Saint-Aubin, C. Viel, E. Sprokkereef, Progress in the implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for operational flood forecasting. *ECMWF Newslet.* **121**, 20–24 (2009)
- M. Coustau, C. Bouvier, V. Borrell-Estudina, H. Jourde, Flood modelling with a distributed event-based parsimonious rainfall-runoff model: case of the karstic Lez river catchment. *Nat. Hazards Earth Syst. Sci.* **12**, 1119–1133 (2012). <https://doi.org/10.5194/nhess-12-1119-2012>
- M.D. Cranston, R. Maxey, A.C.W. Tavendale, P. Buchanan, A. Motion, S. Cole, A. Robson, R.J. Moore, A. Minett, Countrywide flood forecasting in Scotland: challenges for hydrometeorological model uncertainty and prediction, in *Weather Radar and Hydrology, Proceedings of Symposium*, Exeter (2011), April 2011, IAHS Publ. No. 351
- L. Cuo, C.P. Thomas, Q.J. Wang, A review of quantitative precipitation forecasts and their use in short- to medium-range streamflow forecasting. *J. Hydrometeorol.* **12**, 713–728 (2011). <https://doi.org/10.1175/2011JHM1347.1>
- D.P. Dee, S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold et al., The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. R. Meteorol. Soc.* **137** (656), 553–597 (2011). <https://doi.org/10.1002/qj.828>
- J. Demargne, L. Wu, S.K. Regonda, J.D. Brown, H. Lee, M. He, D.-J. Seo, R. Hartman, H.D. Herr, M. Fresch, J. Schaake, Y. Zhu, The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* 79–98 (2014a), in press
- J. Demargne, L. Wu, S. Regonda, J.D. Brown, H. Lee, M. He, D.-J. Seo, R. Hartman, M. Fresch, J. Schaake, Y. Zhu, The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* **95**, 79–98 (2014b). <https://doi.org/10.1175/BAMS-D-12-00081.1>

- B. Desaint, P. Nogues, C. Perret, R. Garçon, La prévision hydrométéorologique opérationnelle: l'expérience d'Electricité de France [Operational hydro-meteorological forecasting: the experience of Electricité de France]. *La Houille Blanche* **5**, 39–46 (2009) [in French]
- DHI, *MIKE 11 – A Modelling System for Rivers and Channels, Reference Manual*. DHI Water and Environment, Hørsholm, Danmark (2007)
- R.E. Emerton, E.M. Stephens, F. Pappenberger, T.C. Pagano, A.H. Weerts, A.W. Wood, P. Salamon, J.D. Brown, N. Hjerd, C. Donnelly, C.A. Baugh, H.L. Cloke, Continental and global scale flood forecasting systems, WIREs Water, Under review (2015)
- T. Follestad, G. Host, A statistical model for the uncertainty of meteorological forecasts with application to the Knappom and Røykenes catchments. HYDRA note. Available from NVE, Oslo (1998)
- F. Fundel, M. Zappa, Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. *Water Resour. Res.* **47**, W09520 (2011). <https://doi.org/10.1029/2010WR009996>
- F. Fundel, S. Joerg-Hess, M. Zappa, Monthly hydrometeorological ensemble prediction of streamflow droughts and corresponding drought indices. *Hydrol. Earth Syst. Sci.* **395–407** (2013). <https://doi.org/10.5194/hess-17-395-2013>
- R. Garçon, Modèle global pluie-débit pour la prévision et la prédétermination des crues. *La Houille Blanche* **7–8**, 88–95 (1999)
- C. Gebhardt, S.E. Theis, M. Paulat, Z. Ben Bouallègue, Uncertainties in COSMO-DE precipitation forecasts introduced by model perturbations and variation of lateral boundaries. *Atmos. Res.* **100**(2–3), 168–177 (2011)
- U. Germann, M. Berenguer, D. Sempere-Torres, M. Zappa, REAL – ensemble radar precipitation for hydrology in a mountainous region. *Q. J. R. Meteorol. Soc.* **135**, 445–456 (2009). <https://doi.org/10.1002/qj.375>
- H.Y. He, H.L. Cloke, F. Wetterhall, F. Pappenberger, J. Freer, M. Wilson, Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Meteorol. Appl.* **16**(1), 91–101 (2009)
- F.-K. Holle, Hochwasservorhersage in Bayern, Modellumgebung und DV-Vernetzung, in *ÖWAV – Seminar Hochwässer, Bemessung, Risikoanalyse und Vorhersage, Wien 26. Mai 2009. Wiener Mitteilungen, Wasser – Abwasser – Gewässer*, Bd. 216, ed. by Institut für Wasserbau und Ingenieurhydrologie TU-Wien (2009). ISBN: 978-3-85234-108-8
- P. Javelle, C. Foucher, P. Arnaud, J. Lavabre, Flash flood warning at ungauged locations using radar rainfall and antecedent soil moisture estimations. *J. Hydrol.* **394**(1–2), 267–274 (2010)
- P. Javelle, J. Demargne, D. Defrance, J. Pansu, P. Arnaud, Evaluating flash-flood warnings at ungauged locations using post-event surveys: a case study with the AIGA warning system. *Hydrol. Sci. J.* **59**(7), 1390–1402 (2014). <https://doi.org/10.1080/0262667.2014.923970>
- A. Kauffeldt, F. Wetterhall, F. Pappenberger, First step towards a multi-model probabilistic flood forecasting system for Europe, <http://hepex.irstea.fr/existing-continental-hydrological-models/>. Last accessed 29 Apr 2014
- L.S. Kuchment, A.N. Gelfan, A study of effectiveness of the ensemble long-term forecasts of spring floods issued with physically-based models of river runoff formation. *Russ. Meteorol. Hydrol.* **34**, 100–109 (2009)
- L.S. Kuchment, A.N. Gelfan, Assessment of extreme flood characteristics based on a dynamic-stochastic model of runoff generation and the probable maximum discharge. *J. Flood Risk Manage.* **4**, 115–127 (2011)
- L.S. Kuchment, V.N. Demidov, Y.G. Motovilov, A physically-based model of the formation of snowmelt and rainfall runoff, in *Modeling Snowmelt-Induced Processes*, ed. by E.M. Morris, vol 155 (IAHS Publications, Budapest, 1986), pp. 27–36
- M. Le Lay, J. Gailhard, P. Bernard, R. Garçon, Prévisibilité hydrométéorologique à l'échelle intrasaisonnière: l'expérience d'EDF-DTG. Ateliers de Modélisation de l'Atmosphère, Météo-

- France, Toulouse. (2013). http://www.meteo.fr/cic/meetings/2013/AMA/resumes/pres_027.pdf. [In French]
- G. Lindström, B. Johansson, M. Persson, M. Gardelin, S. Bergström, Development and test of the distributed HBV-96 hydrological model. *J. Hydrol.* **201**(1), 272–288 (1997)
- K. Ludwig, M. Bremicker, The water balance model LARSIM, in *Freiburger Schriften zur Hydrologie*, Band 22 (Institut für Hydrologie der Universität Freiburg, 2006). pp. 1–141
- H. Maden, C. Skotner, Adaptive state updating in real-time river flow forecasting – a combined filtering and error forecasting procedure. *J. Hydrol.* **308**, 302–312 (2005)
- R. Marty, I. Zin, C. Obled, G. Bontron, A. Djeboua, Towards real-time daily PQPF by an analog sorting approach. Application to flash flood catchments. *J. Appl. Meteorol. Climatol.* **51**, 505–520 (2012). <https://doi.org/10.1175/JAMC-D-11-011.1>
- R. Maxey, M. Cranston, A. Tavendale, P. Buchanan, The use of deterministic and probabilistic forecasting in countrywide flood guidance in Scotland. *Br. Hydrol. Soc.* (2012). <https://doi.org/10.7558/bhs.2012.ns33>
- D. Meißner, S. Rademacher, Die verkehrsbezogene Wasserstandsvorhersage für die Bundeswasserstraße Rhein – Verlängerung des Vorhersagezeitraums und Steigerung der Vorhersagequalität. *Korrespondenz Wasserwirtschaft*, **9**, 531–537 09/10 (2010)
- D. Meißner, S. Gebauer, A.H. Schumann, M. Pahlow, S. Rademacher, Analyse radarbasierter Niederschlagsprodukte als Eingangsdaten verkehrsbezogener Wasserstandsvorhersagen am Rhein. *Hydrol. Wasserbewirtsch.* **1**, 02/2012 (2012). https://doi.org/10.5675/HyWa_2012_1_2
- J. Neal, I. Villanueva, N. Wright, T. Willis, T. Fewtrell, P. Bates, How much physical complexity is needed to model flood inundation? *Hydrol. Process.* (2011) <https://doi.org/10.1002/hyp.8339>
- Ø. Langsrud, A. Frigessi, G. Høst, Pure model error for the HBV model. HYDRA note. Available from NVE, Oslo (1998a)
- Ø. Langsrud, G. Høst, T. Follestad, A. Frigessi, Quantifying uncertainty in HBV runoff forecasts by stochastic simulations. HYDRA note. Available from NVE, Oslo (1998b)
- Pagano et al., Challenges of operational river forecasting. *J. Hydrometeorol.* 2014; e-View <https://doi.org/10.1175/JHM-D-13-0188.1>
- F. Pappenberger, TIGGE and the multi-model approach in hydrology (a brief review) (2014). <http://hepex.irstea.fr/tigge-and-the-multi-model-approach-in-hydrology-a-brief-scientific-review/>. Last accessed 29 Apr 2014
- F. Pappenberger, J. Thielen, M. del Medico, The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrol. Process.* **25**(7) (2011). <https://doi.org/10.1002/hyp.7772,2010>
- C. Perrin, C. Michel, V. Andréassian, Improvement of a parsimonious model for streamflow simulation. *J. Hydrol.* **279**, 275–289 (2003). [https://doi.org/10.1016/S0022-1694\(03\)00225-7](https://doi.org/10.1016/S0022-1694(03)00225-7)
- M. Renner, M.G.F. Werner, S. Rademacher, E. Sprokkrat, Verification of ensemble flow forecasts for the River Rhine. *J. Hydrol.* **376**, 463–475 (2009). <https://doi.org/10.1016/j.jhydrol.2009.07.059>
- J. Roe, C. Dietz, P. Restrepo, J. Halquist, R. Hartman, R. Horwood, W. Olsen, H. Opitz, R. Shedd, E. Welles, NOAA's Community Hydrologic Prediction System. Presented at the 2nd Joint Federal Interagency Conference in Las Vegas, June 27 – July 1, 2010. Link:http://acwi.gov/sos/pubs/2ndJFIC/Contents/7E_Roe_12_28_09.pdf
- H. Roux et al., A physically-based parsimonious hydrological model for flash floods in Mediterranean catchments. *Nat. Hazards Earth Syst. Sci.* **11**(9), 2567–2582 (2011)
- N.R. Sælthun, The “Nordic” HBV model. Norwegian Water Resources and Energy Directorate Report No. 7 (1996)
- J. Schaake, J. Demargne, R. Hartman, M. Mullusky, E. Welles, L. Wu, H. Herr, X. Fan, D.J. Seo, Precipitation and temperature ensemble forecasts from single-value forecasts. *Hydrol. Earth Syst. Sci.* **4**, 655–717 (2007)
- Y. Seity, P. Brousseau, S. Malardel, G. Hello, P. Bernard, F. Bouttier, C. Lac, V. Masson, The AROME-France convective-scale operational model. *Mon. Weather Rev.* **139**, 976–991 (2010)
- Y. Seity, P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, V. Masson, The AROME-France convective-scale operational model. *Mon. Weather Rev.* **139**, 976–991 (2011)

- D.-J. Seo, H.D. Herr, J.C. Schaake, A statistical post-processor for accounting of hydrologic uncertainty in short-range ensemble streamflow prediction. *Hydrol. Earth Syst. Sci.* **3**, 1987–2035 (2006)
- D. Shin, A. Schepen, T. Peatey, S. Zhou, A. MacDonald, T. Chia, J. Perkins, N. Plummer, WAFARI: a new modelling system for seasonal streamflow forecasting service of the Bureau of Meteorology, in *19th International Congress on Modelling and Simulation (MODSIM)*, (Perth, 2011), 12–16 Dec 2011
- W. Sittner, C. Schauss, J. Monro, Continuous hydrograph synthesis with an API-type hydrologic model. *Water Resour. Res.* **5**(5), 1007–1022 (1969)
- T. Skaugen, C. Onof, A rainfall-runoff model parameterized from GIS and runoff data. *Hydrol. Process.* (2013) <https://doi.org/10.1002/hyp.9968> (in press)
- SPC-Allier Réglement de surveillance, de prévision et de transmission de l'Information sur les Crues (RIC) *Direction Régionale de l'Environnement, de l'Aménagement et du Logement/Auvergne*, 2013. Link http://www.vigicrues.ecologie.gouv.fr/ftp/RIC/RIC_SPC_AL_2013.pdf
- SPC-LCI Réglement de surveillance, de prévision et de transmission de l'Information sur les Crues (RIC) *Direction Régionale de l'Environnement, de l'Aménagement et du Logement/Centre*, 2013. Link http://www.vigicrues.ecologie.gouv.fr/ftp/RIC/RIC_SPC_LCI_2013.pdf
- J. Thielen, K. Bogner, F. Pappenberger, M. Kalas, M. del Medico, A. de Roo, Monthly-, medium- and short range flood warning: testing the limits of predictability. *Meteorol. Appl.* **16**(1), 77–90 (2009)
- V. Thiemig, B. Bisselink, F. Pappenberger, J. Thielen, A pan-African flood forecasting system. *Hydrol. Earth Syst. Sci. Discuss.* **11**, 5559–5597 (2014)
- N.K. Tuteja, D. Shin, R. Laugesen, U. Khan, Q. Shao, E. Wang, M. Li, H. Zheng, G. Kuczera, D. Kavetski, G. Evin, M. Thyre, A. Macdonald, T. Chia, B. Le, *Experimental Evaluation of the Dynamic Seasonal Streamflow Forecasting Approach*. Technical report, (Bureau of Meteorology, Canberra, 2011)
- S.J. Van Andel, R. Price, A. Lobbrecht, F. van Kruiningen, R. Mureau, W. Cordero, Framework for anticipatory water management. *J. Water Resour. Plan. Manag.* **140**, 533–542 (2014). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000254](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000254)
- J.M. van der Knijff, J. Younis, A.P.J. de Roo, LISFLOOD: a GIS-based distributed model for river-basin scalewater balance and flood simulation. *Int. J. Geogr. Inf. Sci.* **24**(2), 189–212 (2010)
- B. Vehviläinen, The watershed simulation and forecasting system in the National Board of Waters and Environment. Publications of the Water and Environment Research Institute. National Board of Waters and the Environment, Finland. No. 17. (1994)
- Verkade et al., Development path of operational forecasting systems. (2014)
- B. Vincendon, V. Ducrocq, G.M. Saulnier, L. Bouilloud, K. Chancibault, F. Habets, J. Noilhan, Benefit of coupling the ISBA land surface model with a TOPMODEL hydrological model version dedicated to Mediterranean flash-floods. *J. Hydrol.* (Impact Factor: 2.96). 01/2010; **394**, 256–266 (2010). <https://doi.org/10.1016/j.jhydrol.2010.04.012>
- B. Vincendon, V. Ducrocq, O. Nuissier, B. Vié, Introducing perturbation in rainfall fields for an ensemble forecasting of flash-flood. *NHESS* **11**, 1529–1544 (2011)
- D. Viviroli, M. Zappa, J. Gurtz, R. Weingartner, An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. *Environ. Model. Softw.* **24**(10), 1209–1222 (2009). doi:10.1016/j.envsoft.2009.04.001
- Q.J. Wang, D.E. Robertson, Multisite probabilistic forecasting of seasonal flows for streams with zero value occurrences. *Water Resour. Res.* **47**, W02546 (2011). doi:10.1029/2010WR009333
- Q.J. Wang, D.E. Robertson, F.H.S. Chiew, A Bayesian joint probability modelling approach for seasonal forecasting of streamflows at multiple sites. *Water Resour. Res.* **45**, W05407 (2009a). <https://doi.org/10.1029/2008WR007355>
- Y. Wang et al., The Central European limited area ensemble forecasting system: ALADINLAEF. RC LACE report (2009b). <http://www.rclace.eu/File/Predictability/2009/laef4lace.pdf>

- A.H. Weerts, H.C. Winsemius, J.S. Verkade, Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales). *Hydrol. Earth Syst. Sci.* **15**, 255–265 (2011). <https://doi.org/10.5194/hess-15-255>
- M. Werner, M. van Dijk, J. Schellekens, Delft-FEWS: an open shell flood forecasting system, in *6th International Conference on Hydroinformatics*, ed. by S.Y. Lioung, K. Phoon, V. Babovic (World Scientific Publishing Company, Singapore, 2004), pp. 1205–1212
- M. Werner, J. Schellekens, P. Gijsbers, M. van Dijk, O. van den Akker, K. Heynert, The delft-FEWS flow forecasting system. *Environ. Model. Softw.* **40**, 65–77 (2013). <https://doi.org/10.1016/j.envsoft.2012.07.010>
- F. Wetterhall, F. Pappenberger, L. Alfieri, H.L. Cloke, J. Thielen-del Pozo, S. Balabanova, J. Daňhelka, A. Vogelbacher, P. Salamon, I. Carrasco, A.J. Cabrera-Tordera, M. Corzo-Toscano, M. Garcia-Padilla, R.J. Garcia-Sanchez, C. Ardilouze, S. Jurela, B. Terek, A. Csik, J. Casey, G. Stankūnavičius, V. Ceres, E. Sprokkereef, J. Stam, E. Anghel, D. Vladikovic, C. Alionte Eklund, N. Hjerdt, H. Djerv, F. Holmberg, J. Nilsson, K. Nyström, M. Sušnik, M. Hazlinger, M. Holubecka, HESS opinions “Forecaster priorities for improving probabilistic flood forecasts”. *Hydrol. Earth Syst. Sci.* **17**, 4389–4399 (2013). <https://doi.org/10.5194/hess-17-4389-2013>
- L. Wu, D.-J. Seo, J. Demargne, J.D. Brown, S. Cong, J. Schaake, Generation of ensemble precipitation forecast from single-valued quantitative precipitation forecast via meta-Gaussian distribution models. *J. Hydrol.* **399**(3–4), 281–298 (2011)
- J. Younis, M.H. Ramos, J. Thielen, EFAS forecasts for the March-April 2006 flood in the Czech part of the Elbe River Basin – a case study. *Atmos. Sci. Lett.* **9**(2), 88–94 (2008)
- I. Zalachori, M.H. Ramos, R. Garçon, T. Mathevet, J. Gailhard, Statistical processing of forecasts for hydrological ensemble prediction: a comparative study of different bias correction strategies. *Adv. Sci. Res.* **8**, 135–141 (2012)
- M. Zappa, M.W. Rotach, M. Arpagaus, M. Dorninger, C. Hegg, A. Montani, R. Ranzi, F. Ament, U. Germann, G. Grossi, S. Jaun, A. Rossa, S. Vogt, A. Walser, J. Wehrhan, C. Wunram, MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems. *Atmos. Sci. Lett.* **2**, 80–87 (2008). <https://doi.org/10.1002/asl.183>



Flash Flood Forecasting Based on Rainfall Thresholds

Lorenzo Alfieri, Marc Berenguer, Valentin Knechtl, Katharina Liechti,
Daniel Sempere-Torres, and Massimiliano Zappa

Contents

1	Introduction	1225
2	Flash Flood Monitoring at the European Scale with EPIC	1226
2.1	EPIC	1226
2.2	EPIC in the Operational EFAS Monitoring	1227
2.3	Case Study: Flash Floods in Sardinia (Italy) in November 2013	1231
2.4	Performance in Early Detection of Extreme Storms and Flash Floods	1231
2.5	An Outlook to Future Flash Flood Early Warning in Europe	1233
3	Deterministic and Ensemble Flash Flood Early Warning in Southern Switzerland	1234
3.1	Estimation of Return Periods	1234
3.2	Combined Use of Station Information and Gridded Data for IDF Estimations	1236
3.3	Operational Implementation Forced by Deterministic and Ensemble NWP	1238
3.4	Case Study in October 2014	1240
3.5	An Outlook to Future Flash Flood Early Warning in Switzerland	1244
4	Flash Flood Detection in Catalonia	1245
4.1	Probabilistic Rainfall Inputs	1245
4.2	Flash Flood Hazard Assessment	1246

L. Alfieri (✉)

Directorate for Space, Security and Migration, European Commission – Joint Research Centre,
Ispra, VA, Italy

e-mail: Lorenzo.Alfieri@jrc.ec.europa.eu

M. Berenguer · D. Sempere-Torres

Center of Applied Research in Hydrometeorology, Universitat Politècnica de Catalunya, Barcelona,
Spain

e-mail: marc.berenguer@crahi.upc.edu; daniel.sempere@crahi.upc.edu

V. Knechtl · K. Liechti · M. Zappa

Swiss Federal Research Institute WSL, Birmensdorf, Switzerland

e-mail: valentinknechtl@gmail.com; kaethi.liechti@wsl.ch; massimiliano.zappa@wsl.ch

4.3	Implementation in Catalonia (NE Spain) and Case Studies	1247
4.4	An Outlook to Future Flash Flood Early Warning in Catalonia	1254
5	Conclusions	1256
	References	1256

Abstract

Extreme rainstorms often trigger catastrophic flash floods in Europe and in several areas of the world. Despite notable advances in weather forecasting, most operational early warning systems for extreme rainstorms and flash floods are based on rainfall observations derived from rain gauge networks and weather radars, rather than on forecasts. As a result, warning lead times are bounded to few hours, and warnings are usually issued when the event is already taking place.

This chapter illustrates three recently developed systems that use information on observed and forecasted precipitation to issue flash flood warnings. The first approach is an indicator for heavy precipitation events, developed to complement the flood early warning of the European Flood Awareness System (EFAS) and targeted to short and intense events, possibly leading to flash flooding in small catchments. The system is based on the European Precipitation Index Based on Simulated Climatology (EPIC), which in EFAS is computed using COSMO-LEPS ensemble weather forecasts and a 20-year consistent reforecast dataset.

The second system is a flash flood early warning tool developed based on precipitation statistics. A total of 759 sub-catchments in southern Switzerland is considered. Intensity-duration-frequency (IDF) curves for each catchment have been calculated based on gridded precipitation products for the period 1961–2012 and gridded reforecast of the COSMO-LEPS for the period 1971–2000. The different IDF curves at the catchment level in combination with precipitation forecasts are the basis for the flash flood early warning tool. The forecast models used are COSMO-2 (deterministic, updated every 3 h and with a lead time of 24 h) and COSMO-LEPS (probabilistic, 16-member and with a lead time of 5 days).

The third system (FF-EWS) uses probabilistic high-resolution precipitation products generated from the observations of the weather radar network to monitor situations prone to trigger flash floods in Catalonia (NE Spain). These ensemble precipitation estimates and nowcasts are used to calculate the basin-aggregated rainfall (that is, the rainfall accumulated upstream of each point of the drainage network), which is the variable used to characterize the potential flash flood hazard.

Examples of successful and less skilful forecasts for all three systems are shown and commented to highlight pros and cons.

Keywords

Extreme precipitation events · Numerical weather predictions · Flash flood early warning · EPIC · IDF · Ensemble forecasting · Reforecasts

1 Introduction

Flash flood forecasting is an important field of applied research because flash floods are cause of major damages (Liechti et al. 2013a; Gaume et al. 2009). Several fatalities linked to flash floods have been reported around the world (French et al. 1983; Jonkman 2005; Gaume et al. 2009). Norbiato et al. (2008) adopt the term of “flash” to indicate situations where very shortly after a triggering precipitation event the level of the rivers rapidly increase. The delay between the rainfall and the peak discharge is so short that almost no mitigation action is possible. Typically flash floods are observed in small torrents showing very rapid runoff concentration delays (Norbiato et al. 2008). The torrential character of the concerned rivers might imply that cascading hazards such as mobilization of sediments and driftwood are possible. Alfieri and Thielen (2015) report on the concurrency of flash floods with landslides and debris flows. Such mass movements might increase damages downstream due to wood and debris blocking the river cross section and causing inundation of the surrounding areas.

Mitigation measures are possible if such situations are predicted some hours in advance. Javelle et al. (2010) summarizes that the challenge of accurate prediction of precipitation (Collier 2007), the very limited area of the affected basins and the reduced time available to make complex model simulations are limiting constraints for a successful application of operational forecasting systems to anticipate flash floods. Thus, although advanced early warning systems using forecasted weather radar estimates can be conceived and evaluated (e.g., Liechti et al. 2013a; Versini et al. 2014), their operational deployment might be slowed down by the delays needed to transfer data from observing systems to the meteorological services and later on the chain to the operator of hydrological models.

One way to cope with this is to provide flash flood early warning by adopting only precipitation data to estimate possible flood damage. One of the most popular approaches is the flash flood guidance (FFG) method. FFG is widely used in the United States and basically keeps track of the rainfall depth and intensity needed to trigger a flood at the outlet of a specific catchment (e.g., Georgakakos 2006). It is calculated by a hydrological model run in inverse mode.

Recent developments focus on ensemble early warning systems relying on precipitation information only. This reduces the need for calibration (e.g., Alfieri et al. 2011). One disadvantage of such approaches is the neglecting of the initial conditions (Javelle et al. 2010). This chapter presents three recent approaches devoted to the early prediction of flash floods. They are presented according to the lead time of the forecasting tools. In Sect. 2 a method deployed at the European scale is evaluated. In Sect. 3 an application designed for very small headwater basins in southern Switzerland is presented. Section 4 presents the approach used in the region of Catalonia (Spain). Section 5 summarizes the experience of the three systems and gives insight on possible further developments.

2 Flash Flood Monitoring at the European Scale with EPIC

2.1 EPIC

The European Precipitation Index Based on Simulated Climatology (EPIC, Alfieri et al. 2011) is an indicator to monitor the European domain for upcoming severe storms possibly leading to flash floods. EPIC only depends on the Quantitative Precipitation Forecast (QPF) and on the modeled river network, while all other hydrological processes (e.g., initial soil moisture, snow accumulation, and melting, among others) are not considered. Despite some important simplifications, compared to the actual processes involved, the system has no calibration parameters and can be seen as an extreme frequency analysis of the aforementioned indicator. As Guillot and Duband (1967) described in the Gradex method, the gradient of the statistical distribution of discharges tends to follow asymptotically that of rainfall, for high return periods. Consequently, the methodology here described aims to detect severe flood events by linking them to extreme rainfall accumulations at the catchment scale.

EPIC is defined as:

$$\text{EPIC}(t) = \max_{\forall di} \left(\frac{UP_{di}(t)}{\frac{1}{N} \sum_{yi=1}^N \max(UP_{di})_{yi}} \right), \quad (1)$$

where UP_{di} is the upstream cumulated precipitation that is the double summation of the precipitation depth (P) over the upstream area and over a certain duration di preceding the considered time t :

$$UP_{di}(t) = \sum_{t-di}^t \sum_A P; \quad (2)$$

In Eq. 1, UP is calculated for each time step t and then rescaled by the corresponding mean of the annual maxima derived from a consistent dataset of N years, for the same point and rainfall duration. Although EPIC was specifically designed for forecast applications, the same formulation can be applied in real-time monitoring and in post-event analysis, using observed precipitation fields. In the latter case, one has to collect a precipitation dataset consistent with the event measurements, particularly with regard to extreme values. Reforecast datasets made available by weather forecasting centers are particularly suitable options, as they are produced with the same model version of operational forecasts.

According to the rational method for the estimation of peak flows (e.g., Chow et al. 1988), durations of accumulation di depend on the typical response time of the catchment and on the average delay in producing runoff after a rainfall event.

Following the findings from Fiorentino et al. (1987) and from Viglione and Blöschl (2009), critical rainfall durations can be assumed of the same magnitude of the catchment lag time and its time of concentration. In EFAS operational runs, EPIC considers $di = \{6, 12, 24\text{ h}\}$, which are typical durations of intense storms leading to flash floods in catchments of size up to 2000 km^2 (Gaume et al. 2009; Reed et al. 2007).

2.2 EPIC in the Operational EFAS Monitoring

In the European Flood Awareness System (EFAS, Bartholmes et al. 2009; Thielen et al. 2009), ensemble weather predictions to compute EPIC are provided by the Consortium for Small-Scale Modeling (COSMO). The Limited-Area Ensemble Prediction System (LEPS) of the COSMO model (Marsigli et al. 2005) is produced twice per day at 00:00 and 12:00 UTC, with a forecasting range of 132 h. COSMO-LEPS is a 16-member ensemble covering central-southern Europe, stretching as far north as Scotland, Denmark, and Latvia. Maps are provided on a rotated spherical grid with horizontal resolution of 0.0625° ($\sim 7\text{ km}$) and temporal resolution of 3 h. Climatological values are derived by a 20-year reforecast datasets starting in 1989, produced by COSMO with the same model configuration used for operational forecasts (Fundel et al. 2010). The dataset was created by initializing the model every 90 h using the ERA-Interim atmospheric reanalysis dataset (Dee et al. 2011) as initial and boundary conditions.

EPIC is run twice per day (00:00 and 12:00 UTC) using COSMO-LEPS forecasts as input, resulting in ensembles of 16 possible temporal evolutions over the forecast range. It is calculated for each pixel of the river network at 1-km grid resolution within the COSMO-LEPS domain, resulting in more than one million points. Reference values are EPIC = 0, when no rainfall is forecast, and EPIC = 1, when the cumulated upstream precipitation at a point equals the corresponding mean of the annual maxima for at least one of the considered rainfall durations. In hydrological monitoring systems, flood warning thresholds are usually set for specific return periods; hence, a more intuitive representation is to estimate the return period of EPIC and show its values for the selected events. The approach used in EFAS is described as follows:

- At each forecast, a preliminary empirical set of rules selects the most downstream points with $\text{EPIC} > 1$ for at least four members out of 16 (25% probability) and with $\text{EPIC} > 1.5$ for at least three members. An additional criterion is set on the upstream area of points, which is bound in the range $50\text{--}5000\text{ km}^2$, to address the analysis on flash flood prone catchments. The lower value is bounded by the spatial resolution of the weather prediction data.
- For the selected set of points, a 2-parameter gamma distribution is fitted to EPIC ensembles at each time step of the forecast horizon. The probability density function (pdf) of a gamma-distributed random variable x is defined as:

$$f(x) = \frac{x^{\alpha-1} e^{-x/\beta}}{\beta^\alpha \Gamma(\alpha)} \text{ for } x \geq 0 \text{ and } \alpha, \beta > 0, \quad (3)$$

where α is the shape parameter, β the scale parameter, and $\Gamma(\cdot)$ denotes the gamma function. L-moments estimators are used to fit empirical values as described by Hosking (1990). A similar approach is used and described by Alfieri et al. (2012) to fit ensemble streamflow predictions derived by COSMO-LEPS weather forecasts. Results by Alfieri et al. (2012) show that fitting raw 16-member hydrological ensembles with analytical gamma distributions leads to improvements both in the quantitative streamflow estimation and particularly in the threshold exceedance analysis.

- A Gumbel extreme value distribution is hypothesized for the annual maxima of EPIC(di), for durations of 6, 12, and 24 h, derived from the 20-year climatology. Its cumulative distribution function $F(x)$ takes the form:

$$F(x) = \exp\left(-\exp\left(-\frac{x-\xi}{\alpha}\right)\right), \quad (4)$$

where α is the scale parameter and ξ the location parameter of the Gumbel distribution.

- The two parameters of each distribution are estimated by equalling the first two sample L-moments with those of the analytical distribution (λ_1, λ_2):

$$\lambda_1 = \xi + \alpha\gamma, \quad (5)$$

$$\lambda_2 = \alpha\log 2, \quad (6)$$

where γ is Euler's constant: $\gamma = 0.5772$. Return periods T of EPIC(di) are estimated from their analytical cumulative distributions $F(\text{EPIC}(di))$ in Eq. 4 and by recalling the relation $T = 1/(1-F(\text{EPIC}(di)))$.

- The initial set of reporting points is regrouped into three alert classes. Medium alert class (yellow color coding) includes all points having a maximum probability larger than 15% of exceeding the 2-year return period. Similarly, high (in red) and severe (in purple) alert classes include all points having a maximum probability larger than 15% of exceeding the 5-year and 20-year return period, respectively.

EPIC is calculated operationally, and results are visualized in a web interface and monitored on a daily basis to detect small-scale extreme events over Europe. Products have been designed in analogy to those developed for EFAS, as they are specifically targeted to explore and visualize probabilistic forecasts. Products shown include a map of the maximum probability of exceeding the mean annual maximum of EPIC over the forecast range (i.e., $\max_{\forall t} (\Pr(\text{EPIC}(t) > 1))$), where different probabilities are indicated with shades of red. Also, the three layers of alert points defined above are shown with triangles of size proportional to the probability of EPIC to exceed the corresponding alert class. An example of the resulting display of EPIC on the EFAS web interface (www.efas.eu) is shown in Fig. 1 for forecasts of

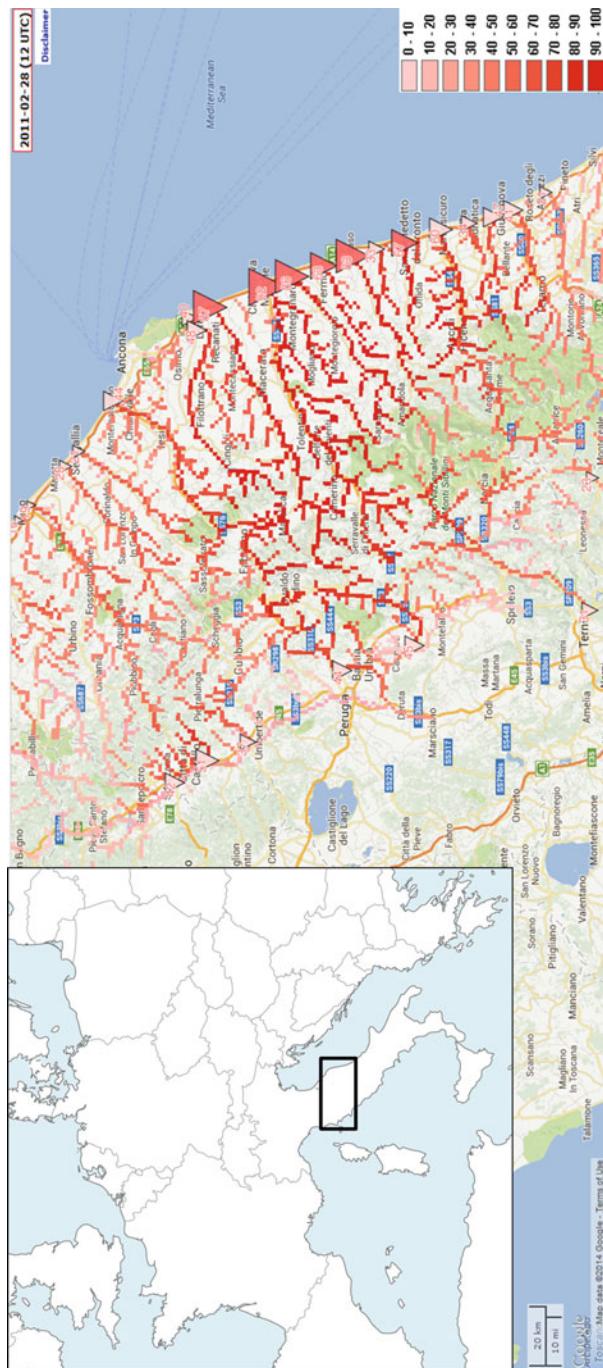


Fig. 1 Map of forecast probability of exceeding the mean annual rainstorm (i.e., EPIC = 1). Model run of 28 February 2011 12:00 UTC. Color saturation of red lines is proportional to the percent probability level (see legend). EPIC reporting points and flash flood alerts are shown with pink and red triangles, respectively

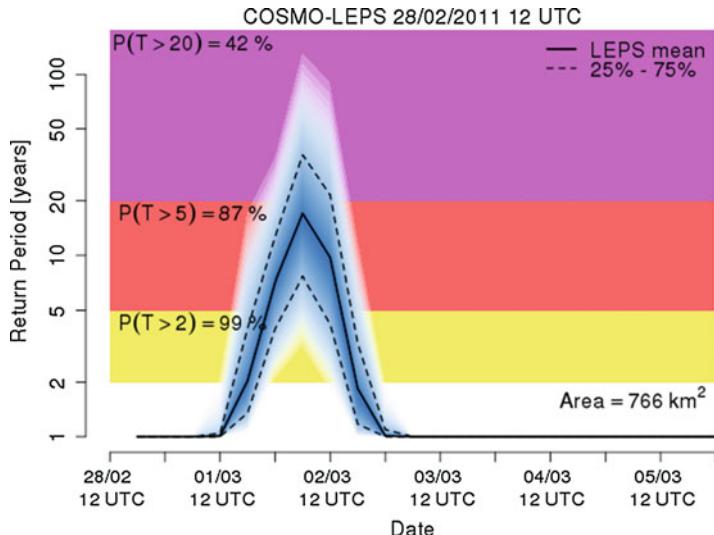


Fig. 2 Forecast return period of EPIC on 28 February 2011 12:00 UTC for a reporting point at the outlet of the Potenza River in central Italy. The maximum probabilities of exceeding the three alert thresholds are shown on the left. The ensemble mean and the interquartile range are also shown with *solid* and *dashed black lines*

the 28 February 2011 12:00 UTC for the Marche region in central Italy. At each reporting point, a time-plot displays with blue shadings the forecast return period of EPIC for a probability range of 5–95% (e.g., see Fig. 2). According to the operational alerting rules currently used in EFAS, these forecasts would have resulted in a flash flood alert for six rivers (indicated by red triangles in Fig. 1), in the Marche and Abruzzo regions. On 02 March 2011, media news confirmed widespread flooding in most rivers of the two regions, due to the most intense rainfall in 40 years (http://www.anas.it/web/notizie/rubriche/english/2011/03/02/visualizza_new.html_1561001936.html, last accessed on 11 July 2016). The location and timing of this event were skillfully captured by EPIC forecasts, together with the river basins where the most extreme features would exhibit.

One can see in Fig. 2 that return period time-plots put the focus on extreme conditions, while rainfall depths below the mean annual maxima tend to converge to the 1-year return period. Vice versa, in such plots the uncertainty spread increases for extreme values, proving the usefulness of the probabilistic information but also showing the difficulty in providing accurate alerts in operational warning systems. As example, the event peak in Fig. 2 has a coefficient of variation (CV) of the predicted EPIC ensemble of $CV_{EPIC} = 0.17$, while the corresponding one calculated on the ensemble of return periods, $CV_T = 0.87$, is about five times larger.

2.3 Case Study: Flash Floods in Sardinia (Italy) in November 2013

During the night between 18 and 19 November 2013, exceptionally high rainfalls fell in Sardinia, Italy. According to the website of ARPA Sardegna, the regional environment agency, more than 350 mm of precipitation was recorded within 24 h in a considerably large region in central-east Sardinia. News reports from the BBC (<http://www.bbc.co.uk/news/world-europe-24996292>, last accessed on 11 July 2016) quote up to 440 mm in 90 min, which place it among the highest measurements on record worldwide.

A total of 18 casualties were reported, the area near Olbia being the most severely affected. Weather forecasts used in EFAS predicted high rainfalls for Sardinia, Corsica, northern Spain, and southern France since 14 November, though forecasts were not persistent and underestimated the actual precipitation totals.

For this event, EPIC showed a medium probability for flash floods across the island about 1 day before the event (see Fig. 3). Only the forecast of 18 November 2013 12:00 UTC indicated a high probability of flash floods which would have resulted in an EFAS warning to authorities, though when the event was ongoing. This example highlights how the performance of EPIC are directly related to those of the underlying weather models in predicting extreme precipitation. In addition, at the grid resolution of 5–10 km the forecast uncertainty is high. Often only few members of the ensemble are able to reproduce extreme conditions potentially leading to flash floods, making the alert detection particularly challenging. A key issue in operating early warning systems based on ensemble or probabilistic forecasts is the definition of minimum probability thresholds used to trigger the alerts. Lower thresholds enable the detection of events with high forecast uncertainty, as in the case described above, and improve the hit rate of the system though at the cost of a higher average false alarm rate. Ideally, probability thresholds should be calibrated depending on the ratio between the cost of issuing an alert and the economic losses in case a flash flood occurs.

2.4 Performance in Early Detection of Extreme Storms and Flash Floods

By its definition, EPIC is not designed to detect all types of floods, but rather those in small-size catchments (with a lower boundary depending on the resolution of the NWP) induced by short and intense rainfall events. The collection of quantitative discharge data and of flood thresholds in small rivers throughout Europe, for validation purpose, is a huge and painstaking task. Flash floods usually occur in ungauged catchments, where the only source of information is post-event descriptive reports. Besides, even when gauging stations are available, they are sometimes damaged and made inoperative by the rage of the flood flow. Performance in operational monitoring of the early warning system described above was assessed through a qualitative approach, by selecting the strongest recorded signals of

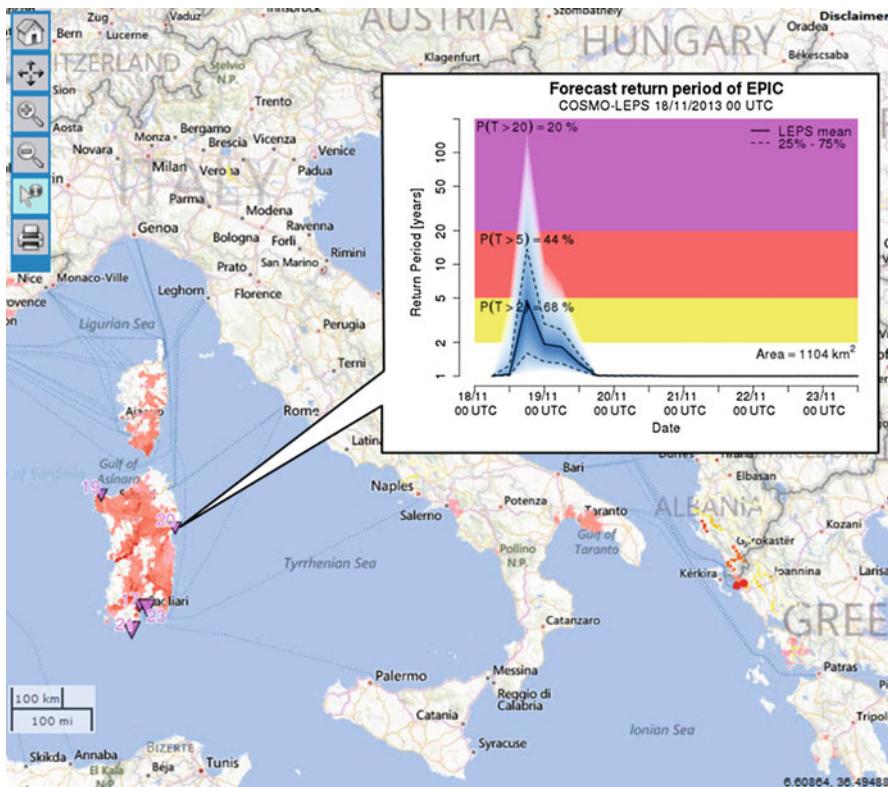


Fig. 3 EPIC flash flood forecast of 18 November 2013 00:00 UTC, at the outlet of the Cedrino River, Sardinia

upcoming severe events from EPIC and verifying the actual occurrence of flooding events in the areas where they were forecast.

A research work by Alfieri and Thielen (2015) analyzed 22 months of forecasts of EPIC driven by COSMO-LEPS forecasts, ending in September 2011. They obtained a threshold for flash flood alerts corresponding to 60% probability of exceeding the 5-year return period of EPIC. Such criterion enabled the detection of 363 alerts out of the overall set of reporting points (see Fig. 4), belonging to 50 distinct events. Their occurrence was investigated by searching for reported news on the internet. The main source of information used is the flooding section of the European Media Monitor (EMM, <http://emm.jrc.it/>, last accessed on 19 August 2015). EMM NewsBrief was developed at the Joint Research Centre of the European Commission. It is a summary of news from the world in several languages, which is generated automatically by software algorithms. EMM news have been complemented by the Emergency Events Database (EM-DAT, <http://www.emdat.be/>, last accessed on 11 July 2016) of the Centre for Research on the Epidemiology of Disasters (CRED) and by targeted internet searches on national and regional news

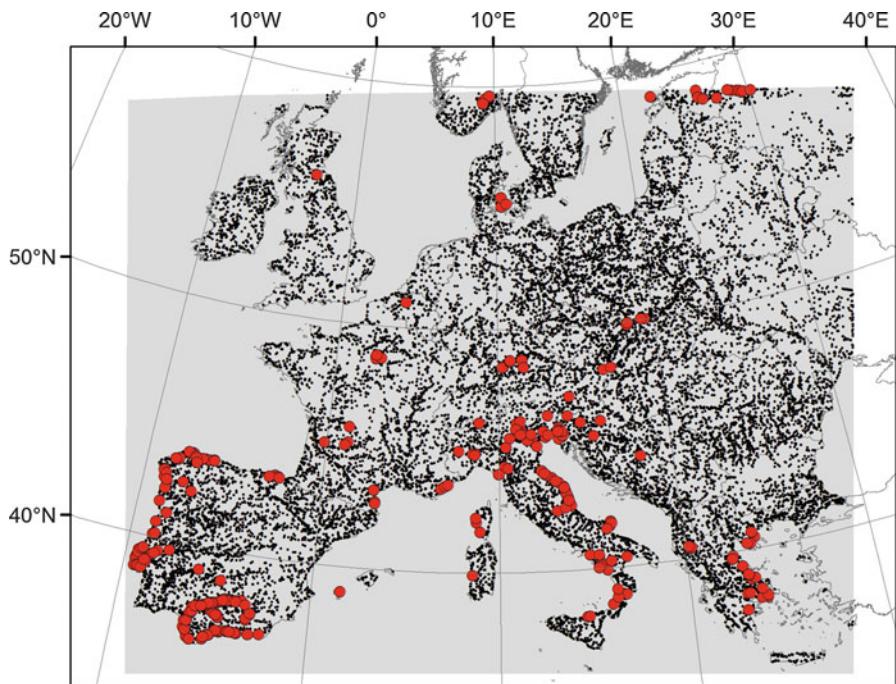


Fig. 4 EPIC reporting points (black dots) and flash flood alerts (red circles) between December 2009 and September 2011. COSMO-LEPS spatial domain is also indicated with a gray shaded area (From Alfieri and Thielen (2012), Copyright © 2012 Royal Meteorological Society, first published by John Wiley & Sons, Ltd.)

websites. Out of 50 events, in 42 cases reported, news of rainstorms and economic losses were found, due to a combination of floods, flash floods, surface water flooding, debris flow, landslides, hail, lightning, sea waves, or wind storms.

The remaining eight cases were not confirmed by media news as hazardous events. Reasons were attributed to errors in the event severity (two cases), to a shift in their location (three cases), and to boundary issues for regions at the edge of the simulation domain (three cases).

2.5 An Outlook to Future Flash Flood Early Warning in Europe

EPIC is used for flash flood early warning as part of the operational EFAS suite. A semiautomated procedure was set up to speed up the preparation of alert emails and reduce the component of human error due to stress and time pressure. A fully automated procedure to send flash flood alerts is envisaged for the near future, in order to maximize the warning lead time after the latest forecast results are ready. In addition, since 2014 EFAS flash flood, alert emails include information on the landslide susceptibility of the affected catchments (from Günther et al. 2013), to

help identify possible hotspots of upcoming debris flow events and rainfall-driven landslides.

Research work by Raynaud et al. (2015) showed how the performance in the event detection can be improved through a modified version of EPIC based on runoff instead of precipitation. It takes into account the initial soil moisture conditions and geomorphological features to weigh the contribution of rainfall on the severity of the forecast event. Similarly, Alfieri et al. (2014) extended the formulation of EPIC to detect floods in a wider range of basin size, using a nonparametric approach relying exclusively on the output of a state-of-the-art global circulation model coupled with a land-surface scheme. They defined the extreme runoff index (ERI), which is designed to detect extreme accumulations of surface runoff over critical flood durations for each section of the river network.

3 Deterministic and Ensemble Flash Flood Early Warning in Southern Switzerland

The flash flood early warning tool deployed in southern Switzerland (canton of Ticino) is running in real time since July 2013. For some regions of the target area, real-time hydrological ensemble predictions have been established since 2007 (Zappa et al. 2008, 2013) including the use of ensemble weather radar Quantitative Precipitation Estimates (QPEs) (Germann et al. 2009; Zappa et al. 2011; Liechti et al. 2013b). Such real-time systems based on the hydrological model PREVAH (Viviroli et al. 2009) need to be calibrated against observed streamflow and are difficult to be configured for small ungauged areas.

The examples presented in Sects. 2 and 4 demonstrate the potential of early warning based on accumulated precipitation. Thus, a tailored tool based on these principles was developed for southern Switzerland, too. First-order catchments are used as baseline spatial unit to accumulate precipitation. To calculate the areas of the first-order catchments a digital terrain model with a spatial resolution of 200 m has been used and processed following the algorithms presented in Binley and Beven (1992). In the target area, 759 first-order catchments have been isolated (Fig. 5, left).

3.1 Estimation of Return Periods

In the Swiss application, return periods have been estimated following the generalized extreme value (GEV) distribution introduced by Jenkinson (1955). It incorporates three types of commonly used extreme value distributions in a single function, the Gumbel distribution (type I), the Fréchet distribution (type II), and the Weibull distribution (type III) (Martins and Stedinger 2000). The standardization of the three types simplifies the extreme value analysis substantially. No subjective decision is

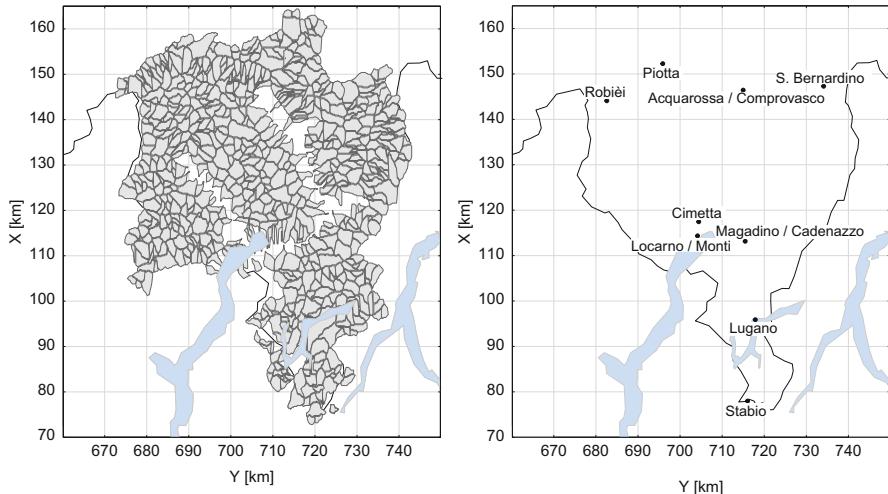


Fig. 5 *Left:* First-order catchments, southern Switzerland (Ticino river basin). 759 first-order catchments exist with an area between 0.02 and 29.7 km^2 . *Right:* Location of the nine MeteoSwiss rain gauge stations with a temporal resolution of 10 min

needed as the data itself defines the matching type of extreme value distribution through the shape parameter (Leadbetter et al. 1983).

Three parameters define the GEV: location μ , scale σ , and shape ξ . It reduces to one of the three distribution types for different ranges of the shape parameter: the Gumbel distribution with $\xi = 0$, the Fréchet distribution with $\xi > 0$, and the Weibull distribution with $\xi < 0$. The cumulative distribution function of the GEV is written as:

$$G(x) = \exp\left\{-\left[1 + \xi\left(\frac{x - \mu}{\sigma}\right)\right]^{-1/\xi}\right\} \text{ for } \xi \neq 0 \quad (7)$$

$$G(x) = \exp\left\{-\exp\left[-\left(\frac{x - \mu}{\sigma}\right)\right]\right\} \text{ for } \xi = 0 \quad (8)$$

where $1 + \xi(x - \mu)/\sigma > 0$ and $-\infty < \mu < \infty$, $\sigma > 0$, and $-\infty < \xi < \infty$ (Coles 2001). The parameters of the GEV were estimated by the maximum likelihood method (Aldrich 1997). As the GEV does not satisfy the regularity conditions required for the usual asymptotic properties associated with the maximum likelihood estimator to be valid, Smith (1985) gave the following guidelines:

- $\xi < 0.5$, the maximum likelihood estimation is valid; it has the usual asymptotic properties.
- $1 > \xi > 0.5$, the maximum likelihood estimation has a result, but the standard asymptotic properties are not fulfilled.

- $\xi > 1$, maximum likelihood estimation is unlikely.

For extreme values $\xi \geq 0.5$ is rare. Therefore, the theoretical restriction of the maximum likelihood method is normally no obstacle in practice (Coles 2001). For the analysis of series of yearly maxima from natural processes shape parameters above 0.5 indicate problems with the data or different processes involved. For analysis, the block maxima approach was chosen. Homogeneity and stationarity were tested by trend analysis using the Mann-Kendall test (Kendall 1970) and the Run test (Wald and Wolfowitz 1940).

3.2 Combined Use of Station Information and Gridded Data for IDF Estimations

Similar to EPIC, also the early warning tool used in southern Switzerland adopts intensity-duration-frequency (IDF) curves. Different data sources are evaluated according to the IDF formulation described in Koutsoyiannis et al. (1998). This IDF equation relates the three indicators related to heavy precipitation: intensity, duration, and return period. IDF are generally created to assess the frequency of a certain rainfall intensity for a certain event duration. Once the relation is established, then a real-time product can be evaluated with respect to its probability of recurrence. In a first step, intensity-duration-frequency (IDF) curves for each catchment have been calculated based on:

- Rain gauge data for the period 1980–2013 from the monitoring network SwissMetNet of MeteoSwiss. Nine stations within the target area observing precipitation with a temporal resolution of 10 min have been used (Fig. 5, right).
- The gridded RhiresD dataset (Schiemann et al. 2010; MeteoSchweiz 2013). RhiresD has a spatial resolution of 1 km and a temporal resolution of 1 day. It has been calculated based on approximately 420 rain gauge measurements covering the entire territory of Switzerland. It is available from 1961 to 2012.
- Thirty-year hindcast of the COSMO-LEPS numerical weather prediction model. The hindcast is available for the time period between 1971 and 2001. The used data has a temporal resolution of 1 day. The grid size is 10 km (Fundel et al. 2010; Jörg-Hess et al. 2015).

The gridded data have been downscaled (nearest-neighbor, Fundel and Zappa 2011) to generate local integral time series for each of the evaluated first-order catchments. Finally, IDF curves have been calculated for each precipitation dataset (single gauges, RhiresD, and COSMO-LEPS) and first-order catchment. A GEV fit (see Sect. 4.1) has been estimated for every duration for which the relation between precipitation intensity and the return period is of interest (from 10 min up to 10 days in the specific case of the early warning system for southern Switzerland).

Since only the single gauges provide information on sub-daily precipitation intensities, a methodology has been developed to combine the gridded datasets

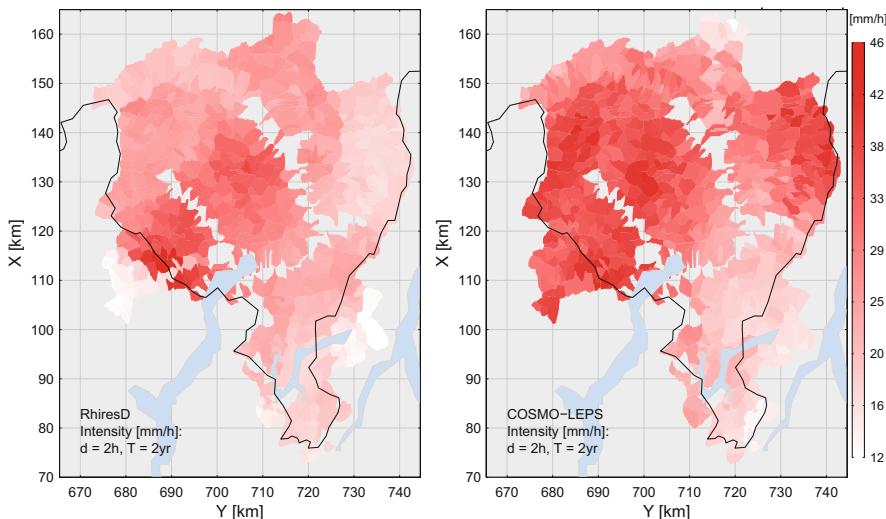


Fig. 6 Catchment-specific rainfall intensities [mm/h] for a duration of 2 h and a return period of 2 years. *Left:* RhiresD. *Right:* COSMO-LEPS

with the information stemming from the ground stations (Knecht 2013). The information of the 10-min rain gauge measurements has been adopted to extend the catchment IDF functions. Knecht (2013) associated each catchment to one of the local rain gauges. The association is realized basing on the similarity between the IDF of the stations and of the sub-areas in case of durations of 1–10 days and return periods of 2–20 years. The assumption is that similar IDF in this interval is a sign that the IDF is similar for all durations and thus the local information provided by the station can be used to integrate the information from the two gridded datasets.

Figure 6 is an example of the resulting spatial intensity distribution based on the gridded precipitation datasets RhiresD (left) and COSMO-LEPS (right) for the duration 2 h and the return period of 2 years. For the RhiresD dataset, the weakest intensities occur in the eastern and southern part of the canton of Ticino. Medium intensities are shown in the northern part and in the central Ticino. The strongest intensities are in the north of the Lago Maggiore. Furthermore, there is a strong intensity gradient between the Italian catchments and the Swiss catchments in the west of the Lago Maggiore.

Concerning the COSMO-LEPS hindcast (right-hand side of Fig. 6), relatively weak intensities occur in the southern Ticino, medium intensities occur in the northern Ticino, and highest intensities occur in the east and western part of the target area.

The spatial distribution of precipitation intensities between the RhiresD and the COSMO-LEPS dataset is clearly different. These differences are due on how these datasets have been elaborated. The COSMO-LEPS hindcast dataset contains the

COSMO-LEPS-related model biases, and the RhiresD contains biases owed to the limitations of the spatial representativity of rain gauges. Furthermore, the strong intensity gradient between the Italian and Swiss catchments in the west of the Lago Maggiore only appear for the RhiresD dataset whereas there is no such gradient for the COSMO-LEPS dataset. The resulting IDF equations for the Italian catchments are erroneous due to missing data in the RhiresD dataset outside Switzerland. To a weaker extent, the same phenomenon occurs in the southeast where Italian sub-catchments show also a weaker intensity compared to the neighboring Swiss sub-catchments for the RhiresD dataset. This limitation could be eliminated only if a homogenous transnational dataset would be available.

In the presented application, the availability of separate IDF analysis for different data sources allows switching between different IDFs in the case of real-time operations. Alerts based on observations are triggered by comparison to the RhiresD IDF, while alerts for the next days are triggered by comparison to the COSMO-LEPS IDF. This reduces the problem of inhomogeneity and bias between the raw COSMO-LEPS output and the observation-based products (Fundel et al. 2010).

3.3 Operational Implementation Forced by Deterministic and Ensemble NWP

The obtained IDFs associated to COSMO-LEPS and RhiresD long time series are the core of the deployed early warning system. Real-time forecasts and gridded observations are obtained from MeteoSwiss. Three datasets are used:

- COSMO-LEPS forecast. COSMO-LEPS is a Limited-Area Ensemble Prediction System developed and run by the Consortium for Small-Scale Modeling (Marsigli et al. 2005). This probabilistic numerical weather forecast model has 16 members. It is available for lead times up to 132 h at a spatial resolution of 7 km (Addor et al. 2011; Zappa et al. 2008). Forecasts are initialized at 12:00 UTC and are delivered approximately 10 h later. Thus, analyses are completed only for 120 h starting from 00:00 UTC of the next day, 12 h after initialization.
- COSMO-2 is a high-resolution realization of COSMO-LEPS and includes explicit calculation of small-scale convection. The grid resolution of the model is 2.2 km and the forecast lead time is 33 h with a temporal resolution of 1 h. A new forecast is calculated every 3 h (Zappa et al. 2008; Ament et al. 2011).
- CombiPrecip is the observation-based dataset used within the Swiss flash flood early warning tool. This dataset combines rain gauge with radar measurements. The aim of this dataset is to combine the quantitatively accurate gauge measurement data with the radar measurement that covers a large area (Sideris et al. 2014). It is operationally available at a temporal resolution of 1 h. Different accumulations can be computed.

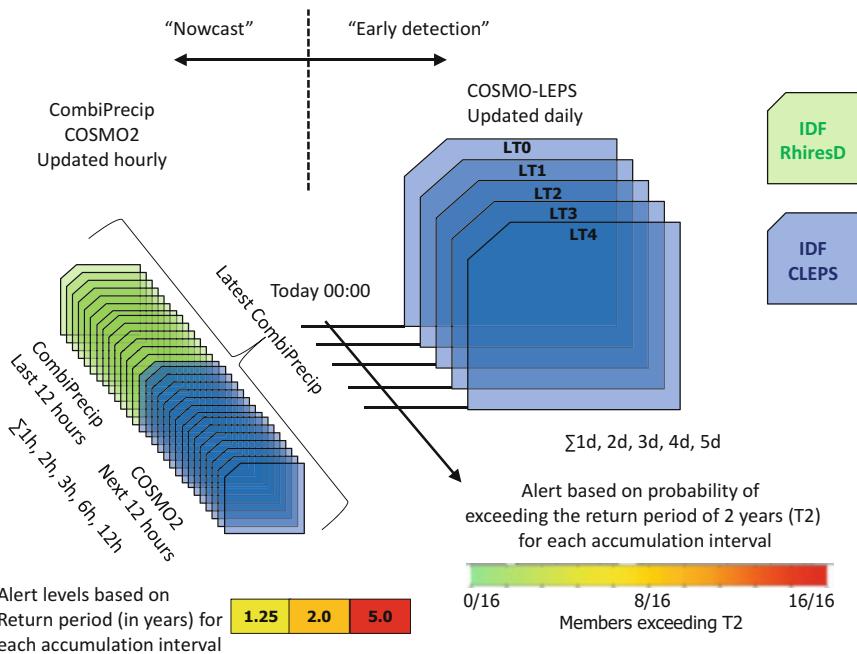


Fig. 7 Sketch of the real-time Swiss flash flood early warning tool consisting of an early detection component (right part) and a nowcast component (left side). See main text for further information

In operational mode, COSMO-2 is nudged to real-time data of a weather radar precipitation obtained from CombiPrecip.

Figure 7 presents the setup of the real-time tool consisting of an “early detection” component based on COSMO-LEPS, which is evaluated against the COSMO IDF for accumulated rainfall of 1–5 days. A second “nowcast” component consists of the combination of the latest 12 hourly fields of CombiPrecip and the forecasted precipitation fields of COSMO-2 for the next 12 h. The accumulated rainfall is evaluated for different intervals (from 1 to 12 h) and depending on the source of the data either the IDF curves of RhiresD or the one based on COSMO-LEPS is used. The nowcast component is updated each hour, while the early detection component runs once every day.

Knechtl (2013) evaluated this setup against observed events. These events are either discharge peaks in gauged sub-areas or reports of damages caused by flash flood events. The hypothesis that it is possible to detect hydrological events with the flash flood early warning tool could be partly confirmed. The highest skill is obtained if the return period of CombiPrecip is assessed at hourly time scale. With this, it was possible to confirm most of the damage events that occurred in 2010 and 2011. The prototype tool is affected by several false alarms. This is because initial conditions of the soils are not considered.

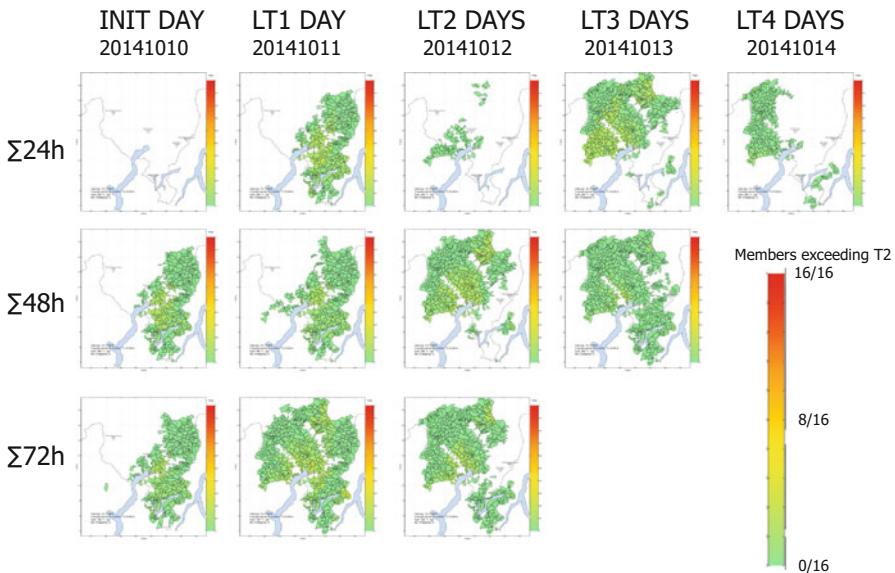


Fig. 8 Early detection of flash floods in southern Switzerland on the basis of operational COSMO-LEPS daily rainfall fields. The examples display alert based on the probability of exceeding the return period of 2 years (T2) for each accumulation interval (24, 48 and 72 h). The situation relates to a forecast issued on 10 October 2014 and indicating possible floods for the next 4 days (see text for additional information)

3.4 Case Study in October 2014

The early warning system presented in Fig. 7 is operational since July 2013. In the period between the start of operations and December 2014 damages related to floods (Hilker et al. 2009) have been reported for more than 20 calendar days (Andres et al. 2015). Numerous damage events occurred on 13 October 2014. In this section, we present the situation as seen by the tool for this particular day. Figure 8 shows the outcomes of the early detection component as available on 10 October 2015 (about 72 h ahead of the event). The analysis of the probabilistic COSMO-LEPS rainfall fields indicates that there is locally some probability to exceed the 2-year return period (T2) for 24, 48, and 72 accumulated rainfall. The 24-h maps indicate that in the northwest areas the intense rainfall should already occur on 11 October 2014. In the southeastern areas, the main rainfall has to be expected for 13 October 2014. The 48-h maps present a band extending from southwest to northeast with probabilities of about 30% (five to six COSMO-LEPS members) to exceed T2 for the period 12–13 October 2014. A similar interpretation can be achieved when inspecting the map of 72-h accumulated rainfall for the period 11–13 October 2014.

The event itself was characterized by a period of intense precipitation between 01:00 UTC and 13:00 UTC on 13 October 2014. Figure 9 presents three different

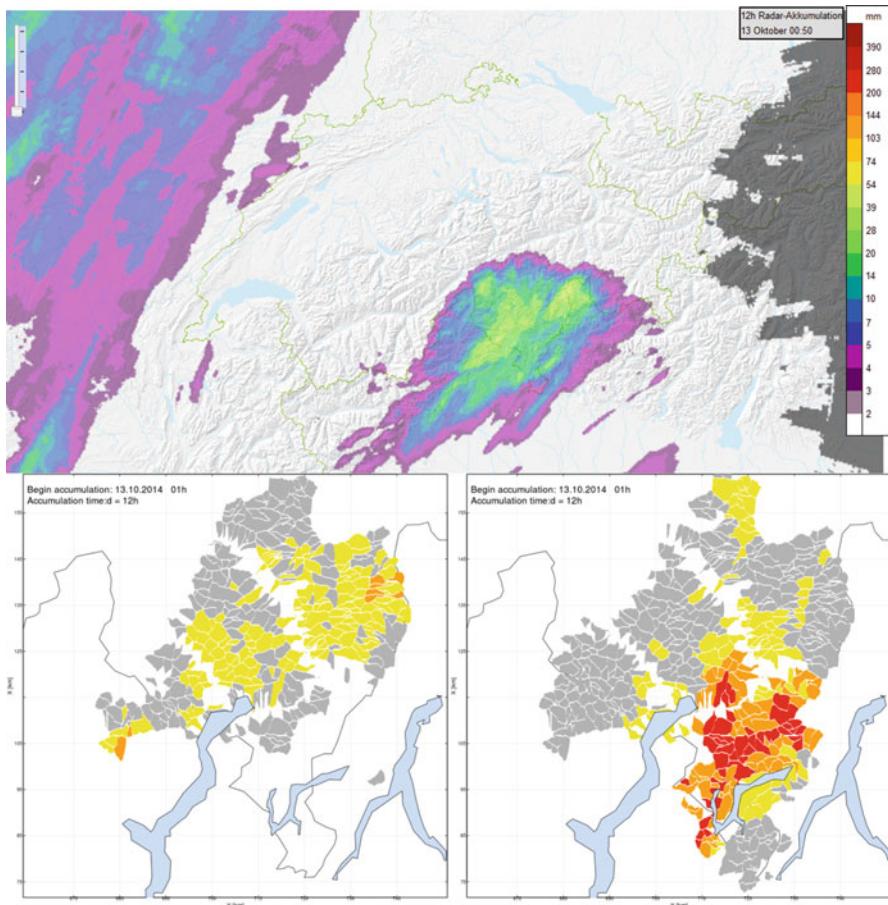


Fig. 9 Heavy precipitation event in southern Switzerland on 13 October 2014. Visualization of analyses of accumulated precipitation between 01:00 UTC and 13:00 UTC. *Top panel*: accumulated rainfall from the weather radar QPE of MeteoSwiss. *Bottom left panel*: evaluation of the CombiPrecip accumulated rainfall. *Bottom right panel*: evaluation of the COSMO-2 forecast available at 20:00 UTC of 12 October 2014. See also Fig. 7 and the text for further information

visualizations of rainfall information during this 12-h interval. The top panel (source MeteoSwiss via <https://www.gin.admin.ch/>, last accessed on 19 August 2015; Heil et al. 2014) presents the accumulated rainfall from the weather radar QPE for a duration of 12 h. The lower left panel presents the evaluation of the CombiPrecip accumulated rainfall during the same time interval under consideration of the RhiresD IDF (Fig. 6, left). The lower right panel of Fig. 10 illustrates the evaluation of the COSMO-2 forecast available at 20:00 UTC of 12 October 2014 (5 h prior to the beginning of the relevant accumulation interval). The return period of the COSMO-2 accumulated rainfall was assigned after comparison to the COSMO-LEPS IDF (Fig. 13, right).

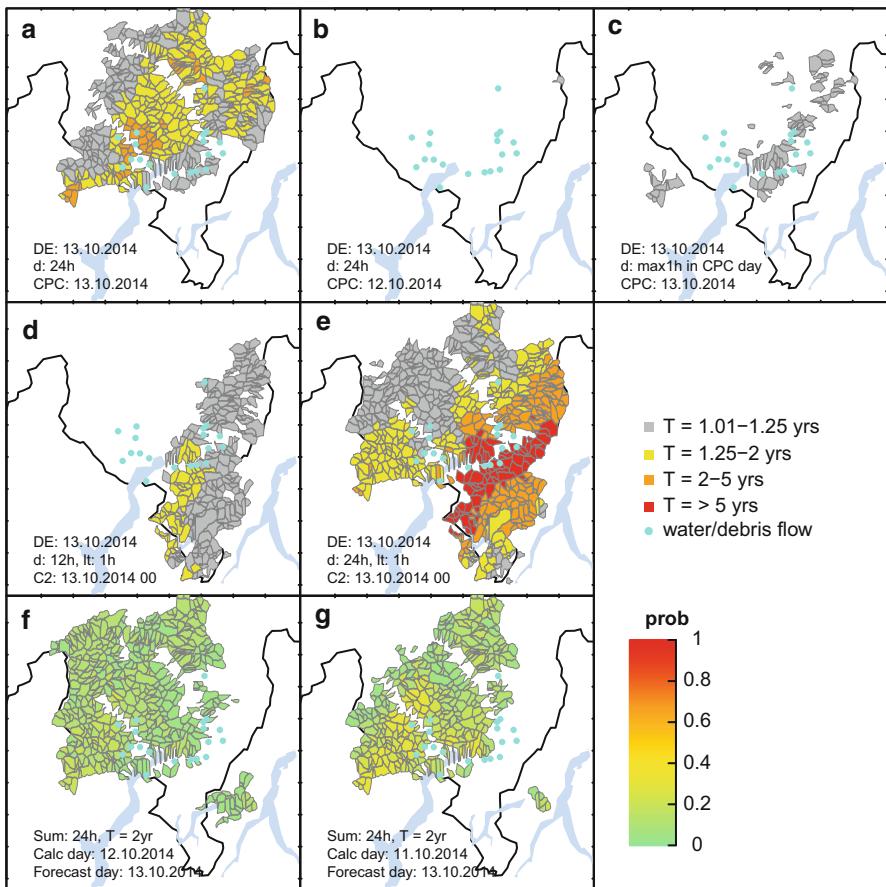


Fig. 10 Visualization of locations affected by floods and debris flow (blue dots) and correspondent alerts issued by the flash flood early warning tool for southern Switzerland on 13 October 2014. *Top row*: alerts related to CombiPrecip accumulated rainfall. *Central row*: alerts related to COSMO-2 rainfall predictions. *Bottom row*: probabilistic alerts based on COSMO-LEPS. See Fig. 7 and the text for full information

Since CombiPrecip is a “post-processed” version of the radar QPE, the general shape of the evaluated CombiPrecip rainfall and of the radar accumulated QPE is very similar. The early warning tool highlights several spots with high precipitation intensities. These spots are within an ellipse stretching from southwest to northeast. This is very similar to the indications obtained from the early prediction obtained on 10 October 2014 by evaluating COSMO-LEPS (Fig. 9). The COSMO-2 prediction presents a very high rainfall intensity exceeding the return period of 5 years in many areas of the target region. The location of the spots with highest intensity is about 30 km more south than the observed fields (radar QPE and CombiPrecip). Nevertheless, the information of

COSMO-2 was available with some hours of advance and could have been used to anticipate the event and trigger mitigation measures.

All flood and debris flow events occurring in Switzerland are documented in a database collecting information from newspapers and online sources (Hilker et al. 2009). Andres et al. (2015) evaluated the damages occurred in the target area on 13 October 2014 and provided the exact coordinates of the locations affected by damages. This information has been used to roughly evaluate the potential of the early warning tool (Fig. 10).

In Fig. 10, the damage locations are plotted within the field of seven flash flood indicators (panels (a) to (g)):

- (a) Return period of 24-h accumulated CombiPrecip precipitation for the event day (13 October 2014)
- (b) Return period of 24-h accumulated CombiPrecip precipitation for the day antecedent to the event
- (c) Maximal return period of hourly rainfall intensity obtained from CombiPrecip during the event day
- (d) Return period of 12-h accumulated COSMO-2 precipitation forecast for the run started at 00:00 of the event day
- (e) Return period of 24-h accumulated COSMO-2 precipitation forecast for the run started at 00:00 of the event day
- (f) Probability of exceeding a return period of 2 years for 24-h accumulated rainfall on the event day as obtained from the 16-member forecast of COSMO-LEPS delivered 1 day ahead
- (g) Probability of exceeding a return period of 2 years for 24-h accumulated rainfall on the event day as obtained from the 16-member forecast of COSMO-LEPS delivered 2 days ahead

The inspection of Fig. 10 indicates that both the rainfall observed 1 day ahead (Fig. 10b) and the maximal 1-hour intensity of CombiPrecip during the event (Fig. 10c) would have been poor indicators for locating the damage spots. The 24-h cumulated CombiPrecip rainfall (Fig. 10a) seems to be better correlated to the locations affected by damages. This might indicate that the events are not triggered by local high-intensity events. The cause for the event is most probably due to long-lasting rainfall, as also indicated by the early prediction component of the tool (Fig. 9). Figure 11 also indicates that the event is characterized by high-intensity 24-h accumulated rainfall. The COSMO-2 evaluation (Fig. 10e) well identifies the regions that were finally affected by damages.

The probabilistic early predictions with the tool based on COSMO-LEPS present higher probability of high 24-h rainfall on the event day in the case of the forecast delivered 2 days in advance (Fig. 10g), while the evaluation available 1 day ahead (Fig. 10f) shows lower probability of exceeding a 2-year return period for 24-h cumulated rainfall in the areas affected by damages according to the newspaper reports collected by Andres et al. (2015).

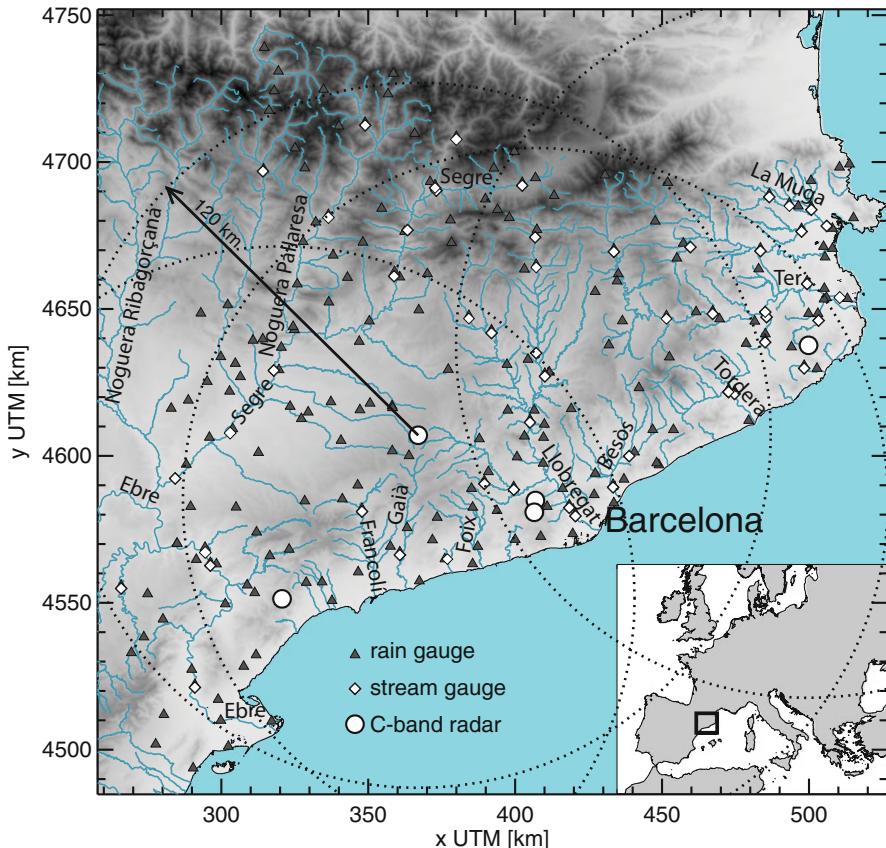


Fig. 11 Network of hydrometeorological sensors available in Catalonia: The *circles* indicate the location of the 4 C-band radars of XRAD and the C-band radar of the Spanish Agency of Meteorology (AEMET); the *gray triangles* and *white diamonds* show the location of the rain gauges and stream-level sensors, respectively

3.5 An Outlook to Future Flash Flood Early Warning in Switzerland

An initial assessment of the value of the early warning tool demonstrated the potential of precipitation-based indexes for flash flood prediction. The current situation concerning the tool implemented in southern Switzerland can be summarized in four points:

- Real-time collection of forecasts is still ongoing since July 2013. The next step will be the verification of this first period of real-time operations.
- The realization of a second prototype in northern Switzerland is planned.

- Further efforts will focus on using probabilistic extrapolation of weather radar QPE (Mandapaka et al. 2012; Foresti and Seed 2014) for hydrological applications (Liechti et al. 2013a).
- Finally, real-time hydrological information as obtained from the application of high-resolution distributed models should be used to estimate initial conditions and reduce the number of false alarms observed in this prototype application (Knecht 2013). This will make use of concepts linked to the mapping of dominant runoff processes (Antonetti et al. 2015).

4 Flash Flood Detection in Catalonia

The flash flood forecasting system adopted in Catalonia (Corral et al. 2009; Alfieri et al. 2011; Versini et al. 2014) is based on rainfall estimates and forecasts generated from weather radar observations. This system, named FF-EWS, is designed as a tool for monitoring the hazard of intense rainfall situations in the context of a flash flood early warning system.

Radar rainfall inputs depict the evolution of the rainfall field with a resolution (of the order of 1 km and 5–10 min) adapted to monitoring the precipitation phenomena that produce flash floods. Also, extrapolation of radar rainfall observations has been successfully used for forecasting the evolution of the rainfall field for a few hours.

With these rainfall inputs, flash flood hazard assessment is based on the assumption that the rainfall accumulated upstream of a point of the drainage network (i.e., the basin-aggregated rainfall) can be used to characterize the flash flood hazard, especially for high return periods, when the pdf of discharges and the pdf of precipitation tend to have the same slope (Guillot and Duband 1967).

4.1 Probabilistic Rainfall Inputs

Quantitative Precipitation Estimates (QPEs) and Quantitative Precipitation Forecasts (QPFs) are produced from weather radar observations with the Integrated Tool for Hydrometeorological Forecasting (Corral et al. 2009). The use of probabilistic rainfall inputs allows us to characterize the uncertainty in QPE and QPF and assess its impact in the estimated hazard. Most of the existing techniques to generate radar-based probabilistic rainfall products are based on the ensemble approach (e.g., Bowler et al. 2006; Llort et al. 2008; Germann et al. 2009; Villarini et al. 2009; Berenguer et al. 2011; Panziera et al. 2011; Quintero et al. 2012): they produce a number of realistic rainfall scenarios (members) compatible with radar observations and that at the same time respect the spatial and temporal structure of the rainfall field. The latter is a crucial aspect to properly assess the impact of rainfall uncertainty on hazard assessment.

Ensuring the quality of the QPE maps is fundamental to guarantee the good performance of the system. This requires processing radar observations with a chain

of algorithms to reduce the effect of the sources of uncertainty affecting radar QPE (e.g., Corral et al. 2009; Villarini and Krajewski 2009). In the Integrated Tool for Hydrometeorological Forecasting (Corral et al. 2009), the production of QPE maps includes (1) mitigating the effects of the interception of the radar beam with the terrain (Delrieu et al. 1995), (2) eliminating non-meteorological echoes (Berenguer et al. 2005; Park and Berenguer 2015), (3) identifying precipitation types in volumetric radar data, (4) extrapolating elevated radar observations to the surface with a vertical profile of reflectivity that depends on the type of precipitation (as described by Franco et al. 2006, 2008), and (5) converting reflectivity into rain rate using a relationship also adapted to the type of precipitation. From instantaneous rainfall maps, rainfall accumulations are computed with a resolution of $1 \times 1 \text{ km}^2$ considering the motion of the precipitation systems and the evolution of rainfall intensities between consecutive radar scans.

Radar QPE ensembles (EQPE) are obtained by perturbing the deterministic QPE with the method of Llort et al. (2008), which considers the space-time structure of errors affecting radar QPE. Each member of the EQPE is a possible realization of the unknown precipitation field given radar measurements.

Deterministic rainfall nowcasting is based on Lagrangian extrapolation (i.e., advection of most recently observed rainfall map with the estimated motion field, neglecting the evolution of rainfall intensities), which has proven to generate useful rainfall forecasts for lead times up to a few hours (see, e.g., Germann et al. 2006; Berenguer et al. 2005, 2012). It is composed of two modules for:

- Rainfall tracking: The algorithm implemented to estimate the motion field of precipitation is based on matching three rainfall maps within 24 min with a modified version of the COTREC algorithm (Li et al. 1995).
- Extrapolation of rainfall observations: The last observed rainfall field is advected in time according to the motion field estimated with the mentioned tracking technique. The motion field is kept stationary in time along the series of generated forecasts.

The uncertainty in rainfall nowcasting is characterized with the SBMcast technique (Berenguer et al. 2011). It generates an ensemble of realistic future rainfall scenarios that evolve from the most recent QPE field, assuming the String of Beads model (Pegram and Clothier 2001) to characterize the space-time variability of the rainfall field.

4.2 Flash Flood Hazard Assessment

Hazard assessment uses estimated and forecasted 30-min rainfall accumulations. This accumulation period is thought to be relevant at point scale (e.g., for urban drainage or in sensitive points of the road network).

For each point of the drainage network, the rainfall inputs available at a given time are used to compute the basin-aggregated rainfall accumulated over a duration

corresponding to the concentration time of the catchment (the computations are made for durations between 0.5 and 24 h and for catchments between 4 and 2000 km²).

Hazard assessment (expressed in terms of probability of occurrence or as return period) is based on comparing the computed rainfall accumulations with the values of the available intensity-duration-frequency (IDF) curves. In the case of basin-aggregated rainfall, point IDF values are reduced with a scaling factor that depends on the area of the drained catchment. Hazard assessment is recalculated every time a new QPE map is available with a resolution of 1 × 1 km² and for lead times between t + 0 and t + 3 h.

Because of the simplicity of this flash flood hazard assessment approach the results sometimes do not reproduce what could be obtained with a system based on a complete rainfall-runoff simulation. The simplification of relating the probability of occurrence of rainfall with the probability of occurrence of discharges neglects some hydrological variables that have an important role in the catchment response (such as the initial moisture state of the catchment or the presence of accumulated snow). On the other hand, the main advantage of the implemented approach is that it does not use parameters that require calibration. This is an important advantage for an operational implementation over large domains, where the aim of the system is to detect flash flood events in small and medium catchments that are often ungauged.

4.3 Implementation in Catalonia (NE Spain) and Case Studies

The FF-EWS is used operationally for flash flood hazard assessment at the control center of the Catalan Water Agency (Barcelona, Spain) for monitoring the evolution of rainfall situations that might lead to flash floods in Catalonia (NE of Spain – see Fig. 11).

In this region, the littoral and pre-littoral mountain ranges (approximately parallel to the coast) act as natural barriers to the warm and humid air from the Mediterranean Sea favoring the genesis of convective processes that lead to intense rains. Yearly rainfall accumulations range from 400 to 1200 mm, but some individual events contribute significantly to the yearly totals (the 10-year return period daily accumulation exceeds 100 mm, and almost every year accumulations over 200 mm in 24 h are recorded somewhere in the Spanish Mediterranean coast). The response times of the mountainous catchments are rather short due to the steep slopes of the streams and the urbanization of the flood plains especially near the coast (where some ephemeral streams have become streets, which increases the risk of flash floods). These factors combined with heavy rainfall events are the ingredients that lead to flash floods in this region.

Operationally, the FF-EWS uses the observations of the rain gauges and stream-level sensors of the Automatic Hydrological Information System and of the four C-band radars of the XRAD (the radar network of the Meteorological Service of Catalonia – see Fig. 11). These are processed with the Integrated Tool for

Hydrometeorological Forecasting to generate the radar-based QPE and QPF ensembles. The IDF curves applied to estimate of the return period of the measured rainfall are those used for river planning and in flooding studies at the Catalan Water Agency (ACA 2003).

4.3.1 Case Study: 12–14 September 2006

This case was a typical autumn event during which several mesoscale convective systems crossed Catalonia from southeast to northwest. The maximum rain gauge accumulations were reported in Constantí (near Tarragona) with 267 mm and in the area of the Gulf of Roses (north of Girona) with 256 mm, 216 mm in 24 h (Fig. 12a).

This event caused significant material losses (intense flooding in urban areas in the regions of Tarragona and Barcelona and failure of the road and railway networks due to flooding and landslides) and one casualty. The intense rainfall produced flash floods in ephemeral torrents near the coast and in some sub-basins of the main rivers as, for example, in the lower part of the rivers Llobregat (5000 km^2) and La Muga (850 km^2).

The estimated rainfall accumulation map for 13 September 2006 is shown in Fig. 12b. The comparison between radar QPE and the available rain gauge records shows a reasonable agreement, except for the largest accumulations (probably more affected by attenuation of the radar signal due to intense rain).

Figure 12c shows the summary of the maximum hazard level estimated at each point of the drainage network throughout the event and how the system was able to successfully detect the importance of the event and identify the areas most affected by intense rainfall and flash floods (marked with red dashed circles in Fig. 12c; see also Fig. 11 of Barnolas et al. 2008). Figure 13 shows the variety of hazard assessment products generated by the system every time a new radar QPE map is available: besides the hazard assessment based on radar observations (Fig. 13a), the system forecasts the expected evolution of the hazard level based on radar QPF (the hazard level forecasted on 13 September 2006 at 15:00 UTC for a lead time of 2 h is shown in Fig. 13b). Additionally, the system estimates the uncertainty in the forecasted hazard level in terms of probability to exceed a given return period (Fig. 13c, d show, respectively, the probability to exceed a return period of 2 and 5 years for the case shown in Fig. 13b). As presented in Sect. 4.1, this product describes how the uncertainty in rainfall QPE and QPF affects the hazard level estimated with the FF-EWS system.

4.3.2 Case Study: 02 November 2008

On 02 November 2008 in the early morning, a convective system coming from the Mediterranean moved toward the interior of Tarragona. The intense rain and strong winds caused significant damages in the coastal area. Some flash floods occurred in the streams near the coastal city of Salou (the closest rain gauge recorded 50 mm between 01:00 and 03:00 UTC). Also, the Segre river flooded some areas in the region of Lleida. The maximum rain gauge accumulation exceeded 130 mm in the village of Prades, with over 90 mm recorded on 02 November 2008 between 02:00 and 05:00 UTC. In the area of Girona, the rainfall was more sustained along the day,

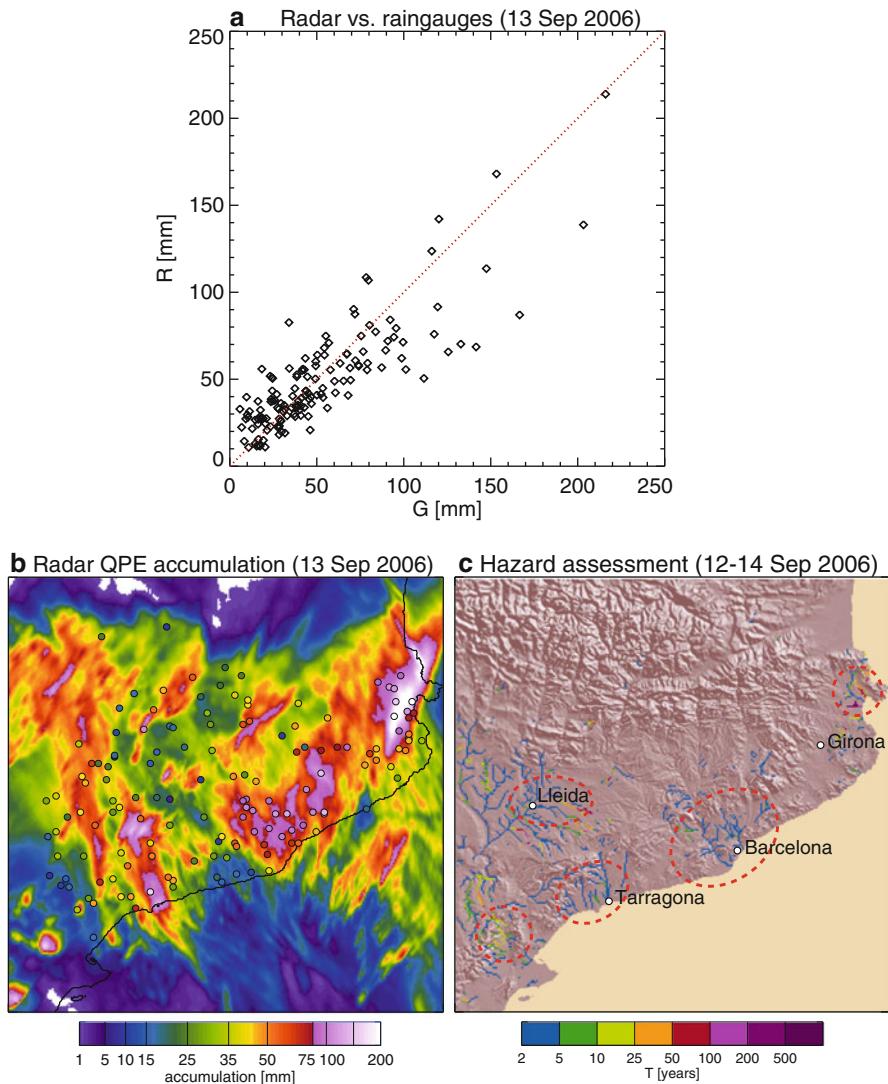


Fig. 12 Results obtained with the radar-based hazard assessment system during the event of 12–14 September 2006. (a) Scatterplot of radar QPE (R) versus rain gauge measurements (G) and (b) radar QPE accumulation for 13 September 2006 (the circles indicate rain gauge accumulations); (c) hazard assessment for the entire event (the red dashed ellipses indicate the areas where floods were reported)

and maximum accumulations reached up to 125 mm, resulting in numerous calls to the fire brigades for flooding in cities like Girona or Olot.

The radar QPE accumulation shows a good correspondence with collocated rain gauge measurements (Fig. 14a, b). Figure 14c shows that the system identified

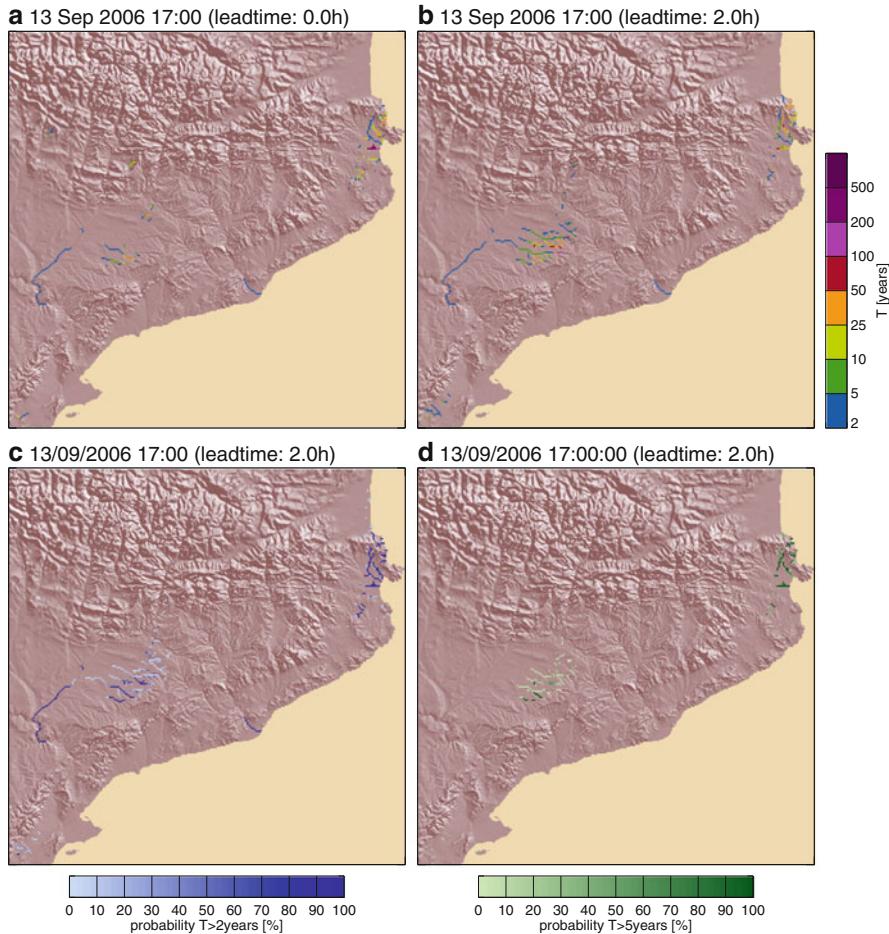


Fig. 13 Hazard assessment for 13 September 2006 at 17:00 UTC: (a) obtained radar rainfall observations; (b) obtained for a lead time of 2 h. (c) and (d) Probability of exceeding a return period of 2 and 5 years, respectively, as estimated from probabilistic rainfall forecasts with a lead time of 2 h

significant hazard in the main spots where floods occurred. In the zone of Tarragona, the area affected by large rainfall accumulations is a 15-km wide band along which an intense convective cell developed and propagated. The system was able to identify significant hazard level in the torrents around Salou, where the most damaging flash floods occurred (see Figs. 14c and 15a). Figure 15b–e show the evolution of the hazard map forecasted for 06:00 UTC. The first signal of possible hazard in this area was detected 1.5 h ahead (i.e., with the forecast generated at 04:30). Thirty minutes later, the hazard map forecasted with a lead time of 1 h is almost identical to what was finally diagnosed from radar observations (Fig. 15a).

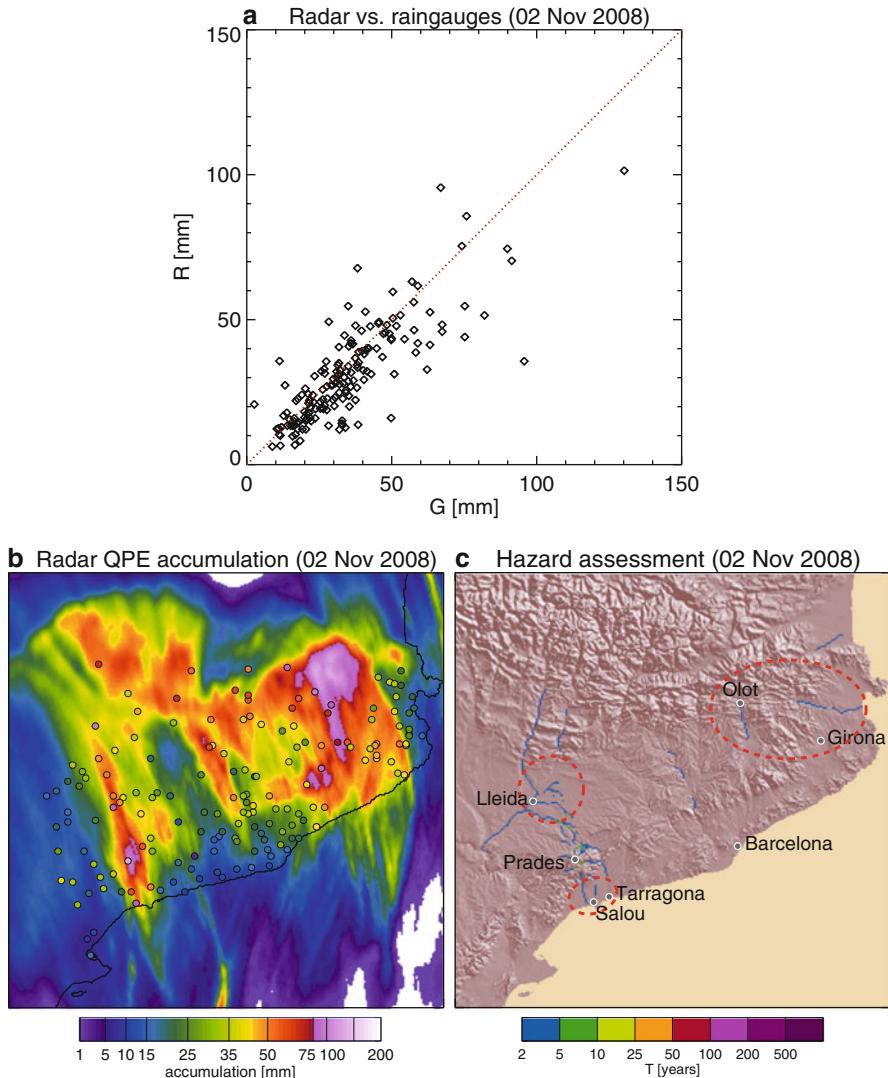


Fig. 14 Same as Fig. 13, but for the event of 02 November 2008

For this case, the European Flood Awareness System's flash flood component (EPIC; see Sect. 2), forecasted the possibility of a significant event in the coastal streams of Tarragona almost 3 days in advance (based on the NWP forecasts produced on 31 September 2008 at 12:00 UTC). However, the differences in the exact location where the different members of the NWP Ensemble Prediction System used in EPIC forecasted the intense rainfall event resulted in a wide region with some (low) probability of flash flood occurrence (Alfieri et al. 2011).

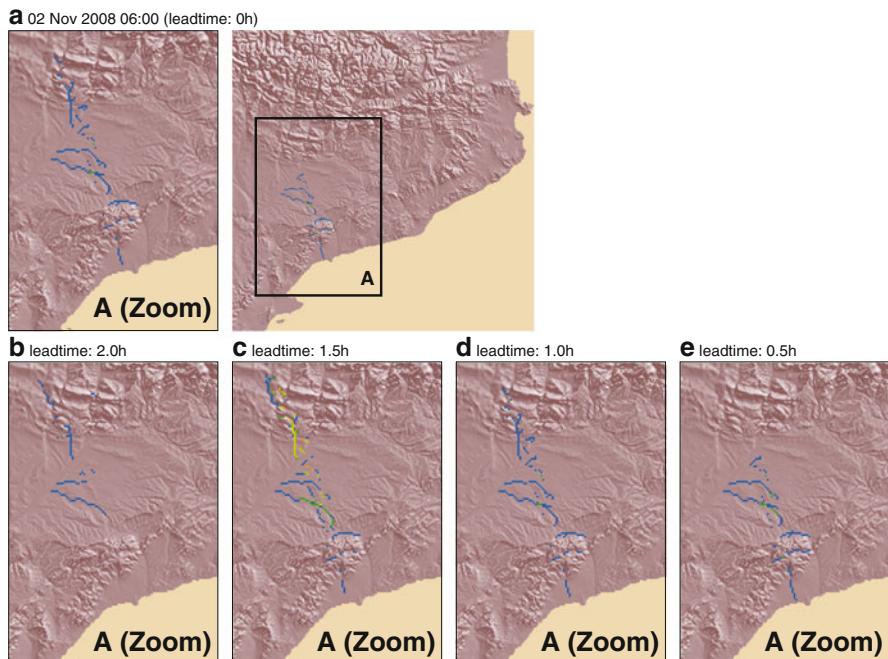


Fig. 15 (a) Hazard estimated on 02 November 2008 at 06:00 UTC. (b–e) Evolution of the hazard assessment product for lead times of 2.0, 1.5, 1.0, and 0.5 h

This case illustrates the complementarity of a system like EPIC (based on NWP forecasts) and the radar-based flash flood hazard system (with a time horizon of a few hours): once a system like EPIC has determined the possibility of flash floods over a certain area a few days ahead, radar-based QPE and QPF allow us to increase the resolution of hazard assessment in the context of flash flood monitoring.

4.3.3 Case Study: 17–19 June 2013

This event caused significant losses all throughout the central Pyrenees (both in the southern and the northern sides). In Catalonia, the northwestern counties were the most affected, with serious flooding in the head of the Garona and Noguera Pallaresa catchments, in the Pyrenees (with mountains above 3000 m amsl). Many of the mountain torrents and the rivers Garona and Noguera Pallaresa overflowed and flooded several villages producing important damages: some roads collapsed, the bridges of Salardú, Arties, and Llavorsí collapsed, several houses were destroyed by the waters in Arties and Bossòst, and 400 people were evacuated (see <http://www.catalannewsagency.com/society-science/item/severe-floods-in-the-north-western-catalan-pyrenees>). In the county of Val d’Aran, the total damages were estimated in 100 M €. Daily rainfall accumulations reached 115 mm in Vielha (the capital of this county), with maximum rainfall intensities of 12 mm in 30 min.

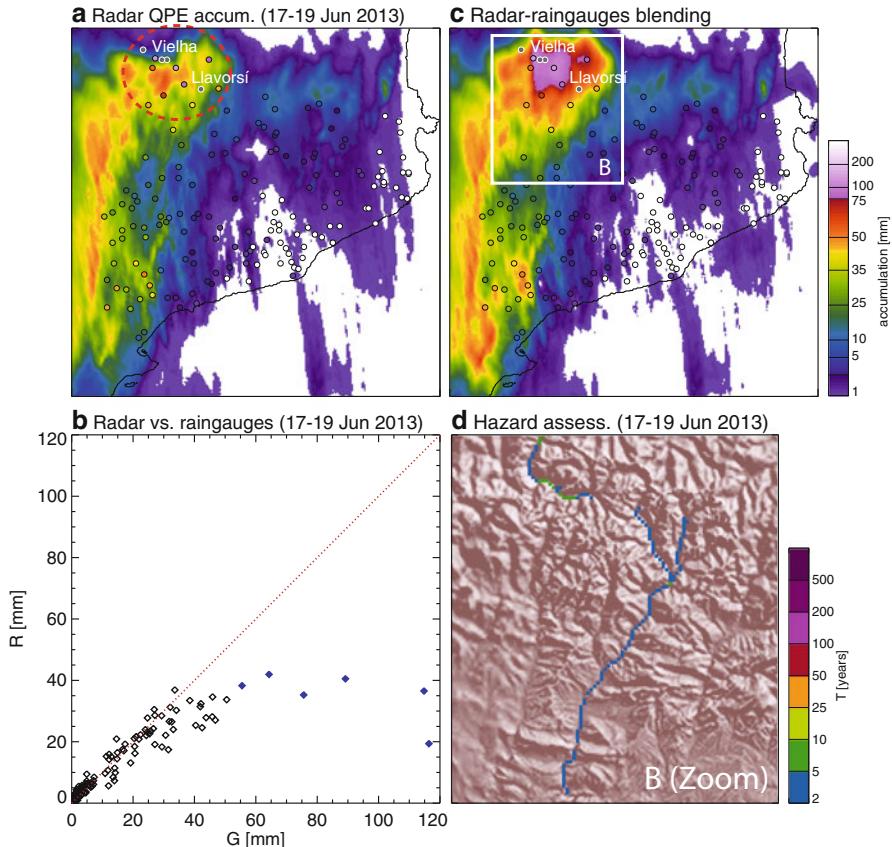


Fig. 16 Results obtained with the radar-based hazard assessment system during the event of 17–19 June 2013. **(a)** Radar QPE accumulation between 17 June 2013 at 20:00 UTC and 19 June 2013 at 24:00 UTC; the red dashed ellipse indicates the location where radar QPE significantly underestimated precipitation and where floods occurred (the black dots correspond to the villages mentioned in the text) (from west to east, Salardú, Vielha, Arties, Bossòst and Llavorsí). **(b)** Radar-rain gauge blending accumulation. **(c)** Scatterplot of radar QPE versus rain gauge measurements (the blue diamonds correspond to the locations where radar QPE significantly underestimated rainfall accumulations). **(d)** Hazard assessment for the entire event based on radar-rain gauge blending in the domain indicated in panel **(b)**

This event is presented to illustrate two of the limitations of the presented hazard assessment system:

1. The location of the affected area falls beyond the limits of what can be considered the maximum range for radar Quantitative Precipitation Estimation (as can be seen in Figs. 11 and 16, the closest radar is located over 100 km away) and is a very mountainous area affected by visibility problems from the radar perspective due to beam blockage and beam

overshooting (e.g., Pellarin et al. 2002). Consequently, the quality of radar QPE in these areas is in general rather poor and underestimates rain gauge accumulations (see Fig. 16a, b).

2. Although this was quite a significant rainfall event, the severe response of the basin was strongly influenced by the accelerated melting of a good part of the snow accumulated in the mountains at the end of spring (after the event, the snow depth recorded in stations located above 2200 m showed a reduction of up to 700 mm).

These factors resulted in serious underestimation of the hazard level in this area using radar-based QPE (not shown here). To improve the quality of radar QPE in such mountainous areas, several authors have proposed to use small X-band radars to fill the gaps of regional radar networks (e.g., Beck and Bousquet 2013; Campbell and Steenburgh 2014). When these small radars are not available (as in the case of the Catalan Pyrenees), the estimated QPE field can be improved by blending radar QPE maps with rain gauge observations (e.g., Velasco-Forero et al. 2009; Schiemann et al. 2011; Sideris et al. 2014 and references therein). These are constrained by rain gauge observations and reproduce the structure of the rainfall field as depicted by radar. The blended radar-rain gauge QPE (obtained with the technique of Velasco-Forero et al. 2009) has been implemented to recalculate the hazard level along the event. Figure 16c, d shows, respectively, the resulting total rainfall accumulation and the summary of the maximum hazard level estimated along the event. Now, the areas most affected by floods show significant flood hazard level (above a return period of 5 years in some parts of the drainage network). However, the damages indicate that the real magnitude of the event was significantly higher (the return period was most likely above 25 years), with an important role of snow melting (not considered in the current version of the hazard assessment system).

4.4 An Outlook to Future Flash Flood Early Warning in Catalonia

The hazard assessment system presented here uses radar-based probabilistic QPE and QPF inputs. The system is used in real time in the control center of the Catalan Water Agency for monitoring the evolution of rainfall events that might lead to flash floods. Some of its outputs are also disseminated to the general public through the website “Water in Real Time” (AETR – <http://aca-web.gencat.cat/aetr/aetr2/>, last accessed on 17 July 2016).

The main advantage of the system is its high resolution (1 km and 6 min), which enables very precise determination of the areas at risk at the expense of shorter lead times (in Catalonia, up to 3 h). Consequently, extending this time horizon with NWP models with lead times beyond 1 day is necessary to enable earlier preparedness and trigger effective emergency and response plans. The best practice is to use the radar-based FF-EWS in the context of monitoring potentially hazardous rainfall situations as detected with systems based on NWP (as suggested by Alfieri et al. 2011; Versini et al. 2014).

Current work in progress focuses on extending the coverage of the system for flash flood hazard assessment in the context of the EC Civil Protection project European Demonstration of a Rainfall- and Lightning-Induced Hazard Identification Nowcasting Tool (EDHIT, www.edhit.eu, last accessed on 17 July 2016). This project is implementing a similar system at European scale based on the European radar mosaics generated by the EUMETNET project OPERA (Huuskonen et al. 2014). The goal of EDHIT is to demonstrate the potential of the European radar mosaic for rainfall-induced hazard assessment in the context of Civil Protection. A prototype of the system has been operating since mid-2012, and the first analyses show promising results. Figure 17 shows an example of the regional hazard assessment based on radar

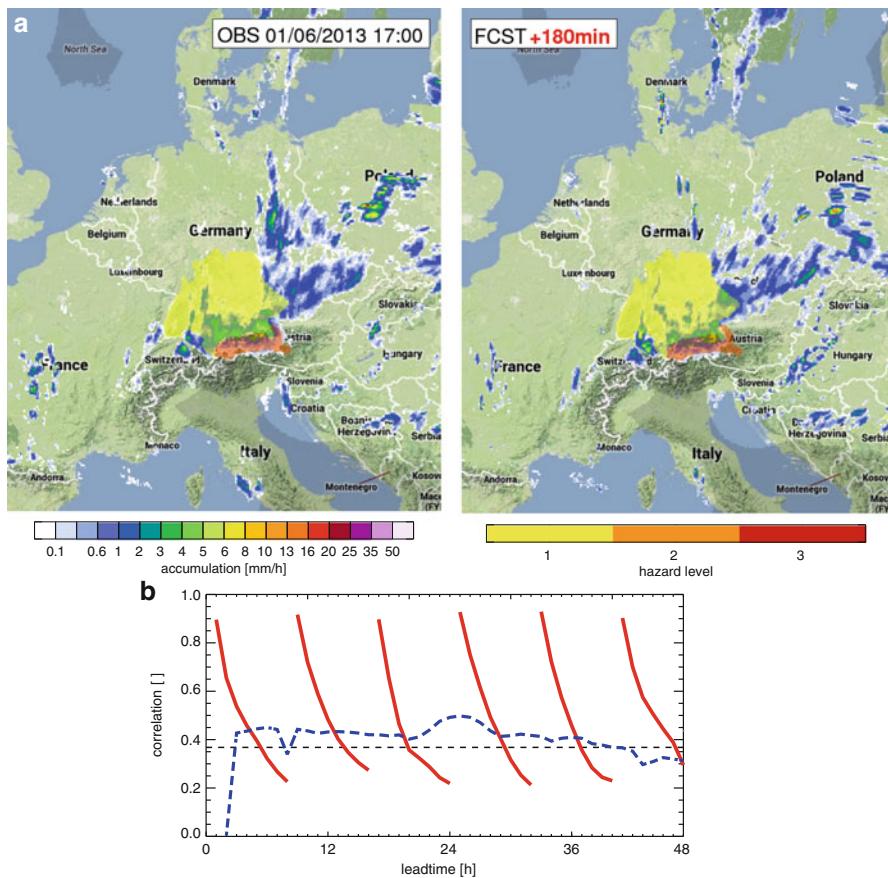


Fig. 17 (a) Hazard level overlaid on hourly rainfall accumulations for 01 June 2013 at 17:00 obtained from radar observations (*left*) and from rainfall forecasts with a lead time of 3 h (*right*). (b) Correlation between hourly accumulations estimated from the OPERA radar mosaic and rainfall forecasts obtained with the HIRLAM run corresponding to 01 June 2013 at 00:00 UTC (*dashed blue line*) and with the analyzed nowcasting system run on 01 June 2013 at 00:00, 08:00, 16:00 UTC and on 02 June 2013 at 00:00, 08:00 UTC (*red lines*)

observations during the event of May 2012 in central Europe, compared with the results obtained with an operational numerical weather prediction system.

In parallel, current research focuses on how to account for soil moisture conditions in hazard assessment: the chosen approach is based on adapting the rainfall thresholds used for hazard assessment to the soil moisture conditions depicted with the rainfall-runoff model LISFLOOD (van der Knijff et al. 2010), applied operationally at European scale in the context of EFAS.

5 Conclusions

Three early warning systems for flash flood early warning have been presented in this chapter. The systems show a number of similarities though making different uses of probabilistic information. EPIC and the Swiss tool use probabilistic forecasts of COSMO-LEPS and consistent hindcast datasets for obtaining spatially distributed information on possible upcoming floods. The Catalan FF-EWS system makes full use of ensemble weather radar QPE and QPF, while the Swiss tool uses an advanced weather radar product combined to high-resolution numerical weather predictions for the nowcasting of flash floods.

A clear challenge in flash flood forecasting based on NWP is the accurate prediction of the location and timing of extreme storms, given their features of small-scale, short, and intense events. Probabilistic approaches and ensemble forecasting can help us address this issue, so that if some chance of extreme events is predicted, regional systems are triggered to monitor the evolution of the event as it approaches and develops, making use of more detailed information of regional monitoring networks and of the expertise of local flood forecasters.

Interestingly, ongoing developments of all the three systems are aimed to integrating initial soil moisture conditions and to better characterize the timing and magnitude of the flood events.

All three systems present and target applications in areas extending across different countries. Problems of data homogeneity are often not easy to address, particularly when merging data from different national and regional networks and in the blending of point observations with gridded output from numerical models. EPIC is not affected by such issue as it is based on modeled precipitation from a single data product. On the other hand, both the Swiss tool and the large-scale application of the Catalan approach might suffer from using transnational data. Hence, the implementation of these latter applications in different areas requires a careful and tailored design based on the local data availability.

References

- ACA, *Recomanacions tècniques per als estudis d'inundabilitat d'àmbit local* (Agència Catalana de l'Aigua, Barcelona, 2003), p. 106
- N. Addor, S. Jaun, F. Fundel, M. Zappa, An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* **15**, 2327–2347 (2011). <https://doi.org/10.5194/hess-15-2327-2011>

- J.R.A. Aldrich, Fisher and the making of maximum likelihood 1912–1922. *Stat. Sci.* **12**(3), 162–176 (1997)
- L. Alfieri, J. Thielen, A European precipitation index for extreme rain-storm and flash flood early warning. *Meteorol. Appl.* **22**(1), 3–13 (2015). <https://doi.org/10.1002/met.1328>
- L. Alfieri, D. Velasco, J. Thielen, Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. *Adv. Geosci.* **29**, 69–75 (2011). <https://doi.org/10.5194/adgeo-29-69-2011>
- L. Alfieri, J. Thielen, F. Pappenberger, Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland. *J. Hydrol.* **424–425**, 143–153 (2012). <https://doi.org/10.1016/j.jhydrol.2011.12.038>
- L. Alfieri, F. Pappenberger, F. Wetterhall, The extreme runoff index for flood early warning in Europe. *Nat. Hazards Earth Syst. Sci.* **14**(6), 1505–1515 (2014). <https://doi.org/10.5194/nhess-14-1505-2014>
- F. Ament, T. Weusthoff, M. Arpagaus, Evaluation of MAP D-PHASE heavy precipitation alerts in Switzerland during summer 2007. *Atmos. Res.* **100**(2–3), 178–189 (2011)
- N. Andres, A. Badoux, C. Hegg, Unwetterschäden in der Schweiz im Jahre 2014. Rutschungen, Murgänge, Hochwasser und Sturzereignisse. *Wasser Energie Luft* **107**(1), 47–54 (2015)
- M. Antonetti, R. Buss, S. Scherrer, M. Margreth, M. Zappa, Mapping dominant runoff processes: an evaluation of different approaches using similarity measures and synthetic runoff simulations, *Hydrol. Earth Syst. Sci. Discuss.* **12**, 13257–13299 (2015). <https://doi.org/10.5194/hessd-12-13257-2015>
- M. Barnolas, A. Atencia, M.C. Llasat, T. Rigo, Characterization of a Mediterranean flash flood event using rain gauges, radar, GIS and lightning data. *Adv. Geosci.* **17**, 35–41 (2008). <https://doi.org/10.5194/adgeo-17-35-2008>
- J.C. Bartholmes, J. Thielen, M.H. Ramos, S. Gentilini, The European flood alert system EFAS – part 2, statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.* **13**(2), 141–153 (2009)
- J. Beck, O. Bousquet, Using gap-filling radars in mountainous regions to complement a national radar network: improvements in multiple-doppler wind syntheses. *J. Appl. Meteorol. Climatol.* **52**, 1836–1850 (2013). <https://doi.org/10.1175/JAMC-D-12-0187.1>
- M. Berenguer, C. Corral, R. Sanchez-Diezma, D. Sempere-Torres, Hydrological validation of a radar-based nowcasting technique. *J. Hydrometeorol.* **6**, 532–549 (2005). <https://doi.org/10.1175/JHM433.1>
- M. Berenguer, D. Sempere-Torres, G.G.S. Pegram, SBMcast – an ensemble nowcasting technique to assess the uncertainty in rainfall forecasts by Lagrangian extrapolation. *J. Hydrol.* **404**, 226–240 (2011). <https://doi.org/10.1016/j.jhydrol.2011.04.033>
- M. Berenguer, M. Surcel, I. Zawadzki, M. Xue, F. Kong, The diurnal cycle of precipitation from continental radar mosaics and numerical weather prediction models, part II: intercomparison among numerical models and with nowcasting. *Mon. Weather Rev.* **140**, 2689–2705 (2012). <https://doi.org/10.1175/mwr-d-11-00181.1>
- A. Binley, K. Beven, Three dimensional modelling of hillslope hydrology. *Hydrol. Process.* **6**(3), 253–368 (1992)
- N.E. Bowler, C.E. Pierce, A.W. Seed, STEPS: a probabilistic precipitation forecasting scheme which merges an extrapolation nowcast with downscaled NWP. *Q. J. Roy. Meteorol. Soc.* **132**, 2127–2155 (2006). <https://doi.org/10.1256/qj.04.100>
- L.S. Campbell, W.J. Steenburgh, Finescale orographic precipitation variability and gap-filling radar potential in little Cottonwood Canyon, Utah. *Weather Forecast.* **29**, 912–935 (2014). <https://doi.org/10.1175/WAF-D-13-00129.1>
- V.T. Chow, D.R. Maidment, L.W. Mays, *Applied Hydrology*. (McGraw-Hill Science/Engineering/Math., McGraw-Hill: New York. ISBN 0-07-010810-2, 1988)
- S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, vol. 208 (Springer, London, 2001)
- C.G. Collier, Flash flood forecasting: what are the limits of predictability? *Q. J. Roy. Meteorol. Soc.* **133**(622), 3–23 (2007)
- C. Corral, D. Velasco, D. Forcadell, D. Sempere-Torres, Advances in radar-based flood warning systems. The EHMI system and the experience in the Besòs flash-flood pilot basin, in *Flood*

- Risk Management: Research and Practice*, ed. by P. Samuels, S. Huntington, W. Allsop, J. Harrop (Taylor & Francis, London, 2009), pp. 1295–1303
- D.P. Dee, S.M. Uppala, A.J. Simmons, P. Berrisford, P. Poli, S. Kobayashi, U. Andrae, M.A. Balmaseda, G. Balsamo, P. Bauer, P. Bechtold, A.C.M. Beljaars, L. van de Berg, J. Bidlot, N. Bormann, C. Delsol, R. Dragani, M. Fuentes, A.J. Geer, L. Haimberger, S.B. Healy, H. Hersbach, E.V. Hólm, L. Isaksen, P. Kållberg, M. Köhler, M. Matricardi, A.P. McNally, B.M. Monge-Sanz, J.-J. Morcrette, B.-K. Park, C. Peubey, P. de Rosnay, C. Tavolato, J.-N. Thépaut, F. Vitart, The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q. J. Roy. Meteorol. Soc.* **137**(656), 553–597 (2011). <https://doi.org/10.1002/qj.828>
- G. Delrieu, J.D. Creutin, H. Andrieu, Simulation of radar mountain returns using a digitized terrain model. *J. Atmos. Oceanic Tech.* **12**, 1038–1049 (1995). [https://doi.org/10.1175/1520-0426\(1995\)012<1038:SORMRU>2.0.CO;2](https://doi.org/10.1175/1520-0426(1995)012<1038:SORMRU>2.0.CO;2)
- M. Fiorentino, F. Rossi, P. Villani, Effect of the basin geomorphoclimatic characteristics on the mean annual flood reduction curve. *Proc. IASTED Int. Conf. Model. Simul.* **5**, 1777–1784 (1987)
- L. Foresti, A. Seed, The effect of flow and orography on the spatial distribution of the very short-term predictability of rainfall from composite radar images. *Hydrol. Earth Syst. Sci.* **18**, 4671–4686 (2014). <https://doi.org/10.5194/hess-18-4671-2014>
- M. Franco, R. Sánchez-Diezma, D. Sempere-Torres, Improvements in weather radar rain rate estimates using a method for identifying the vertical profile of reflectivity from volume radar scans. *Meteorol. Z.* **15**, 521–536 (2006). <https://doi.org/10.1127/0941-2948/2006/0154>
- M. Franco, R. Sánchez-Diezma, D. Sempere-Torres, I. Zawadzki, Improving radar precipitation estimates by applying a VPR correction method based on separating precipitation types, 5th European Conference on Radar in Meteorology and Hydrology (Helsinki, 2008), P14.16
- J. French, R. Ing, S. Von Allmen, R. Wood, Mortality from flash floods: a review of national weather service reports, 1969–81. *Public Health Rep.* **98**(6), 584 (1983)
- F. Fundel, M. Zappa, Hydrological ensemble forecasting in mesoscale catchments: sensitivity to initial conditions and value of reforecasts. *Water Resour. Res.* **47**, W09520 (2011). <https://doi.org/10.1029/2010WR009996>
- F. Fundel, A. Walser, M.A. Liniger, C. Appenzeller, Calibrated precipitation forecasts for a limited-area ensemble forecast system using reforecasts. *Mon. Weather Rev.* **138**(1), 176–189 (2010)
- E. Gaume, V. Bain, P. Bernardara, O. Newinger, M. Barbuc, A. Bateman, L. Blaškovičová, G. Blöschl, M. Borga, A. Dumitrescu, I. Daliakopoulos, J. Garcia, A. Irimescu, S. Kohnova, A. Koutoulis, L. Marchi, S. Matreata, V. Medina, E. Preciso, D. Sempere-Torres, G. Stancalie, J. Szolgay, I. Tsanis, D. Velasco, A. Viglione, A compilation of data on European flash floods. *J. Hydrol.* **367**(1–2), 70–78 (2009)
- K.P. Georgakakos, Analytical results for operational flash flood guidance. *J. Hydrol.* **317**(1), 81–103 (2006)
- U. Germann, I. Zawadzki, B. Turner, Predictability of precipitation from continental radar images, part IV: limits to prediction. *J. Atmos. Sci.* **63**, 2092–2108 (2006). <https://doi.org/10.1175/JAS3735.1>
- U. Germann, M. Berenguer, D. Sempere-Torres, M. Zappa, REAL-Ensemble radar precipitation estimation for hydrology in a mountainous region. *Q. J. Roy. Meteorol. Soc.* **135**, 445–456 (2009). <https://doi.org/10.1002/qj.375>
- P. Guillot, D. Duband, La méthode du Gradex pour le calcul de la probabilité des crues à partir les pluies. *AISH Publ.* **84**, 560–569 (1967)
- A. Günther, M. Van Den Eeckhaut, J.P. Malet, P. Reichenbach, J. Hervás, The European landslide susceptibility map ELSUS 1000 Version 1, EGU General Assembly Conference Abstracts, 15, 10071 (2013), <http://adsabs.harvard.edu/abs/2013EGUGA..1510071G>. Last accessed 19 Aug 2015
- B. Heil, I. Petzold, H. Romang, J. Hess, The common information platform for natural hazards in Switzerland. *Nat. Hazards* **70**(3), 1673–1687 (2014)
- N. Hilker, A. Badoux, C. Hegg, The Swiss flood and landslide damage database 1972–2007. *Nat. Hazards Earth Syst. Sci.* **9**, 913–925 (2009)
- J.R.M. Hosking, L-moments: analysis and estimation of distributions using linear combinations of order statistics. *J. R. Stat. Soc. Ser. B Methodol.* **52**(1), 105–124 (1990)

- A. Huuskonen, E. Saltikoff, I. Holleman, The operational weather radar network in Europe. *Bull. Am. Meteorol. Soc.* **95**, 897–907 (2014). <https://doi.org/10.1175/BAMS-D-12-00216.1>
- P. Javelle, C. Fouchier, P. Arnoud, J. Lavabre, Flash flood warning at un-gauged locations using radar rainfall and antecedent soil moisture estimations. *J. Hydrol.* **394**, 267–274 (2010)
- A.F. Jenkinson, The frequency distribution of the annual maximum (or minimum) values of meteorological elements. *Q. J. Roy. Meteorol. Soc.* **81**(348), 158–171 (1955)
- S.N. Jonkman, Global perspectives on loss of human life caused by floods. *Nat. Hazards* **34**(2), 151–175 (2005)
- S. Jörg-Hess, S. B. Kempf, F. Fundel, M. Zappa, The benefit of climatological and calibrated reforecast data for simulating hydrological droughts in Switzerland, *Meteorological Applications*, **22**(3), 444–458 (2015) <https://doi.org/10.1002/met.1474>
- M.G. Kendall, *Rank Correlation Methods* (Griffin, London, 1970). ISBN 0-85264-199-0
- V. Knechtl, Flash-flood early warning tool. Use of intensity-duration-frequency curves for flash-flood warning in southern Switzerland and forecast skill evaluation, Master Thesis, ETH Zürich, 2013
- D. Koutsoyiannis, D. Kozonis, A. Manetas, A mathematical framework for studying rainfall intensity-duration-frequency relationships. *J. Hydrol.* **206**, 118–135 (1998)
- M.R. Leadbetter, G. Lindgren, H. Rootzén, *Extremes and Related Properties of Random Sequences and Processes* (Springer, New York a.o., XII, 336 pp, 1983)
- L. Li, W. Schmid, J. Joss, Nowcasting of motion and growth of precipitation with radar over a complex orography. *J. Appl. Meteorol.* **34**, 1286–1300 (1995). [https://doi.org/10.1175/1520-0450\(1995\)034<1286:NOMAGO>2.0.CO;2](https://doi.org/10.1175/1520-0450(1995)034<1286:NOMAGO>2.0.CO;2)
- K. Liechti, L. Panziera, U. Germann, M. Zappa, The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrol. Earth Syst. Sci.* **17**, 3853–3869 (2013a). <https://doi.org/10.5194/hess-17-3853-2013>
- K. Liechti, M. Zappa, F. Fundel, U. Germann, Probabilistic evaluation of ensemble discharge nowcasts in two nested Alpine basins prone to flash floods. *Hydrol. Process.* **27**, 5–17 (2013b). <https://doi.org/10.1002/hyp.9458>
- X. Llort, C.A. Velasco-Forero, J. Roca-Sancho, D. Sempere-Torres, Characterization of uncertainty in radar-based precipitation estimates and ensemble generation, 5th European Conference on Radar in Meteorology and Hydrology (Helsinki, 2008)
- P.V. Mandapaka, U. Germann, L. Panziera, A. Hering, Can Lagrangian extrapolation of radar fields be used for precipitation nowcasting over complex Alpine orography? *Weather Forecast.* **27**, 28–49 (2012). <https://doi.org/10.1175/WAF-D-11-00050.1>
- C. Marsigli, F. Boccanera, A. Montani, T. Paccagnella, The COSMO-LEPS mesoscale ensemble system: validation of the methodology and verification. *Nonlinear Processes Geophys.* **12**(4), 527–536 (2005)
- E.S. Martins, J.R. Stedinger, Generalized maximum-likelihood generalized extreme-value quantile estimators for hydrologic data. *Water Resour. Res.* **36**(3), 737–744 (2000)
- Meteoschweiz, Documentation of MeteoSwiss Grid-Data Products – Daily Precipitation (final analysis): RhiresD (2013), http://www.meteoschweiz.admin.ch/content/dam/meteoswiss/de/service-und-publikationen/produkt/raeumliche-daten-niederschlag/doc/ProdDoc_RhiresD.pdf. Last accessed 13 July 2015
- D. Norbiato, M. Borga, S. Esposti, E. Gaume, S. Anquetin, Flash flood warning based on rainfall thresholds and soil moisture conditions: an assessment for gauged and ungauged basins. *J. Hydrol.* **362**, 274–290 (2008)
- L. Panziera, U. Germann, M. Gabella, P.V. Mandapaka, NORA–Nowcasting of Orographic Rainfall by means of Analogues. *Q. J. Roy. Meteorol. Soc.* **137**, 2106–2123 (2011). <https://doi.org/10.1002/qj.878>
- S. Park, M. Berenguer, Adaptive reconstruction of radar reflectivity in clutter-contaminated areas by accounting for the space–time variability. *J. Hydrol.* **520**, 407–419 (2015). <https://doi.org/10.1016/j.jhydrol.2014.11.013>
- G.G.S. Pegram, A.N. Clothier, High-resolution space-time modelling of rainfall: the “String of Beads” model. *J. Hydrol.* **241**, 26–41 (2001). [https://doi.org/10.1016/S0022-1694\(00\)00373-5](https://doi.org/10.1016/S0022-1694(00)00373-5)
- T. Pellarin, G. Delrieu, G.M. Saulnier, H. Andrieu, B. Vignal, J.D. Creutin, Hydrologic visibility of weather radar systems operating in mountainous regions: case study for the Ardeche catchment

- (France). *J. Hydrometeorol.* **3**, 539–555 (2002). [https://doi.org/10.1175/1525-7541\(2002\)003<0539:hvows>2.0.co;2](https://doi.org/10.1175/1525-7541(2002)003<0539:hvows>2.0.co;2)
- F. Quintero, D. Sempere-Torres, M. Berenguer, E. Baltas, A scenario-incorporating analysis of the propagation of uncertainty to flash flood simulations. *J. Hydrol.* **460–461**, 90–102 (2012). <https://doi.org/10.1016/j.jhydrol.2012.06.045>
- D. Raynaud, J. Thielen, P. Salamon, P. Burek, S. Anquetin, L. Alfieri, A dynamic runoff co-efficient to improve flash flood early warning in Europe: evaluation on the 2013 central European floods in Germany. *Meteorol. Appl.* **22**(3), 410–418 (2015). <https://doi.org/10.1002/met.1469>
- S. Reed, J. Schaake, Z. Zhang, A distributed hydrologic model and threshold frequency-based method for flood forecasting at ungauged locations. *J. Hydrol.* **337**, 402–420 (2007)
- R. Schiemann, M. Liniger, C. Frei, Reduced space optimal interpolation of daily rain gauge precipitation in Switzerland. *J. Geophys. Res.* **115**, D14109 (2010). <https://doi.org/10.1029/2009JD013047>
- R. Schiemann, R. Erdin, M. Willi, C. Frei, M. Berenguer, D. Sempere-Torres, Geostatistical radar-rain gauge combination with nonparametric correlograms: methodological considerations and application in Switzerland. *Hydrol. Earth Syst. Sci.* **15**, 1515–1536 (2011). <https://doi.org/10.5194/hess-15-1515-2011>
- I.V. Sideris, M. Gabella, R. Erdin, U. Germann, Real-time radar-rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Q. J. Roy. Meteorol. Soc.* **140**, 1097–1111 (2014). <https://doi.org/10.1002/qj.2188>
- R.L. Smith, Maximum likelihood estimation in a class of non-regular cases. *Biometrika* **72**(1), 67–90 (1985)
- J. Thielen, J. Bartholmes, M.H. Ramos, A. de Roo, The European flood alert system – part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**(2), 125–140 (2009)
- J.M. van der Knijff, J. Younis, A. de Roo, A GIS-based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.* **24**, 189–212 (2010)
- C.A. Velasco-Forero, D. Sempere-Torres, E.F. Cassiraga, J.J. Gómez-Hernández, A non-parametric automatic blending methodology to estimate rainfall fields from rain gauge and radar data. *Adv. Water Resour.* **32**, 986–1002 (2009)
- P. Versini, M. Berenguer, C. Corral, D. Sempere-Torres, An operational flood warning system for poorly gauged basins: demonstration in the Guadalhorce basin (Spain). *Nat. Hazards* **71**, 1355–1378 (2014). <https://doi.org/10.1007/s11069-013-0949-7>
- A. Viglione, G. Blöschl, On the role of storm duration in the mapping of rainfall to flood return periods. *Hydrol. Earth Syst. Sci.* **13**(2), 205–216 (2009). <https://doi.org/10.5194/hess-13-205-2009>
- G. Villarini, W. Krajewski, Review of the different sources of uncertainty in single polarization radar-based estimates of rainfall. *Surv. Geophys.* **31**, 107–129 (2009). <https://doi.org/10.1007/s10712-009-9079-x>
- G. Villarini, W. Krajewski, G. Ciach, D. Zimmerman, Product-error-driven generator of probable rainfall conditioned on WSR-88D precipitation estimates. *Water Resour. Res.* **45**, W01404 (2009). <https://doi.org/10.1029/2008wr006946>
- D. Vivioli, M. Zappa, J. Gurtz, R. Weingartner, An introduction to the hydrological modelling system PREVAH and its pre- and post-processing-tools. *Environ. Model. Software* **24**(10), 1209–1222 (2009)
- A. Wald, J. Wolfowitz, On a test whether two samples are from the same population. *Ann. Math. Stat.* **11**, 147–162 (1940)
- M. Zappa, M. Rotach, M. Arpagaus, M. Dorninger, C. Hegg, A. Montani, R. Ranzi, F. Ament, U. Germann, G. Grossi, S. Jaun, A. Rossa, S. Vogt, A. Walser, J. Wehrhan, C. Wunram, MAP D-PHASE: real-time demonstration of hydrological ensemble prediction systems. *Atmos. Sci. Lett.* **9**(2), 80–87 (2008)
- M. Zappa, S. Jaun, U. Germann, A. Walser, F. Fundel, Superposition of three sources of uncertainties in operational flood forecasting chains. *Atmos. Res.* **100**(2–3), 246–262 (2011). <https://doi.org/10.1016/j.atmosres.2010.12.005>. Thematic Issue on COST731
- M. Zappa, F. Fundel, S. Jaun, A “Peak-Flow Box” approach for supporting interpretation and evaluation of operational ensemble flood forecasts. *Hydrol. Process.* **27**, 117–131 (2013). <https://doi.org/10.1002/hyp.9521>



Medium Range Flood Forecasting Example

EFAS

Jutta Thielen-del Pozo, Peter Salamon, Peter Burek,
Florian Pappenberger, C. Alionte Eklund, Eric Sprokkereef,
M. Hazlinger, M. Padilla Garcia, and R. Garcia-Sanchez

Contents

1	Introduction	1262
2	EFAS – A Novel Concept for Improving Preparedness for Flooding in Europe	1264
2.1	Hydrological Model and System Setup	1265
2.2	Data and How It Is Used	1267
2.3	Concepts and Methodologies	1269
2.4	Postprocessing	1271

J. Thielen-del Pozo (✉)

European Commission, Joint Research Centre, Ispra, Italy

e-mail: jutta.thielen@jrc.ec.europa.eu

P. Salamon

European Commission, Joint Research Centre (JRC), Institute for Environment and Sustainability (IES), Climate Risk Management Unit, Ispra, VA, Italy

P. Burek

Water Program (WAT), International Institute for Applied System Analysis (IIASA), Laxenburg, Austria

F. Pappenberger

European Centre for Medium-Range Weather Forecast, ECMWF, Reading, UK

C. Alionte Eklund

Swedish Meteorological and Hydrological Institute, Norrköping, Sweden

E. Sprokkereef

Ministry of Infrastructure and the Environment, Water Management Centre of the Netherlands, River Forecasting Service, Lelystad, The Netherlands

M. Hazlinger

Slovak Hydrometeorological Institute, Bratislava, Slovakia

M. P. Garcia

REDIAM, Sevilla, Spain

R. Garcia-Sanchez

ELIMCO SISTEMAS S.L., Sevilla, Spain

2.5 Calculation of Scores	1272
2.6 Training	1274
2.7 Partner Network and Connection to International Initiatives	1274
3 Summary and Conclusion	1274
References	1275

Abstract

Europe repeatedly observes flood events that affect several countries at the same time and which require the coordination of assistance at the European level. The European Flood Awareness System (EFAS) has been developed specifically to respond to the need for forecasting transnational floods with sufficient lead time to allow coordination of aid at the European level in case the national capacities for emergency management are exceeded. In order to achieve robust and reliable flood forecasting at the continental scale with lead times up to 10 days, EFAS promotes probabilistic forecasting techniques based on multiple numerical weather prediction inputs including ensemble prediction systems. Its aim is to complement existing national flood forecasting services with added value information and to provide European decision makers with coherent overviews on ongoing and upcoming floods in Europe for better planning and coordination of aid. To date, EFAS is a unique system providing daily, probabilistic flood forecast information for the entire of Europe on a single platform. Being a complementary system to national ones, EFAS predicts the probabilities for exceeding critical flood thresholds rather than quantitative information on stream flows. By maintaining a dedicated, multinational partner network of EFAS users, novel research could be transferred directly to the operational flood forecasting centers in Europe. EFAS development started in 2003, and the system has become fully operational under the umbrella of Emergency Management Service of the European Copernicus Space Program in 2011.

Keywords

Continental · Ensemble prediction · Early flood warning

1 Introduction

With around 40 cross-border rivers, Europe is particularly exposed to floods which require bilateral or international cooperation for effective management of flood prevention and preparedness measures. Most European rivers cross two to three countries, except the Danube River which is shared with as many as 18 different countries and controlled by many more authorities (<http://www.icpdr.org/main/publications/15-years-managing-danube-basin>).

During the last 20 years, several critical transnational flood events took place in Europe, for example, in the river basins of the Odra (1997), Rhine and Meuse (1993, 1995), and Elbe and Danube (2002, 2006 and 2013). In particular, the 2002 floods in the Elbe and Danube were disastrous and affected eight countries simultaneously (Brazdil et al. 2005; Yiou et al. 2006; Toothill 2002). While many of those countries exceeded

their national capacity to fight against the flooding and to reduce the impacts of the event, the planning and coordination of aid between the countries and in particular at the European level proved to be difficult because decision makers were faced with noncoherent flood warning information from different sources and of variable quality and information level. In fact, despite the high number of transnational river basins and the exposure to floods, fully integrated, transnational flood forecasting systems for European systems are rare, and information exchange on discharge and water level data is still a challenge between cross-border authorities. However, without actually sharing the same forecasting model for upstream and downstream river sections or using the same information platform, appropriate planning is a challenge. Bakker (2007) suggests that transboundary flood events can be particularly hazardous and result in higher number of victims and financial damages than national rivers which can be partially attributed to the lack of integrated information platforms and coherent management plans.

Therefore, the European Commission initiated the development of the European Flood Awareness System (EFAS). The aim of EFAS is to operate a unique forecasting system for the entire of Europe capable of forecasting the probability of severe flood events with sufficient lead time for various decision makers (Thielen et al. 2009; Bartholmes et al. 2009) including experts in flood forecasting authorities and international civil protection services. The system is to provide the national authorities with added value information to their own national and higher resolution systems. Added value for national services could include fully catchment-based simulation including river sections from neighboring countries, flood forecasting information based on different weather forecasts, flood forecasting based on another hydrological model, as well as overview information from neighboring river basins for general information. In case of absence of a national flood forecasting system, as is the case in a few countries, EFAS can serve as unique flood forecasting system. For planning and coordination of international aid in support to the affected countries, the system provides the European civil protection community with unique overview information on ongoing and upcoming flood events. EFAS information is provided on a single platform using the same color codes and thresholds for all river basins in Europe and has become an important planning tool for European decision makers (Croke et al. 2013a). A recent study also evaluated the potential monetary value of the system and found a return of up to €400 for every €1 invested (Pappenberger et al. 2015a).

An important driver for floods is rainfall which can only be forecasted skillfully up to a few days ahead with single (deterministic) forecasts (Buizza et al. 1999). Meteorologists have responded to this limitation arising from numerical representation by using ensemble prediction systems (EPS). While ensemble prediction systems (EPS) were already commonly used and applied in the meteorological community in the 1990s for extending the predictability of forecasts (Buizza et al. 1999), hydrological EPS (HEPS) were still novel in hydrological applications and forecasters only started familiarizing themselves with their use and how to deal with the associated uncertainties (Demeritt et al. 2007; Thielen et al. 2011). Over the last decade, HEPS are increasingly seen as a possibility for extending the predictability for flood events and consequently for improved decision-making (Croke et al. 2013b). They are already an integral part in many operational centers in

particular in Europe (Croke and Pappenberger 2009; Croke et al. 2009; Wetterhall et al. 2013) and represent today the state of the art in forecasting science and related disciplines (Schaake et al. 2006; Thielen et al. 2008a, 2011; Collins and Knight 2007; Croke et al. 2013c; Stephens and Croke 2014).

In this chapter, first the setup of the EFAS system including a brief description of the hydrological model, the technical setup, and meteorological inputs will be provided. Furthermore, a subsection is dedicated to the visualization of EFAS results and their communication to the different end users. Finally, calculation of scores and evaluation of EFAS results on event bases are described. A brief summary and outlook to future development avenues will be given at the end of the chapter.

2 EFAS – A Novel Concept for Improving Preparedness for Flooding in Europe

The development of EFAS from a research project to a fully operational system was a continuous process over a time span of approximately 10 years. The system sparked off from a research project called “European Flood Forecasting System (EFFS, 1999–2002)” (Gouweleeuw et al. 2004; 2005; Pappenberger et al. 2005) where for the first time in Europe, the feasibility of using ensemble prediction system inputs for early flood warning was tested for several European river basins. From 2003 to 2011, the Joint Research Centre of the European Commission brought EFAS to an operational state in collaboration with the European hydrological and meteorological institutions in the Member States and the international research community represented through HEPEX.

During the first 2 years of development, the technical skeleton was set up, data collected, and the model calibrated. From 2005 to 2010, EFAS was being tested in real-time mode and continuously improved and adapted to the needs of the National hydrological services and the European Civil Protection which could consult EFAS information through the EFAS information system (EFAS-IS, launched in 2007), a username and password protected web interface to which partners can connect any time to view EFAS results, provide feedback, and share comments and suggestions with others (www.efas.eu).

While at the beginning of the EFAS project, the emphasis was put on products for large-scale riverine floods in transnational river basins, this started changing once the system had proved its added value. Increasingly, EFAS partners also looked to EFAS for additional information for national and smaller rivers and flash flood events. Subsequently, a variety of indicators for rapid flood generating rainfalls were produced (Alfieri et al. 2012a, b, *this issue*).

By 2011 EFAS was sufficiently developed to be transferred to full operations as part of the Emergency Management Service of the Copernicus Initial Operations (Regulation (EU) No 377/2014 of the European Parliament and of the Council of 3 April 2014 establishing the Copernicus Program and repealing Regulation (EU) No 911/2010) in direct support to European Civil Protection (Decision No 1313/2013/EU of the European Parliament and of the Council of 17 December 2013 on a Union Civil Protection Mechanism). All operational components have been

outsourced to competent institutions in a decentralized manner as illustrated in Fig. 1.

The establishment of a partner network with national, regional, and local flood forecasting centers was key during the development of EFAS and remains an integral part of the system management during operations. Direct contact with partners as well as annual meetings with the representatives from the entire network ensured that the development of EFAS products was in line with the evolving needs of the end users.

The core features of EFAS are described comprehensively in the two EFAS publications by Thielen et al. (2009) and Bartholmes et al. (2009), which provide a good overview of the system. Further publications on specific topics and further developments are numerous and referred to in this chapter, which represents a concise summary of the current state of the art of the system.

2.1 Hydrological Model and System Setup

The core hydrological model of EFAS is the LISFLOOD model, a spatially distributed hybrid between a conceptual and a physical rainfall-runoff model combined with a routing module in the river channel, which has been specifically designed for the simulation of hydrometeorological processes in large European river basins (de Roo et al. 2003; van der Knijff et al. 2008). It is programmed in a dynamic Geographical Information System (GIS) language allowing optimal use of data layers of land use, soil type, depth and texture, and river network (Thielen

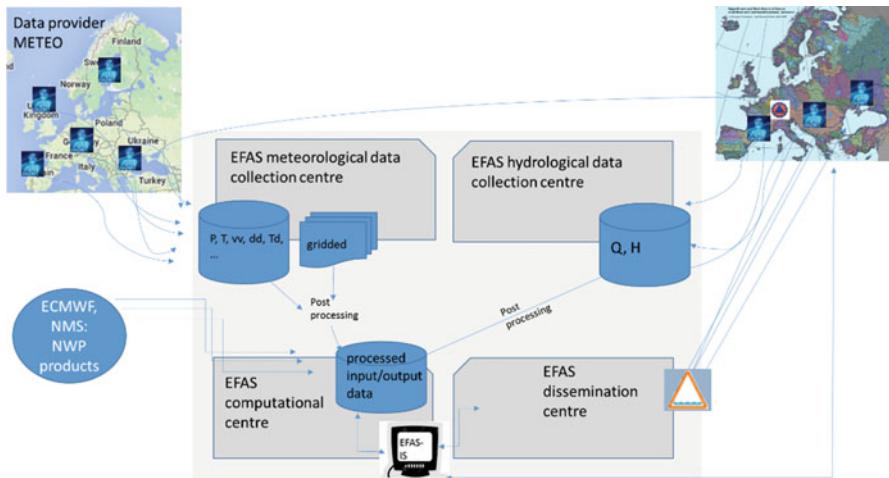


Fig. 1 Illustration of the data and information flow for the operational EFAS. There are four operational centers executing distinct tasks of hydrological and meteorological data collection, computations, postprocessing and visualization, and finally analysis and dissemination of results. Partner networks play major roles as data providers and receivers of EFAS forecast information

et al. 2008b). LISFLOOD has several parameters that require calibration based on observed discharges. Currently 693 stations across Europe are available for the calibration as illustrated in Fig. 2. In case observed discharges are not available, LISFLOOD is run with standard parameters.

For EFAS, LISFLOOD has been set up on a rectangular $5 \times 5 \text{ km}^2$ grid for a geographical domain including all 27 EU countries. Currently, an extension of the

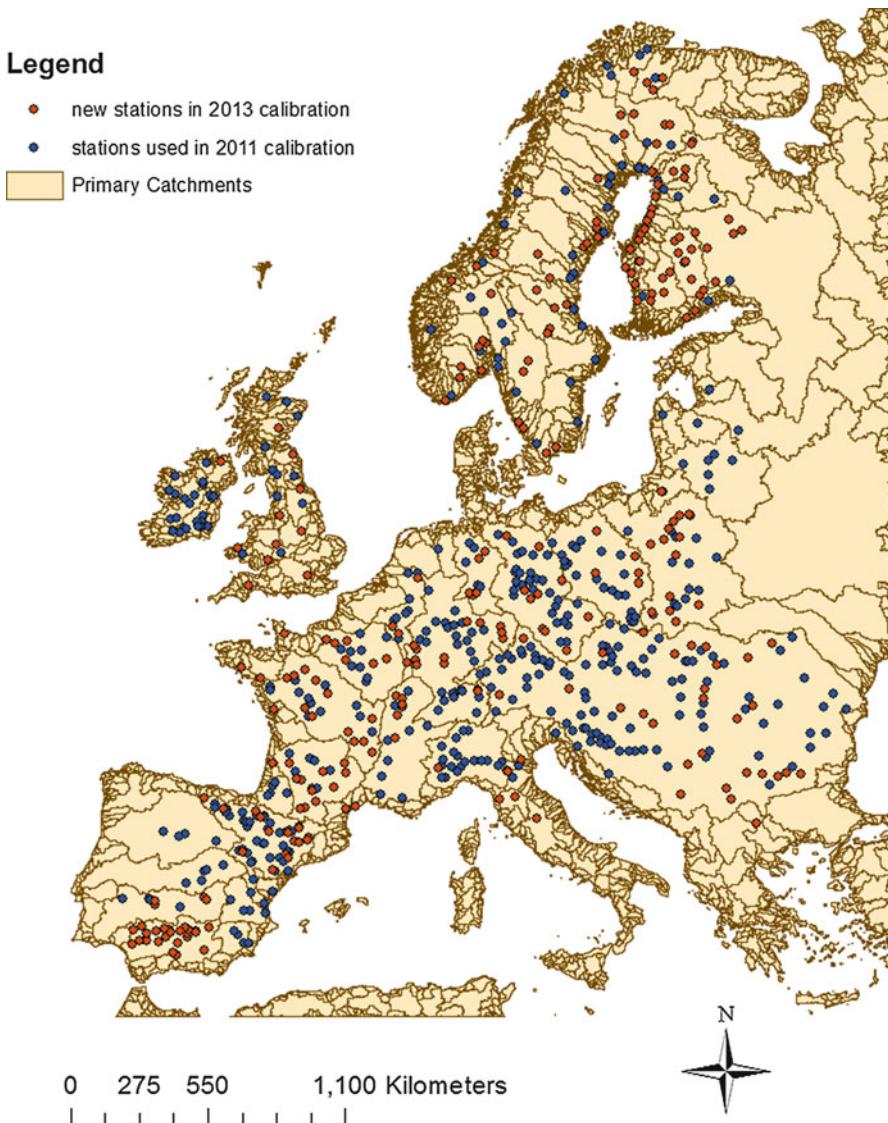


Fig. 2 Distribution of calibration stations used for the calibration conducted in 2011 (blue) and new stations added during the calibration conducted in 2013 (red)

model domain towards Eastern countries is ongoing. Time steps are variable: in order to establish the initial conditions, the model is driven with observed meteorological data with a 24 h time step. The forecasts are driven with various numerical weather prediction (NWP) inputs and executed with either 6 hourly or daily time steps. Experimentally, also hourly time steps for flash flood type applications were implemented (Alfieri et al. 2011, [this issue](#)), in this case combined with a higher grid resolution. In the near future, a multihydrological model setup is also envisaged.

2.2 Data and How It Is Used

2.2.1 Static Data

The LISFLOOD model requires information on the topography such as elevation and slope gradient. General land use information is essential and is complemented with additional maps on fraction of forest and urban areas. Seasonality is accounted for by incorporating monthly information on leaf area index and total water withdrawal demand. Soil is described by soil depth and soil texture for an upper and lower soil layer. The description of the river network is crucial and not always straightforward when upscaling from high resolution to lower resolution in particular for tributaries and rivers at close distance. The channel is defined through maps on width, length, gradient, and slope. Although the model is set up on a $5 \times 5 \text{ km}^2$ grid, information on the river itself is available at higher resolution.

2.2.2 In Situ

For EFAS, both hydrological and meteorological in situ data are required. The data are collected as real-time data for the forecasting applications and as historic time series data for model calibration and validation purposes. After collection of the raw data, the data are submitted to vigorous quality and quantity checks before being aggregated further. Data are typically aggregated to 6 hourly or 24 hourly data and both raw data, flagged data, and aggregated data stored in dedicated databases.

During the calibration and validation phase, the meteorological in situ data are used for calculating long-term simulations which are used to calculate statistics and to study specific flood events. In operational mode, the observed data are used to establish the initial conditions for the next forecast runs. The time from the last available observed data and the next forecast run is filled in with forecasted NWP data.

The observed hydrological data are used for several applications. First of all, with a subset of reliable data the model is calibrated and validated. In real-time mode, all available data are used to compare against the model output for visual evaluation of the model output (and identification of observational errors). For a subset of the stations, the model output is bias and error-corrected to obtain fully probabilistic discharge forecasts (Bogner and Kalas 2008; Bogner and Pappenberger 2011). These postprocessed data are made available for the EFAS partners. Finally, for all stations

where critical discharge or water levels are available, the actual values are compared against the critical thresholds, and where they are exceeded, the station is highlighted with information on time, location, and which critical threshold has been exceeded, as well as a link to the national providers.

2.2.3 Numerical Weather Prediction (NWP)

EFAS promotes the use of ensemble prediction systems (EPS) as well as multimodel inputs to achieve robust and reliable probabilistic outputs (Thielen et al. 2009). NWP products are used from the European Centre for Medium Range Weather Forecasts (ECMWF, deterministic, and EPS), the German Weather Service (DWD, deterministic, dynamically downscaled for days 1–3), and the COSMO-LEPS consortium (COSMO-LEPS, EPS). Products from the UK Meteorological Office (UKMO, EPS) are currently also being tested. Furthermore, all global EPS stored in the Thorpx Interactive Global Grand Ensemble (TIGGE) archive can be fed into EFAS. TIGGE data, so far, have only been applied in research mode and not in operational mode, for example, Pappenberger et al. (2008). Spatial resolutions and lead times in days are illustrated in Table 1.

In EFAS, each ensemble member is treated like a single deterministic forecast. Only after each member has been pushed through the hydrological model, the results are postprocessed to different probabilistic products.

Real-time weather forecasts are used for the operational flood forecasting application twice a day using the 00:00 and the 12:00 forecast. In between time steps such as 03:00 and 09:00 are not included. EFAS forecasts are initiated as soon as a weather forecast is available. Archived forecasts and reforecasts are applied in hindcast studies and to recalculate long-term skill scores (Pappenberger et al. 2011a, b).

Table 1 List of weather forecast products used in EFAS in 2016

Forecast name	Members	Lead time (days)	Spatial resolution (km)	Mode
DWD (German Weather Service)	1	7	~6.5 km (for days 1 to 5) ~15 km (for days 6 to 7)	Operational, twice a day
ECMWF-Deterministic	1	10	~8 km (for days 1–10)	Operational, twice a day
ECMWF – VAREPS	51	15	~18 km	Operational, twice a day
COSMO-LEPS	16	5	~7 km (does not cover most of Scandinavia)	Operational, twice a day
<i>UKMO</i>				<i>Case studies</i>
<i>TIGGE archive including ensembles from at least eight global NWP</i>		<i>Variable:</i> <i>5–30</i>	<i>Variable:</i> 30–110 km	<i>Case studies</i>

2.3 Concepts and Methodologies

By comparing the forecasts to these thresholds, the ensemble streamflow calculations are converted into effective flood forecasts with up to 10 days lead time.

2.3.1 Flood Threshold Exceedance Versus Discharge Forecasts

The role of EFAS is to notify its partners when the system indicates a probability that critical flood levels might be exceeded. This task can be challenging for a continental flood forecasting system which is lacking the spatial and temporal resolution of most national systems, the high-resolution input data to correctly determine the initial conditions, and, most of all, the local knowledge when discharges are becoming critical. Therefore, a different approach needed to be developed. The development team of EFAS opted for a model consistent framework: Using the same model setup and parameterization, long-term discharge simulations are produced from which different return periods are calculated. For EFAS, the 1.5-, 2-, 5-, and 20-year return periods were chosen. These thresholds are calculated for every grid point in the model domain in the same way. When the forecasted discharges are then compared against these thresholds, the streamflow forecasts are effectively converted into dichotomous time series of “flood threshold exceeded”/“flood threshold not exceeded.” This approach has the advantage that systematic over- or underestimations of discharges in the model are compensated for. The only bias remaining then arises from the discrepancies between the observed and forecasted weather inputs. This approach, which has been developed quite early in EFAS, has proved to be very successful as a guidance for decision makers.

Return periods of 1.5 years are classified as “low,” 2 years as “medium,” 5 years as “high,” and 20 years as “severe.” Threshold exceedances also play a role in the visualization and are color coded as green (1.5), yellow (2), red (5), and purple (20).

2.3.2 Incorporating Persistence into Alerting Procedures

The main objective of EFAS is to report large riverine floods, which are typically caused by widespread severe or long-lasting precipitation, snow melting, or rainfalls combined with snow melting. The driving processes for such floods are mostly large synoptic-scale weather phenomena that build up over several days and should therefore be captured repeatedly by the NWP. Therefore, the principle of temporal “persistence” was introduced for notifying EFAS partners: A pixel is flagged as “risk of flooding” only if the discharges in that river pixel exceed the EFAS high or EFAS severe flood threshold in at least three consecutive 12-hourly forecasts. This is equivalent to lagged forecasts. It has been shown that by introducing a criterion of persistence in flood forecasting, the forecast reliability increases (Bartholmes et al. 2009).

2.3.3 Visualization Multiple Forecasts and EPS as Aggregated Exceedance Information

EFAS results are distributed to more than 38 EFAS partners from 23 different countries, most of them with different languages. Since end users include both

experts from hydrological forecasting and civil protection services, the recipients have different level of training and understanding. Therefore, EFAS information must be clear and unambiguous in order to be correctly understood and lead to appropriate action (Demeritt et al. 2007). EFAS results are visualized as time series or maps on the customizable web interface.

Time series are not shown as spaghetti plots as this was dismissed by the EFAS partners as a useful representation (Ramos et al. 2007). Instead, the information is shown as classical box diagrams with min/max whiskers and the 25, 50, and 75 quantiles in a box (Fig. 3) where the y-axis is expressed as return period and color coded according to the threshold. Information is aggregated when possible, for example, the results based on EPS are visualized together with the deterministic forecasts. However, also with the concise representation of box-plot diagrams, visualizing results from two or more HEPS in the same diagram is confusing. Instead, a more condensed view has been developed: First of all, information is aggregated in daily information, then expressed either in terms of threshold exceedance for the deterministic forecasts or in terms of percentages exceeding thresholds for the HEPS. This example is illustrated in Fig. 4.

The x-axis represents the date for which the forecast has been made. Each row represents a forecast. The first two rows illustrate the two deterministic forecasts which are simply expressed in terms of threshold exceedance. For further information, the tendency of rising or falling discharges is indicated by arrows, the peak as a star. The next row represents the percentage of HEPS exceeding the 5-year return period threshold (color coded red) and the 20-year return period (color coded purple). The two rows below show the equivalent for the COSMO-LEPS. In this

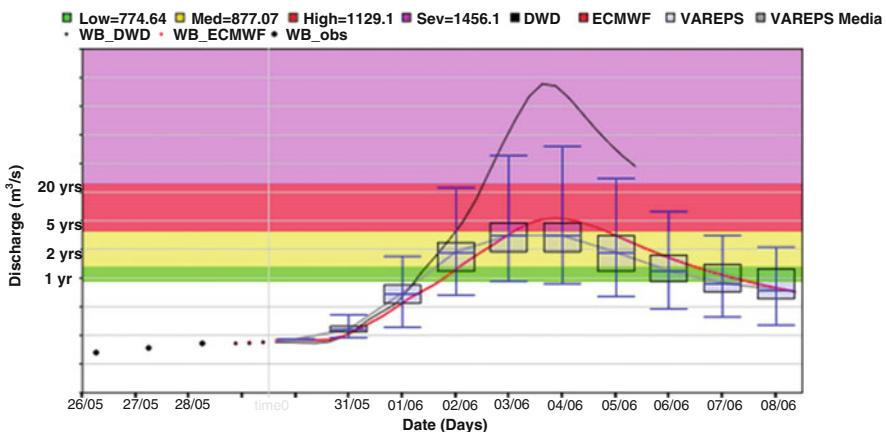


Fig. 3 Box-plot diagram showing the discharge time series in reference to the four EFAS thresholds low (green, 1.5-year return period), medium (yellow, 2-year return periods), high (red, 5-year return periods), and severe (purple, 20-year return period) for the ECMWF EPS. The first 3 days show the simulation of the hydrological conditions based on observations (indicated as dots), while the forecasts are shown for a lead time of 10 days for EPS. In addition, two deterministic forecasts of ECMWF (red line) and of DWD (black line) are shown

Overview of DWD, ECMWF, EUE > HAL, EUE > SAL

Forecast Type	30	31	1	2	3	4	5	6	7	8
DWD		↑	↓	↑	↓	*	↑			
EUD		↑	↓	↑	↓	↓	*	↑	↓	↓
EUE > HAL				10	39	45	20	8		
EUE > SAL					4	6	4			
COS > HAL				31	44					
COS > SAL					19					

Fig. 4 EFAS representation of flood threshold exceedances from multiple forecasts in one diagram, corresponding to the box-plot diagram in Fig. 3. The flood thresholds are color coded in the same way as in Fig. 3. For probabilistic forecasts, the probability of exceeding the 5- or 20-year periods are indicated in percent and color coded by intensity. Arrows indicate if the discharges are increasing or decreasing during the 24 h aggregation and a star symbolizes the peak of the time series

representation, the lead times are clearly visible, for example, the 5-day lead time for COSMO-LEPS as opposed to the 10 days of ECMWF. In this very condensed representation, all information is summarized at one glance and information on further forecasts could easily be added. It also has the advantage that the consistency between the forecasts are easily assessed, for example, are all the forecasts indicating the same day for the peak and is the event forecast with similar severity in all forecasts?

Furthermore, this representation also allows checking back in time if the forecasts are persistent (Fig. 5). The representations of forecasts are shifted so that forecasts where peaks are always forecast for the same day align. Figure 5 clearly shows that the forecasts cluster a higher probability for exceeding the EFAS high alert around the 3–5th of the coming month. It also clearly shows that the probabilities are increasing from forecast to forecast from an initial 4–8% to 45% in the last forecast.

Spatial information is shown in the form of maps and when appropriate follow the same color coding concept – pixels exceeding the 5-year return period thresholds are color coded in red, while those exceeding the 2-year return period are color coded in yellow.

2.4 Postprocessing

Although the concept of threshold exceedance has proven to work well and has been successful, experts in the flood forecasting centers are also interested in quantitative information for better comparison of national results with EFAS. Therefore, at selected points where real-time data are available, the streamflow output of EFAS is postprocessed so that the forecasts start with the same value as the observed

Forecast Day	27	28	29	30	31	1	2	3	4	5	6	7	8
2013-05-27 00:00							2	4	6	6			
2013-05-27 12:00							4	4	4	4			
2013-05-28 00:00								6	8	10	4		
2013-05-28 12:00							2	6	4	4			
2013-05-29 00:00							2	6	18	18	12	4	2
2013-05-29 12:00							2	18	31	35	27	14	4
2013-05-30 00:00							10	39	45	20	8		

Fig. 5 Persistence diagram for the current probabilistic forecasts based on ECWMF Ensemble Prediction System (last line) and the history of previous forecasts (lines above last). The current forecast is the same as the EUE > HAL of Fig. 4

discharges and the information can be readily uploaded into national systems and compared.

For this, a novel method based on wavelet transformation has been developed which is described in Bogner and Pappenberger (2011) and Bogner and Kalas (2008). For specific stations where real-time data are available, the method applies AutoRegressive models with eXogenous input (ARX) to relate the observed streamflow at one time to the previous one with a time lag and the simulated model. This approach will work particularly well for the first time lags. However, model errors usually show spatio-temporal variations that can, for example, depend on the season or local weather conditions. Such errors can be handled most efficiently using wavelet transformations as described in detail by Bogner and Kalas (2008).

The postprocessing method developed for EFAS has further the advantage that information on both model uncertainty (for past days) and total predictive uncertainty (for forecasts) can be provided to the end user as shown in Fig. 6. Currently, developments are ongoing to publish this postprocessed forecasts also as a web service using Open Geospatial Consortium (OGC) standards for a better direct integration into the national forecasting systems.

2.5 Calculation of Scores

Quality control and validation of EFAS results is essential for EFAS (de Roo et al. 2011). While deterministic forecasts can be relatively easily classified as “hit,” “false alarm,” or “missed” when comparing the forecasts to the observed event, for probabilistic forecasts it needs to be verified if the distribution of the hydrographs as well as the probability of the extreme events have been captured correctly (Bartholmes et al. 2009; Cloke and Pappenberger 2009). Since different scores test

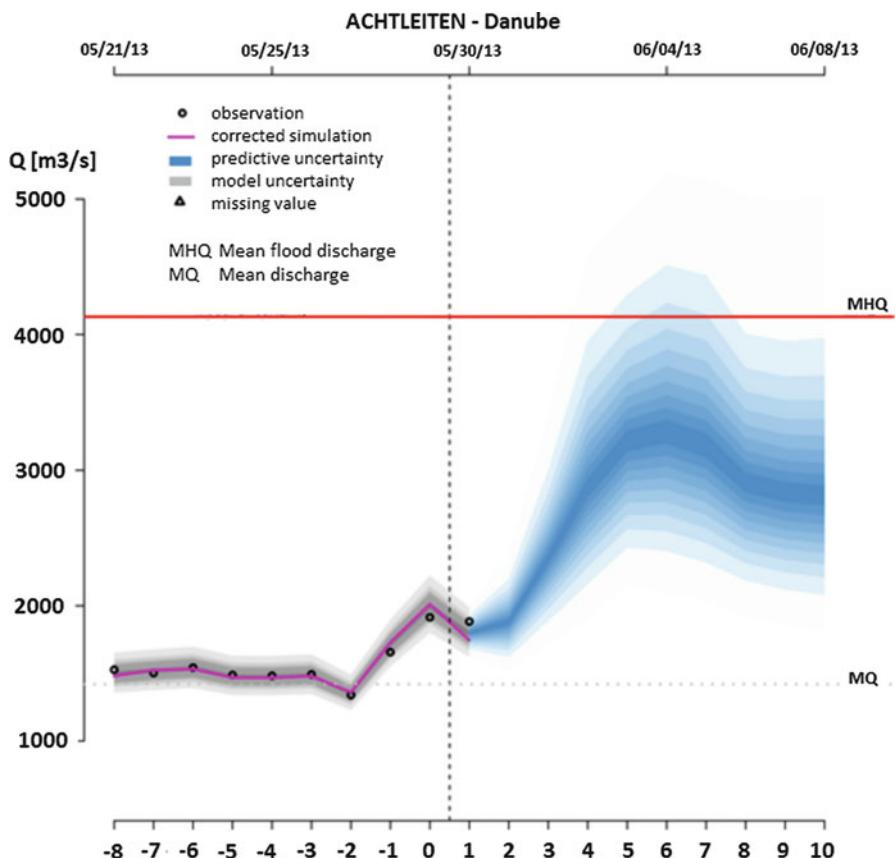


Fig. 6 Postprocessed EFAS results for selected stations indicating the model uncertainty in gray shading for the period where the model is driven with observed meteorological data and the predictive uncertainty in blue shading

different performance attributes such as reliability, resolution, sharpness, spread-skill, and bias, single scores do not capture the performance of the HEPS (Hsu and Murphy 1986; Jolliffe and Stephenson 2003).

For EFAS, the HEPS are validated against the EFAS water balance for each grid point. Average scores are updated on the 13th day of each month. In order to reduce possible effects of seasonality, the scores are calculated over a time window including the corresponding first day of the month and the past 365 days. For EFAS, several scores have been calculated including the rank histogram, root mean square error (RMSE) and the RMSE scaled by mean discharges, also referred to as coefficient of variation of RMSE (CV), Brier (skill) score, continuous ranked probability (skill) score (CRPS), relative operating characteristics (ROC), and the area under the ROC (AROC). Current skill scores are published regularly in the bimonthly EFAS bulletins accessible on www.efas.eu. A longer term assessment of

the skill of EFAS has been performed and published by Pappenberger et al. (2011a,b) and the use of benchmarks in EFAS for skill score analysis evaluated by Pappenberger et al. (2015b).

2.6 Training

Training plays an important role in EFAS. Its concepts, setups, and functionalities of the EFAS interface are demonstrated regularly to the partners during the annual EFAS meetings as well as during dedicated training sessions held at the National hydrological services. In addition, during the EFAS annual meetings, different topics have been addressed such as the concept of probabilistic forecasting, skill scores, making decisions based on uncertain forecasts, and different ways of visualizing results (Pappenberger et al. 2013; Wetterhall et al. 2013).

2.7 Partner Network and Connection to International Initiatives

A dedicated and active partner network has been key to ensuring that EFAS products are designed as added value products. Interaction and feedback from the National hydrological services as well as the civil protection community has ensured that the products are complementary, understood, and used by the various decision makers. The EFAS partner network has grown to 38 dedicated partners and several associated partners. In addition to the operational end users, the development of a continental flood forecasting system has also put EFAS at the forefront of research in probabilistic flood forecasting and served as a testbed for many research studies not directly linked to the operational system and an example for others to follow (for example, Alfieri et al. 2011; He et al. 2009; Pappenberger et al. 2015a, b; Bogner et al. 2012a, b). HEPEX has played an important role for EFAS since the HEPEX community addressed many topics of relevance for the development of the system including preprocessing of meteorological data, bias correction of both meteorological and hydrological data, estimation of uncertainty, visualization of results, and definition of end user needs. Furthermore, close links will be kept with the Global Flood Partnership, a recently launched international initiative to combat the effects of floods on international scale (<http://portal.gdacs.org/Global-Flood-Partnership>).

3 Summary and Conclusion

With the European Flood Awareness System, the first operational, continental flood early warning system for Europe has been developed. It is operated under the umbrella of the European space program Copernicus as part of the Emergency Management Service in support to improved preparedness for floods in Europe and serves both the National hydrological services and the European Civil Protection community.

EFAS is a comprehensive system incorporating several ensemble prediction systems as well as deterministic weather forecasts to assess the probability of critical flood thresholds being exceeded in the coming 10–15 days. By combining different concepts such as the flood threshold exceedance based on a model consistent framework, the use of forecast persistence to increase reliability, easy visualization of forecasts on a twice daily updated web interface for a wide end user basis, sophisticated postprocessing of forecast for an improved quantitative comparison with national results, and a regular calculation and publication of skill scores EFAS has become a unique, continental scale flood forecasting system. It is embedded in a management structure including a strong partner network which provides valuable feedback on the performance and future developments of the system. Continuous training sessions held at the National hydrological services and a strong linkage to research projects and international initiatives such as HEPEX ensure that EFAS products are designed as added value products and maintain the system as state of the art.

References

- L. Alfieri, D. Velasco, J. Thielen Del Pozo, Flash flood detection through a multi-stage probabilistic warning system for heavy precipitation events. *Adv. Geosci.* **29**, 69–75 (2011)
- L. Alfieri, P. Salamon, F. Pappenberger, F. Wetterhall, J. Thielen, Operational early warning systems for water-related hazards in Europe. *Environ. Sci. Pol.* **21**, 35–49 (2012a)
- L. Alfieri, J. Thielen, F. Pappenberger, Ensemble hydro-meteorological simulation for flash flood early detection in southern Switzerland. *J. Hydrol.* **424–425**, 43–153 (2012b)
- L. Alfieri, M. Berenguer, V. Knechtl, K. Liechti, D. Sempere-Torres, M. Zappa, Flash flood forecasting based on rainfall thresholds, in *Handbook of Hydrometeorological Ensemble Forecasting*, ed. by Q. Duan et al. (Springer, Berlin/Heidelberg, this issue). https://doi.org/10.1007/978-3-642-40457-3_49-1
- M.H.N. Bakker, Transboundary river floods: vulnerability of continents, international river basins and countries. Ph.D Dissertation, Oregon State University, 276, (2007), <http://hdl.handle.net/1957/3821>
- J.C. Bartholmes, J. Thielen, M.H. Ramos, S. Gentilini, The European Flood Alert System EFAS – part 2: statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.* **13**, 141–153 (2009)
- K. Bogner, F. Pappenberger, Multiscale error analysis, correction, and predictive uncertainty estimation in a flood forecasting system. *Water. Resour. Res.* (Impact Factor: 3.15) **47** (2011). : <https://doi.org/10.1029/2010WR009137>
- K. Bogner, M. Kalas, Error-correction methods and evaluation of an ensemble based hydrological forecasting system for the Upper Danube catchment. *Atmos. Sci. Lett.* **9**(2), 95–102 (2008)
- K. Bogner, H.L. Cloke, F. Pappenberger, A. De Roo, J. Thielen, Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube Catchment. *Int. J. River Basin Manage.* **10**(1), 1–12 (2012a). ISSN 1814–2060. <https://doi.org/10.1080/15715124.2011.625359>
- K. Bogner, F. Pappenberger, H.L. Cloke, Technical note: the normal quantile transformation and its application in a flood forecasting system. *Hydrol. Earth Syst. Sci.* **16**, 1085–1094 (2012b). ISSN 1027–5606. <https://doi.org/10.5194/hess-16-1085-2012>
- R. Brazdil, C. Pfister, H. Wanner, H. von Storch, J. Luterbacher, Historical climatology in Europe – the state of the art. *Clim. Change* **70**, 363–430 (2005)
- R. Buizza, A. Hollingsworth, F. Lalaurie, A. Ghelli, Probabilistic predictions of precipitation using the ECMWF ensemble prediction system. *Weather Forecast.* **14**, 168–189 (1999)

- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**(3–4), 613–626 (2009)
- H. Cloke, J. Thielen, F. Pappenberger, S. Nobert, G. Balint, C. Edlund, A. Koistinen, C. de Saint-Aubin, E. Sprokkereef, C. Viel, P. Salamon, R. Buizza, Progress in the implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for operational flood forecasting. *ECMWF Newsletter* **121**, 20–24 (2009)
- H. Cloke, F. Pappenberger, J. Thielen, V. Thiemig, Operational European flood forecasting, in Environmental Modelling: Finding Simplicity in Complexity, ed. by J. Wainwright, M. Mulligan, 2nd edn. (Wiley, Chichester, 2013a). <https://doi.org/10.1002/9781118351475.ch25>
- H.L. Cloke, F. Pappenberger, S.J. van Andel, J. Schaake, J. Thielen, M.-H. Ramos, Hydrological ensemble prediction systems. *Hydrol. Process.* **27**, 1–4 (2013b). <https://doi.org/10.1002/hyp.9679>
- H.L. Cloke, F. Wetterhall, Y. He, J.E. Freer, F. Pappenberger, Modelling climate impact on floods with ensemble climate projections. *Q. J. R. Meteorol. Soc.* **139**(671 part B), 282–297 (2013c). ISSN 1477-870X. <https://doi.org/10.1002/qj.1998>
- M. Collins, S. Knight, Ensembles and probabilities: a new era in the prediction of climate change. *Phil. Trans. R. Soc. A.* **365**, (1857), 1471–2962 (2007)
- A. de Roo, J. Thielen, P. Salamon, K. Bogner, S. Nobert, H.L. Cloke, D. Demeritt, J. Younis, M. Kalas, K. Bódis, D. Muraro, F. Pappenberger, Quality control, validation and user feedback of the European Flood Alert System (EFAS). *Int. J. Digital Earth.* **4**(Supplement 1), 77–90 (2011), Special Issue
- A. de Roo, B. Gouweleeuw, J. Thielen, J. Bartholmes, P. Bongioannini-Cerlini, E. Todini, P. Bates, M. Horritt, N. Hunter, K.J. Beven, F. Pappenberger, E. Heise, G. Rivin, M. Hills, A. Hollingsworth, B. Holst, J. Kwadijk, P. Reggiani, M. van Dijk, K. Sattler, E. Sprokkereef, Development of a European Flood Forecasting System. *Int. J. River Basin Manage.* **1**, 49–59 (2003)
- D. Demeritt, H. Cloke, F. Pappenberger, J. Thielen, J. Bartholmes, M.H. Ramos, Ensemble predictions and perceptions of risk, uncertainty, and error in flood forecasting. *Environ. Hazards* **7**, 115–127 (2007)
- B. Gouweleeuw, P. Reggiani, A. De Roo, A European Flood Forecasting System EFFS. Full Report, European Report EUR21208, EC DG JRC & WL Delft Hydraulics, p. 304 (2004)
- B.T. Gouweleeuw, J. Thielen, G. Franchello, A.P.J. de Roo, R. Buizza, Flood forecasting using medium-range probabilistic weather prediction. *Hydrol. Earth Syst. Sci.* **9**(4), 365–380 (2005)
- Y. He, F. Wetterhall, H.L. Cloke, F. Pappenberger, M. Wilson, J. Freer, G. McGregor, Tracking the uncertainty in flood alerts driven by grand ensemble weather predictions. *Met. Apps* **16**, 91–101 (2009). <https://doi.org/10.1002/met.132>
- W.-R. Hsu, A.H. Murphy, The attributes diagram: a geometric framework for assessing the quality of probability forecasts. *Int. J. Forecast.* **2**, 285–293 (1986)
- I.T. Jolliffe, D.B. Stephenson, *Forecast Verification: A practitioner's Guide in Atmospheric Science* (Wiley, Chichester, 2003)
- F. Pappenberger, K.J. Beven, N.M. Hunter, P.D. Bates, B.T. Gouweleeuw, J. Thielen, A.P.J. de Roo, Cascading model uncertainty from medium range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within the European flood forecasting system EFFS. *Hydrol. Earth Syst. Sci.* **9**(4), 381–393 (2005)
- F. Pappenberger, J. Bartholmes, J. Thielen, H.L. Cloke, R. Buizza, A. de Roo, New dimensions in early flood warning across the globe using grand-ensemble weather predictions. *Geophys. Res. Lett.* **35**, L10404 (2008). <https://doi.org/10.1029/2008GL033837>
- F. Pappenberger, J. Thielen Del Pozo, M. Del Medico, The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrol. Process.* **25**(7), 1091–1113 (2011a)
- F. Pappenberger, K. Bogner, F. Wetterhall, H. Yi, H. Cloke, J. Thielen Del Pozo, Forecast convergence score: a forecaster's approach to analysing hydro-meteorological forecast systems. *Adv. Geosci.* **29**, 27–32 (2011b)

- F. Pappenberger, E. Stephens, J. Thielen, P. Salamon, D. Demeritt, S.J. van Andel, F. Wetterhall, L. Alfieri, Visualizing probabilistic flood forecast information: expert preferences and perceptions of best practice in uncertainty communication. *Hydrol. Process.* **27**(1), 132–146 (2013). <http://onlinelibrary.wiley.com/doi/10.1002/hyp.9253/full>
- F. Pappenberger, H.L. Cloke, D.J. Parker, F. Wetterhall, D.S. Richardson, J. Thielen, The monetary benefit of early flood warnings in Europe. *Environ. Sci. Pol.* **51**, 278–291 (2015a). ISSN 1873–6416. :<https://doi.org/10.1016/j.envsci.2015.04.016>
- F. Pappenberger, M.H. Ramos, H.L. Cloke, F. Wetterhall, L. Alfieri, K. Bogner, A. Mueller, P. Salamon, How do I know if my forecasts are better? Using benchmarks in Hydrological ensemble prediction. *J. Hydrol.* **522**, 697–713 (2015b). ISSN 0022–1694. <https://doi.org/10.1016/j.jhydrol.2015.01.024>
- M.-H. Ramos, J. Bartholmes, J. Thielen-del Pozo, Development of decision support products based on ensemble forecasts in the European flood alert system. *Atmos. Sci. Lett.* **8**(4), 113–119 (2007)
- J. Schaake, K. Franz, A. Bradley, R. Buizza, The Hydrological Ensemble Prediction EXperiment (HEPEX). *Hydrol. Earth Syst. Sci. Discuss.* **3**, 3321–3332 (2006)
- E. Stephens, H. Cloke, Improving flood forecasts for better flood preparedness in the UK (and beyond). *Geochem. J.* **180**, 310–316 (2014). <https://doi.org/10.1111/geoj.12103>
- J. Thielen, J. Schaake, R. Hartman, R. Buizza, Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. *Atmos. Sci. Lett.* **9**, 29–35 (2008a)
- J. Thielen, P. Salamon, A. De Roo, “Geographical information systems – An integral part of the European Flood Alert System (EFAS)”. *GeoFocus* (Editorial) (8), 12–16 (2008b). ISSN: 1578–5157
- J. Thielen, J. Bartholmes, M.-H. Ramos, A. de Roo, The European Flood Alert System – part 1: concept and development. *Hydro. Earth Syst. Sci* **13**, 125–140 (2009)
- J. Thielen, F. Pappenberger, P. Salamon, K. Bogner, P. Burek, A. de Roo, State of the art of flood forecasting – from deterministic to probabilistic approaches, in *Flood Hazards: Impacts and Responses for the Built Environment*, ed. by J. Lamond, C. Booth, F. Hammond, D. Proverbs, T. Francis (Francis and Taylor, London, 2011), 371 pp
- J. Toothill, Central European Flooding August 2002, Technical Report EQECAT, ABS Consulting, 21 p, (2002)
- J.M. van der Knijff, J. Younis, A.P.J. de Roo, LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.* **24**, 189–212 (2010)
- F. Wetterhall, F. Pappenberger, H. Cloke, J. Thielen del Pozo et al., Forecasters priorities for improving probabilistic flood forecasts. *Hydrol. Earth Syst. Sci.* **17**, 4389–4399 (2013)
- P. Yiou, P. Ribereau, P. Naveau, M. Nogaj, R. Brazdil, Statistical analysis of floods in Bohemia (Czech Republic) since 1825. *Hydrol. Sci. J.* **51**(5), 930–945 (2006)



Seasonal Drought Forecasting on the Example of the USA

Eric F. Wood, Xing Yuan, Joshua K. Roundy, Ming Pan, and Lifeng Luo

Contents

1 Introduction: Princeton's Seasonal Drought Forecast System	1280
2 Drought Forecast Application over USA	1281
3 Conclusions	1286
References	1287

Abstract

Drought is a slowly developing process and usually begins to impact a region without much warning once the water deficit reaches a certain threshold. Predicting the drought a few months in advance will benefit a variety of sectors for drought planning and preparedness. In response to the National Integrated Drought Information System (NIDIS), the Princeton land surface hydrology group has been working on drought monitoring and forecasting for over 10 years and has developed a seasonal drought forecasting system based on global climate forecast models and a large-scale land surface hydrology model.

E. F. Wood (✉) · M. Pan

Department of Civil and Environmental Engineering, Princeton University, Princeton, NJ, USA
e-mail: efwood@princeton.edu; mpan@princeton.edu

X. Yuan

RCE-TEA, Inst. of Atmosph. Phys., Chinese Academy of Sciences, Beijing, China
e-mail: yuanxing@nju.edu.cn

J. K. Roundy

Department of Civil, Environmental and Architectural Engineering, University of Kansas,
Lawrence, KS, USA
e-mail: jkroundy@ku.edu; jroundy@princeton.edu

L. Luo

Department of Geography, Michigan State University, East Lansing, MI, USA
e-mail: lluo@msu.edu

This chapter will showcase the performances of the system in predicting soil moisture drought area, frequency, and severity over the Conterminous United States (CONUS) at seasonal scales; discuss about the challenges in forecasting streamflow for hydrologic drought; and provide an outlook for future developments and applications.

Keywords

Drought · Seasonal forecast · Hydrology · Soil moisture · Streamflow · Severity · Climate model · Land surface model · CFSv2 · VIC · NMME · Ensemble prediction · Postprocessing · Downscaling · Bayes

1 Introduction: Princeton's Seasonal Drought Forecast System

The central element of Princeton's Seasonal Drought Forecast System (Luo and Wood 2007, 2008) is the Variable Infiltration Capacity (VIC) hydrological model (Liang et al. 1996) that transforms seasonal climate forecast information into hydrological information such as soil moisture and streamflow. VIC is one of the state-of-the-art macroscale hydrological models, and it has been calibrated, validated, and evaluated in numerous studies at grid, basin, and continental scales (e.g., Wood et al. 1997). The forecast system implements a Bayesian merging procedure (Luo et al. 2007) to combine seasonal forecasts from dynamical climate models with observed climatology at the monthly level to obtain posterior distributions for monthly precipitation and temperature at each grid for each month of the forecast period. With the mean and variance of the posterior distribution for the spatially downscaled monthly variables, a hybrid method that includes both the historical-analogue criterion and random selection is used to generate daily forecast time series (Luo and Wood 2008). Finally, a simple scaling method is used to adjust the resulting series according to the monthly ensemble mean of the posterior distribution (Yuan et al. 2013a). Runoff generated within a grid cell is routed to the stream gauge location using the linear routing model developed by Lohmann et al. (2004).

The inputs of the forecast system are monthly precipitation and temperature from seasonal climate forecast models and the initial land surface conditions generated from Princeton's drought monitoring system as part of the North American Land Data Assimilation (NLDAS) system. The outputs include high-resolution (1/8th degree spatially) soil moisture and streamflow at selected stations, which can be converted into percentiles to reflect the agricultural and hydrologic drought conditions, respectively. In the following sections, the system is evaluated for predicting soil moisture drought areas, frequency, and severity. The ensemble streamflow forecast is also assessed using root mean square error skill score (SS_{RMSE}) and Relative Operating Characteristic (ROC) diagram. The latest version of the forecast system (Yuan et al. 2013a), which is based on National Centers for Environmental Prediction (NCEP)'s Climate Forecast System version 2 (CFSv2; Yuan et al. 2011),

has been successfully transitioned to NCEP/EMC for operational drought forecasting (<http://www.em.ncep.noaa.gov/mmb/nldas/forecast/TSM/perc/>).

2 Drought Forecast Application over USA

Since the beginning of 2007, new drought conditions were developing in several large regions within the continental USA. In the West, very little rain fell over much of California during the winter-spring of 2006–2007. The severe-to-extreme drought across the West resulted in the dramatic spread of fire activities in some parts of the region, also putting the rest of the region in high risk for wildfire. Drought conditions were also severe in the Southeast in terms of the impact on agriculture and the local economy. Much of Alabama was in the midst of the worst drought it has experienced in more than one hundred years. Figure 1, from Luo and Wood (2007), shows the evolution of the droughts and their predictions over the West and the Southeast. The prediction is based on a CFSv1/VIC forecast system. Within each region, the number of 1/8th degree grids where the monthly mean soil moisture value is below the 20th percentile threshold is counted for each month. The black solid lines in Fig. 1 are from Princeton's real-time drought monitor and represent the development of the droughts. For the predictions, grids that satisfy the same criteria are counted in each of the seven ensemble members and the counts are averaged to give the mean forecasts (solid green, blue and red lines). The spread of each ensemble forecast is indicated by the dashed color lines as one standard deviation from their mean. As Fig. 1 demonstrates, the predictions are very skillful in capturing the evolution of the droughts over both regions, especially during the first 2 months of each forecast.

NCEP updated its operational seasonal climate forecast system from CFSv1 to CFSv2 in March 2011. Therefore, Princeton's seasonal drought forecasting system

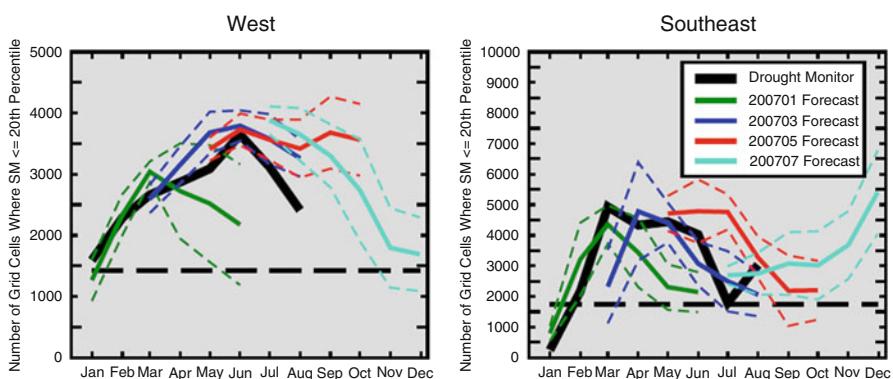


Fig. 1 Drought predictions over the West and the Southeast. Shown is the area (number of grid cells) where soil moisture is below the 20th percentile of the climatological soil moisture distribution from a 50-year offline simulation. The drought index is defined as the soil moisture percentiles (in percent) (Luo and Wood 2007)

was also updated accordingly. Comprehensive validation over the CONUS was carried out to compare CFSv2/VIC with CFSv1/VIC and ESP/VIC, where ESP represents using historical resampling of the meteorology as forecasts. Figure 2 shows the ratio of forecasted over offline simulated soil moisture drought frequency averaged among all forecasts during 1982–2008 (Yuan et al. 2013a). For the droughts that last for at least 1 month (Fig. 2; first row), more than 80% of them are captured by ESP forecasts over central USA, 60–70% over SE, while less than 50% can be predicted over the NE, Ohio, and western coast areas. CFSv1 has improved forecasts over the Lower Mississippi and California, while CFSv2 has even further enhancements, especially over the eastern USA. For the 2-month duration droughts (Fig. 2; second row), offline simulation indicates that the occurrence frequency of these events is less than 10% over the eastern USA and Pacific Northwest. Thus, all three approaches have lower skill in forecasting them over these areas. However, CFSv2 forecasts have a consistent improvement over ESP and CFSv1 across the country. As the drought duration increases from 3 to 6 months, the chance of successfully forecasting their frequency decreases, especially in the eastern USA and west coast areas (Fig. 2; third and fourth rows). In contrast, all three approaches have plausible performance in predicting drought frequency over central USA out to 6 months.

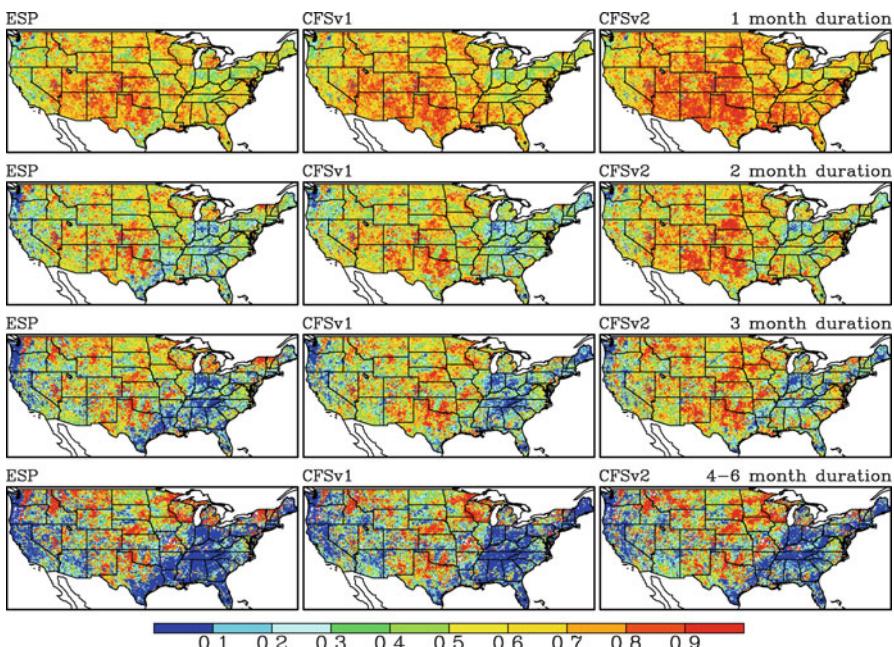


Fig. 2 Ratio of the ensemble mean forecasts of soil moisture drought frequency averaged among all forecasts in each calendar month at 0.5-month lead divided by those from offline simulation. As an example, the frequency of 3-month duration drought is counted by using all forecasts that have continuously dry conditions (less than 20%) in the first 3 months (Yuan et al. 2013a)

Besides the drought area and frequency, the predictive skill for drought severity is also evaluated. We define the regional accumulated severity (S) as $S = \sum_{i=1}^n \sum_{j=1}^t (1 - P_{i,j})$, where n and t represent that the number (n) of drought grid cells with drought durations of t months, and $P_{i,j}$ is the monthly percentile of soil moisture for a specific grid cell in a specific drought month (Yuan et al. 2013a). Note that n and $P_{i,j}$ may be different between the offline simulation and the forecasts; thus, the defined S is used to quantify regional accumulated soil water deficit during the drought period. Figure 3 shows the predictive skill of S for three types of forecasts over the CONUS region for different durations. To be consistent with the 3-month duration droughts which have four different leads based on 6-month forecasts, only the results with lead times up to 3.5 months are shown. The predictive skill of severity tends to be higher during winter due to the strong initial soil moisture control and/or better precipitation prediction skill from the climate forecast models. However, it is not as low as we expected during the summer, while the lowest skill occurs during the spring and fall. Therefore, the predictive skill for short-term drought has different seasonal characteristics as compared with precipitation, indicating the important role of initial soil moisture condition in seasonal forecasting. As compared with ESP, climate models do offer added value, and their advantages

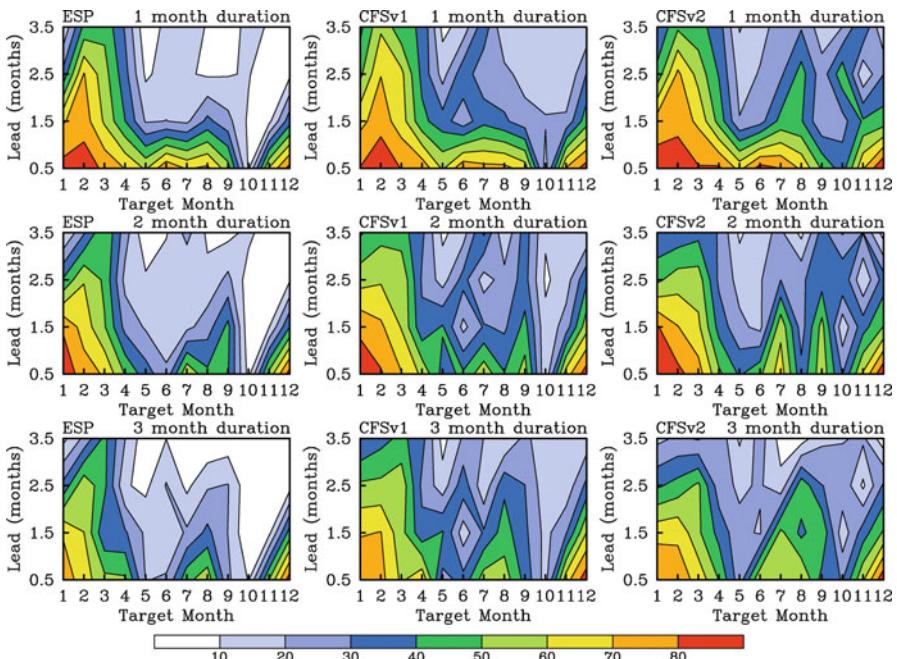


Fig. 3 R^2 (%) of drought severity accumulated over CONUS for different durations as functions of target months and leads (Yuan et al. 2013a)

become clearer at long leads where there is less impact from initial conditions. Though CFSv2 is better than CFSv1 during the winter, summer, and fall seasons, it is slightly worse than the latter during the spring.

The above forecasts are validated against offline simulated soil moisture. Although the uncertainty of the hydrologic model could be reduced through calibration, some discrepancies could be expected for model simulated soil moisture, which could not be easily determined due to lack of soil moisture observation at large scales. On the other hand, the predicted streamflow can be directly validated against gauge observation. Figure 4 shows the root mean square error skill score ($SS_{RMSE} = 1 - RMSE/RMSE_{ESP}$) for ensemble mean monthly streamflow from climate model-based forecasts (Yuan et al. 2013a). The streamflow from ESP is used as the reference forecast. During the first month, CFSv1 and CFSv2 have moderate improvements against ESP over most watersheds, with high skill score (>0.1) occurring over California (Fig. 4a–b), and some over Ohio basin for CFSv2. Based on the skill scores averaged over the 14 large basins, CFSv2 reduces $RMSE$ for streamflow forecasts from ESP by

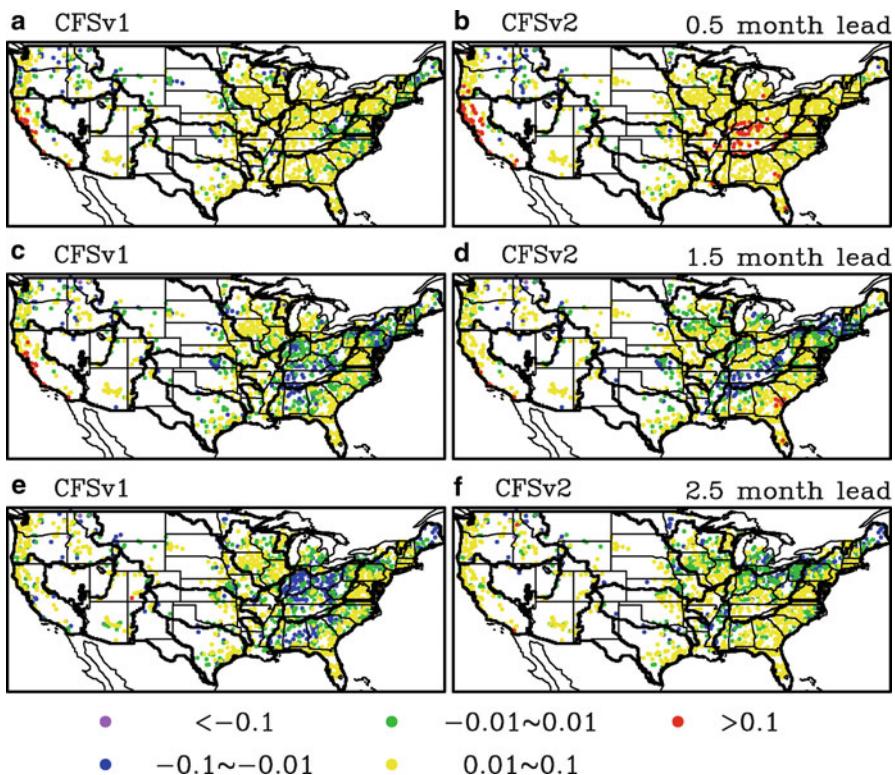


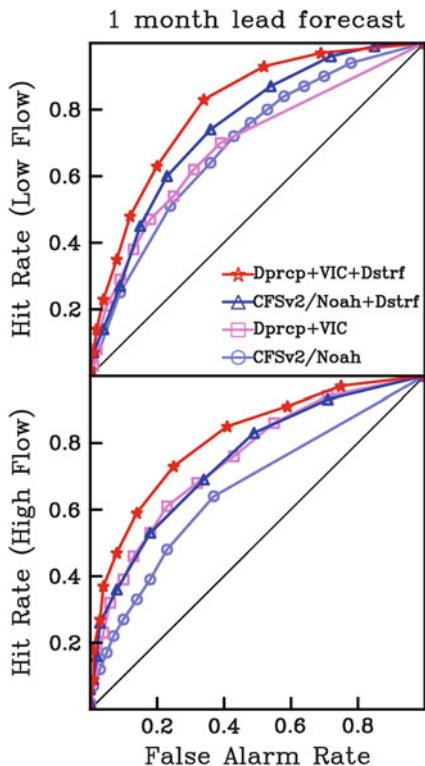
Fig. 4 SS_{RMSE} for ensemble mean monthly streamflow forecasts from CFSv1/VIC and CFSv2/VIC calculated at 1734 gauges. The reference is ESP/VIC (Yuan et al. 2013a)

about 4–7% in the eastern USA and less than 4% over the west, and the errors are reduced by 10% over California. In the second and third month, climate model-based forecasts are comparable to the ESP at more gauges (Fig. 4c–f; green dots). The large basin averaged skill scores indicate that the error reductions are below 3% (except the 1.5-month forecasts over California), and CFSv1 is slightly worse than ESP over the Ohio basin. The blue dots indicate that streamflow forecasts over some small catchments are still quite challenging.

To further diagnose the climate model-based seasonal streamflow ensemble forecast, which is targeted for providing hydrologic drought prediction, a set of seasonal streamflow hindcast experiments for each calendar month during 1982–2008 was carried out based on the CFSv2 reforecasts (Yuan and Wood 2012a). These are (1) bilinearly interpolating daily runoff from CFSv2 to 1/8 degree over Ohio basin, and routing them to the stream gauge location using a linear routing model. This experiment is referred to as CFSv2/Noah, where Noah is the land surface component of CFSv2; (2) downscaling monthly precipitation and temperature from CFSv2 by using the Bayesian method, then driving water-budget version of the VIC land surface model to produce hydrologic forecast with initial conditions provided by offline simulation, and finally routing the runoff forecasts by the same routing model as in experiment CFSv2/Noah. This experiment is referred to as $D_{\text{prep}} + \text{VIC}$, which is the same as current drought forecast system; (3) bias-correcting the monthly streamflow forecasts from CFSv2/Noah using a Bayesian procedure that merges forecast and observed streamflow information. This experiment is referred to as CFSv2/Noah + D_{strf} ; and 4) bias-correcting streamflow forecasts from $D_{\text{prep}} + \text{VIC}$ by the same method as CFSv2/Noah + D_{strf} , which is referred to as $D_{\text{prep}} + \text{VIC} + D_{\text{strf}}$.

Figure 5 shows the Relative Operating Characteristic (ROC) diagrams (Wilks 2011) for month-1 forecasts for low and high flow conditions. As a general rule, forecast with better discrimination exhibits ROC curves approaching the upper-left corner of the diagram more closely. For detecting low flow conditions, $D_{\text{prep}} + \text{VIC}$ only has a moderate advantage against CFSv2/Noah when the false alarm rate is small, this is partly due to the under-forecasting bias of CFSv2/Noah. In contrast, such advantage becomes clearer for high flow conditions. $D_{\text{prep}} + \text{VIC}$ and CFSv2/Noah + D_{strf} have comparable performance in the discrimination, and the latter is better than the former for low flow condition with high false alarm rate. As compared with the above three set of forecasts, $D_{\text{prep}} + \text{VIC} + D_{\text{strf}}$ shows overall improvement for both low and high flow conditions (Fig. 5) and increases the area under ROC curve from CFSv2/Noah + D_{strf} by about 10% (8%) for low (high) flow conditions. Therefore, postprocessing seems to be a critical step to make streamflow forecast more useful (Yuan and Wood 2012a; Yuan et al. 2013a) for hydrologic drought prediction. This analysis also indicates that the uncertainty from the hydrologic model could be comparable to the uncertainty in the precipitation forecasts from the seasonal climate models. Besides the parameter calibration, improving the process parameterizations in the hydrologic models is important for improved seasonal hydrologic forecasting.

Fig. 5 ROC diagrams for low and high flow month-1 forecasts averaged over 50 gauges within Ohio basin and in all seasons during 2001–2008 (Yuan and Wood 2012a)



3 Conclusions

The Princeton's Seasonal Drought Forecasting system has now been extended from CONUS (Luo and Wood 2007; Yuan et al. 2013a) to Africa (Yuan et al. 2013b; Sheffield et al. 2014) for drought early warning and is being tested over major global river basins for its potential contribution to the Regional Hydroclimate Projects (RHP) under the Global Energy and Water Exchanges Project (GEWEX). The climate forecast model has been updated from CFSv1 to CFSv2, and the overall system is being augmented with the suite of seasonal climate forecast models (Yuan and Wood 2012b, 2013) that are participating in the experimental North American Multimodel Ensemble (NMME) project (Kirtman et al. 2014). There is a plan to develop a hyperresolution land surface model to improve streamflow simulation and facilitate other applications (Wood et al. 2011). Moreover, Princeton terrestrial hydrology group also focus on diagnosing the seasonal climate forecast models during drought events, in term of both ocean–atmosphere teleconnection (Yuan and Wood 2013; Kam et al. 2014) and land-atmosphere coupling (Roundy et al. 2014). We are particularly interested in how the drought predictability research can improve the operational drought prediction.

References

- J. Kam, J. Sheffield, X. Yuan, E.F. Wood, Did a skilful prediction of sea surface temperatures help or hinder forecasting of the 2012 Midwestern US drought? *Environ. Res. Lett.* **9**, 034005 (2014). <https://doi.org/10.1088/1748-9326/9/3/034005>
- B.P. Kirtman et al., The North American MultiModels Ensemble (NMME): phase1 seasonal to interannual prediction, phase2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc* **95**, 585–601 (2014). <https://doi.org/10.1175/BAMS-D-12-00050.1>
- X. Liang, E.F. Wood, D.P. Lettenmaier, Surface soil moisture parameterization of the VIC-2L model: evaluation and modifications. *Global Planet. Change* **13**, 195–206 (1996)
- D. Lohmann et al., Streamflow and water balance intercomparisons of four land surface models in the North American Land Data Assimilation System project. *J. Geophys. Res.* **109**, D07S91 (2004). <https://doi.org/10.1029/2003JD003517>
- L. Luo, E.F. Wood, Monitoring and predicting the 2007 U.S. drought. *Geophys. Res. Lett.* **34**, L22702 (2007). <https://doi.org/10.1029/2007GL031673>
- L. Luo, E.F. Wood, Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. *J. Hydrometeorol.* **9**, 866–884 (2008)
- L. Luo, E.F. Wood, M. Pan, Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.* **112**, D10102 (2007). <https://doi.org/10.1029/2006JD007655>
- J.K. Roundy, C.R. Ferguson, E.F. Wood, Impact of land-atmospheric coupling in CFSv2 on drought prediction. *Clim. Dyn.* (2014). <https://doi.org/10.1007/s00382-01301982-7>
- J. Sheffield, E.F. Wood, N. Chaney, K. Guan, S. Sadri, X. Yuan, L. Olang, A. Amani, A. Ali, S. Demuth, L. Ogallo, A drought monitoring and forecasting system for sub-Saharan African water resources and food security. *Bull. Amer. Meteor. Soc* **95**, 861–882 (2014). <https://doi.org/10.1175/BAMS-D-12-00124.1>
- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences*. International Geophysics Series, vol. 100, 3rd edn. (Academic, Oxford/Waltham, 2011), 676pp
- E.F. Wood, D.P. Lettenmaier, X. Liang, B. Nijssen, S.W. Wetzel, Hydrological modeling of continental-scale basins. *Annu. Rev. Earth Planet. Sci.* **25**, 279–300 (1997)
- E.F. Wood et al., Hyperresolution global land surface modeling: meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resour. Res.* **47**, W05301 (2011). <https://doi.org/10.1029/2010WR010090>
- X. Yuan, E.F. Wood, Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. *Water Resour. Res.* **48**, W12519 (2012a). <https://doi.org/10.1029/2012WR012256>
- X. Yuan, E.F. Wood, On the clustering of climate models in ensemble seasonal forecasting. *Geophys. Res. Lett.* **39**, L18701 (2012b). <https://doi.org/10.1029/2012GL052735>
- X. Yuan, E.F. Wood, Multimodel seasonal forecasting of global drought onset. *Geophys. Res. Lett.* **40**, 4900–4905 (2013). <https://doi.org/10.1002/grl.50949>
- X. Yuan, E.F. Wood, L. Luo, M. Pan, A first look at Climate Forecast System version 2 (CFSv2) for hydrological seasonal prediction. *Geophys. Res. Lett.* **38**, L13402 (2011). <https://doi.org/10.1029/2011GL047792>
- X. Yuan, E.F. Wood, J.K. Roundy, M. Pan, CFSv2-based seasonal hydroclimatic forecasts over conterminous United States. *J. Climate* **26**, 4828–4847 (2013a). <https://doi.org/10.1175/JCLI-D-12-00683.1>
- X. Yuan, E.F. Wood, N.W. Chaney, J. Sheffield, J. Kam, M. Liang, K. Guan, Probabilistic seasonal forecasting of African drought by dynamical models. *J. Hydrometeor.* **14**, 1706–1720 (2013b). <https://doi.org/10.1175/JHM-D-13-054.1>



Ensemble Streamflow Forecasts for Hydropower Systems

Marie-Amélie Boucher and Maria-Helena Ramos

Contents

1 Introduction	1290
2 The Overall Context of Hydropower Production and Management	1292
3 Ensemble Streamflow Forecasts for Operational Systems	1295
4 Human Expertise, Quality, and Value of Forecasts	1298
5 Key Issues and Future Challenges	1301
References	1303

Abstract

Hydropower operation and planning requires streamflow forecasts at both short (typically, the first 4–5 days) and long ranges (a few months or a season ahead) over different spatial scales. Operational streamflow forecasting services a variety of decisions, made under conditions of risk and uncertainty, e.g., flood protection, dam safety, system's operation, optimization, and planning of power production. In areas where snow falls in significant quantities during winter, spring freshet poses additional challenges given the uncertainties related to the timing and volume of melt water flowing into hydropower reservoirs. Reservoir levels need to be gradually lowered over the winter to make it possible to store snowmelt water in spring. Reservoirs are thus important regulators of streamflow natural variability. They act as a storage place to water that can be used later to meet periods of higher electricity demands or to sell surplus electricity to the power distribution grid. They also usually are multipurpose, and their operation

M.-A. Boucher (✉)

Civil Engineering Department, Université de Sherbrooke, Sherbrooke, QC, Canada
e-mail: marie-amelie.boucher@usherbrooke.ca

M.-H. Ramos

IRSTEA, National Research Institute of Science and Technology for Environment and Agriculture, UR HBAN, Antony, France
e-mail: maria-helena.ramos@irstea.fr

must take into account the different water uses, which can, in some cases, be conflictual. The importance of accurate and reliable streamflow forecasts is therefore unquestionable. The hydropower sector has long recognized that streamflow forecasting is intrinsically uncertain and the use of ensemble forecasts is progressing fast. Key challenges today are related to the integration of state-of-the-art weather services, the implementation of systematic, advanced data assimilation schemes, to the assessment of the links between forecast quality and value, and to the enhancement of risk-based decision-making.

Keywords

Hydropower · Energy · Water management · Users · Decision-making · Data assimilation · Economic value · Reservoir · Storage · Multipurpose · Dams · Streamflow forecast · Human expertise · Forecast quality

1 Introduction

Energy production from falling waters to fulfill human's needs for growth has a long history, dating back hundreds of years to the use of water mills for agriculture. The first hydropower facilities, built at the end of the nineteenth century, began exploiting the kinetic energy of water masses flowing through rivers for electricity production (Kumar et al. 2011). In order to better manage the natural variability of river flows in time and be able to store water, dams and reservoirs were built in several river basins around the world. The management of hydropower production is then facilitated, but efforts must still be put into the collection of data and the development of modeling tools for the quantification of river flows. Modeling involves not only the forecasting of future inflows but also the integration of real-time data to assess current conditions of areas upstream of power plants and the analysis of historic time series of streamflow data to set up the model and quantify the long-term availability of local water resources. Hydropower operation also requires market price assessment and forecasts for future energy demand, which is also linked to atmospheric conditions and variability, as is the case of river flows.

Forecasting for hydropower production involves the prediction of several weather and hydrologic variables over a wide range of space and time scales. In space, for instance, one may consider the regional joint production of energy coming from hydropower plants installed at different types of facilities, from run-of-the-river hydroelectricity to production from rivers regulated by small to large dams. In time, as shown in Fig. 1, the needs of the hydropower sector for accurate and reliable forecasts span from forecasts up to 2–3 days ahead for flood protection of the population living downstream the facilities and for the security of installations (e.g., Akabari et al. 2014), medium-range forecasts up to 7–15 days ahead for the value of the production in the electricity grid (e.g. Tang et al. 2010), and long-term (months ahead) streamflow forecasts for hydropower optimization, planning, and seasonal water resources management, including dealing with environment protection measures and concurrent water uses during drought periods (e.g., Lu et al. 2017;

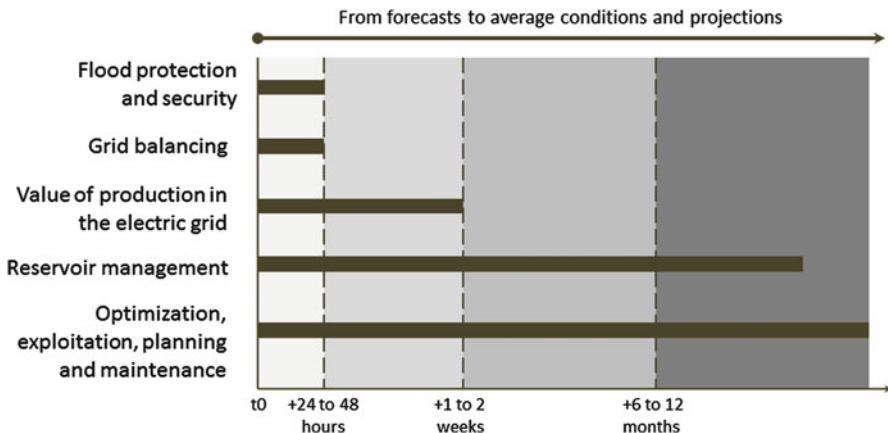


Fig. 1 Time frames in forecasting for hydropower systems

Bazile et al. 2017). The case of multi-year or long lasting droughts represents an especially challenging situation. It results in lower power generation capacities for both run-of-river plants (because of reduced inflows) and storage based plants, as lower reservoir levels translate into lower heads and inefficient water-to-energy conversion (Harto et al. 2011). In addition, drought periods are often associated with an increased demand for energy, attributable to air conditioning (e.g., Vliet et al. 2016). According to Vliet et al. (2016), the most recent recorded droughts throughout the world led to a reduction of 5.2% of hydropower production on average.

Additionally to forecasts, the hydropower industry is also concerned by hydrologic predictions based on future climate conditions and projected trends, as the effects of expected changes in precipitation and temperature may lead to changes in runoff volume, extremes, and seasonality, directly affecting the potential for hydropower generation (Kumar et al. 2011; Schaeffli 2015). These issues need to be considered in the long-term planning of hydropower plants and also in the planning of their interaction with other sources of energy (other renewable climate-related energies or nuclear power plants, for instance), notably in views of better managing the storage capacity provided by the reservoirs of regulated rivers (François et al. 2014). Climate change impacts on the hydropower sector can be addressed in contextualized studies (e.g., Boucher and Leconte 2013; Kiani et al. 2013; Hendrickx and Sauquet 2013), which usually include the assessment of regional trends in hydrometeorological variables, the evaluation of future market and economic scenarios, and the implications of expected changes for current management practices and planned adaptation strategies.

Within this broad context, the importance of accurate forecasts and reliable impact assessments is clear. The hydropower sector has long recognized that hydrometeorological forecasting is an essential part of its operations, and the use of ensemble- or scenario-based forecasts is progressing fast (e.g. Boucher et al. 2010; Fan et al. 2016; Séguin et al. 2017). This chapter focuses on the use of short- to long-

term forecasts for the hydropower sector. The following subsections provide an overview on the general context of hydropower production, the specificities of streamflow ensemble forecasts for hydropower production, and the main issues behind human expertise, forecast quality, and value in operational forecasting. Key challenges for the industry are presented at the end.

2 The Overall Context of Hydropower Production and Management

Water and energy are intricately linked in different ways. For instance, the extraction of shale natural gas requires large amounts of water; the desalination of sea water to make it drinkable typically requires the use of energy. According to the US Department of Energy (2014), “Historically, interactions between energy and water have been considered on a regional or technology-by-technology basis.” The same report stresses the need for reconnecting energy and water management at the national and even international levels in order to reduce the vulnerability of the worldwide population to climate change and disasters.

Hydropower production remains the activity that establishes the most direct link between energy and water. Water flowing in rivers or stored in reservoirs can be diverted through a turbine to drive a generator that converts the mechanical energy into electricity, which is then distributed to users through transmission lines and connected grids. The dependence of hydropower production on atmospheric and hydrologic variables affects not only energy supply but also users’ demand for electricity. Electricity demand is intrinsically related to weather conditions. It depends mostly on temperature, but also on other atmospheric variables such as humidity or wind, and is subject to seasonal fluctuations and variations across the week and during the day. When associated with storage in reservoirs, hydropower can regulate the natural variability of river flows and help in providing the energy balance between production and consumption, particularly when there is a need to quickly respond to peak load demands and grid stability (Fig. 1).

The flexibility offered by storage-based hydropower is a key feature to help in the stability of electrical systems that highly depend on climate-related (variable) renewable energy sources. Hydropower plants generate the largest share of electricity from renewables. According to Kumar et al. (2011), 12.9% of the world’s primary energy sources were renewable in 2008. In terms of electricity, 19% of the global production in 2008 was from renewable sources, with 16% coming from hydropower. The role of hydropower in increasing renewable energy penetration and ensuring energy security is emphasized with the increasing deployment of intermittent renewable sources such as wind and solar power (François et al. 2016).

The management of hydropower is challenged by the fluctuations in space and time of the weather and hydrologic variables that govern its production. Run-of-the-river hydropower systems are highly dependent on the spatial structure of river networks and the time variability of river flows. Their production is marked by the efficiency of the upstream watersheds in transforming rainfall into runoff and flow in

the channels of the rivers. When associated to large reservoirs, hydropower systems can store water for weeks, months, or even years, smoothing river flow temporal variability. Often these are, however, multifunction systems. The same reservoir has multiple purposes and must also comply with regulations and policies for other uses such as drinking water supply, flood and drought control, navigation, tourism, and environment protection, in addition to energy production (e.g., Tilmant et al. 2008; Björnson Gurung et al. 2016). Finally, for some larger catchments, series of reservoirs and hydropower dams are usually installed. The operation of a reservoir upstream may influence the operation of a reservoir downstream. Operation must then be planned in an integrated way. Concertation is key when different owners operate different facilities in the same river catchment. Multi-operator in a multiuser context calls for efficient modeling systems and coordination mechanisms (Anghileri et al. 2013). In these complex situations, hydropower production services will include not only hydrometeorological forecasting systems to provide input to models of reservoir simulation but also reservoir optimization tools to derive optimal and multi-objective operating rules.

The general problem of optimizing hydropower production depends on (1) the forecasts for future streamflows (or reservoir inflows), (2) the type of dam and installed power system (storage capacity, turbine capacity, structural and release constraints), and (3) the forecasts for the energy market prices, which generally reflect variations in electricity demand. Figure 2, adapted from Alemu et al. (2011), illustrates how input forecast data and models are linked in the context of the optimal management of the production of a hydropower plant with storage capacity. In the figure, T represents a forecast horizon expressed in days, which can represent short-,

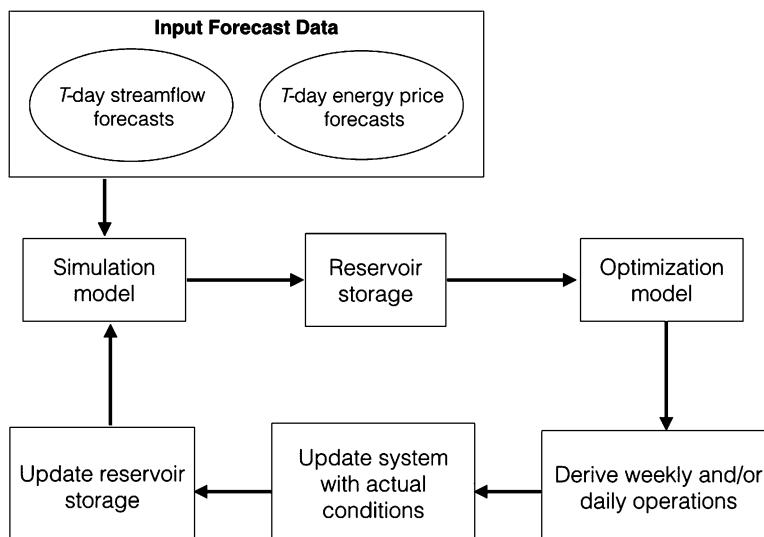


Fig. 2 General framework linking input data and models for the optimal management of hydropower production systems

medium-, or long-term forecasts. A typical hydropower producer aims at satisfying the demand for energy at all times during the year, while selling the energy at the highest possible price and managing the facilities as close as possible to optimal, according to future inflow forecasts and management constraints.

Since electricity itself cannot be stored, the reservoir behind a hydropower dam acts as storage place to water that can be used later to meet periods of higher (or peak) electricity demands or to sell surplus electricity to the power distribution grid. Consequently, the optimization of hydropower production searches to define an optimal decision rule between turning turbines on to use the water to produce energy at the present time or turning them off to save water to be used later, at a moment in the future when it might be more valuable. Structural and management constraints play, however, an important role in the optimization. For instance, the storage capacity of a reservoir is limited by its dimensions; inflows from snowmelt and extreme precipitation events need to be accommodated in the reservoir to avoid spilling water due to the occurrence of more water than expected from the forecasts; the occurrence of water levels over the design high water level of the dam or downstream flooding must be avoided for security reasons; downstream flows must be secured for other water uses and to comply with environmental regulations. Figure 3 presents a simplified, schematic view of the management of an inflow forecast to a water reservoir for the production of energy at the best hours of energy prices while respecting reservoir capacity constraints. The management rule defines when to direct the water flow through the turbines to generate electricity and when to store water in the reservoir. In real-time operation, rules and operations are updated on a day-by-day basis, but should always balance short-term and long-term targets (e.g., Lu et al. 2017).

For run-of-the-river hydropower, a small dam or a low head weir is usually created to raise water levels and facilitate water flow diversion to the water intake at the power plant. These ponds, however, do not store enough water for later use as in storage-based hydropower. The management of run-of-the-river hydropower plants is therefore more directly dependant on river flow variability and the forecasts of inflows, with little flexibility to control intermittency in time and manage volumes from extreme events. Flood forecasting is, nevertheless, an important component of such systems, since flood events can damage facilities and threaten onsite workers if rivers overflow their banks. In this context, accurate streamflow forecasts at short lead times (hourly or less) are preferred in order to fine-tune the operations. In the case of multi-objective and complex systems, with some storage capacity, “win-win” situations for hydropower production and flood control can be sought in order to evaluate if flood forecasts can be improved when hydropower production planning is integrated in real-time modeling and operations (e.g., Addor et al. 2011). Additionally, even if the environmental impacts of a small run-of-the-river dam may be lesser than the impact of a large hydropower dam, the physical and biological impacts of hydropeaking and rapid changes of water levels must also be considered in run-of-the-river hydropower operations. For this, hydrologic data and modeling at both the watershed and the river channel levels are crucial for defining adapted operational strategies and setting up integrated approaches for energy production and ecosystem modeling (Anderson et al. 2015).

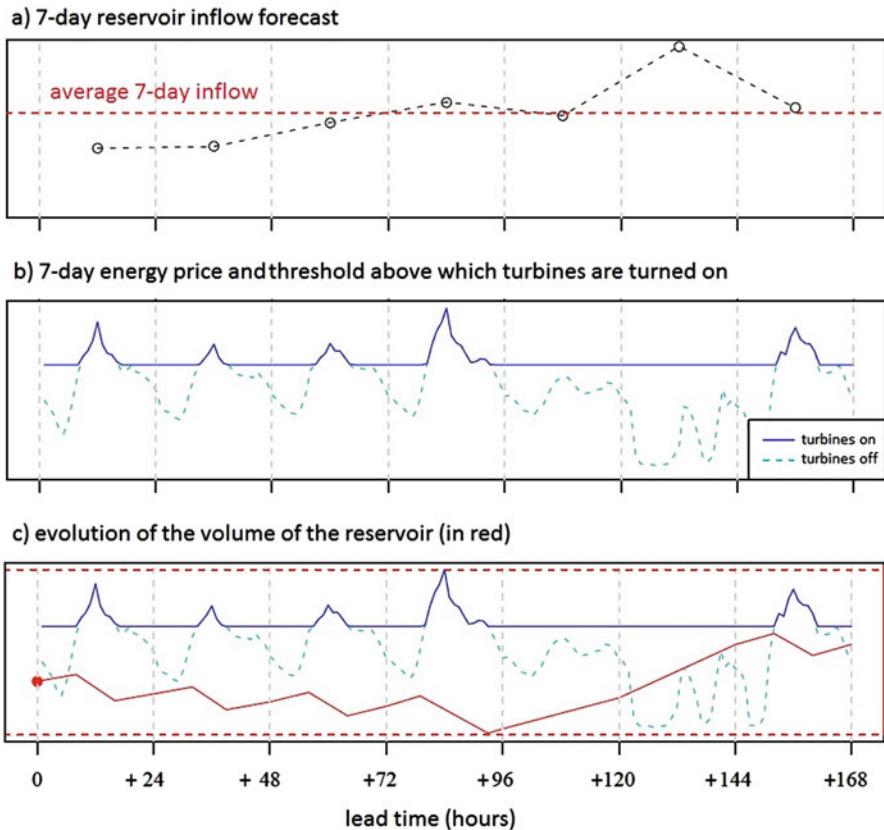


Fig. 3 Simplified scheme of the management of an inflow forecast to a reservoir-based hydropower system: based on a days-ahead inflow forecast (a), and on the evolution of energy prices (b), optimization techniques define the production of electricity over time (i.e., when turbines should be turned on), considering the best hours for production (i.e., when energy prices are high) (b). The volume stored in the reservoir will consequently increase (when turbines are off) or decrease (when turbines are on) (c) and should be managed to respect its capacity constraints (maximum and minimum levels, here represented by dashed red lines) and optimize production (i.e., store water when prices are low and release water to the turbines when prices are high). Rules are often updated on a day-by-day basis. This is a simplified scheme, and several other factors are usually considered in hydropower optimization, such as electricity demand, grid security constraints, the existence of other power plants, infrastructure constraints, turbine capacity, or other water use needs

3 Ensemble Streamflow Forecasts for Operational Systems

The hydropower sector has long recognized the importance of accurate and reliable streamflow forecasts and simulations as input data to the operation and value of production. They have also acknowledged that forecasts are intrinsically uncertain. Since informed decisions with strong social and economic consequences have to be

made, uncertainty needs to be quantified and effectively handled in hydropower operations and planning. It is well known that a single future scenario, no matter how well elaborated it is, is not a good solution to reflect all the possible realities and to base a decision upon. The use of probabilistic- or ensemble-based frameworks in hydrometeorological forecasting, rather than deterministic forecasts, quickly turns out to be an adapted solution to be applied to the modeling challenges of hydropower systems.

In reservoir management and decision-making at seasonal scales, the adoption of probabilistic and ensemble approaches can be dated back to the 1970s. One of the first developers of ensemble approaches for seasonal streamflow forecasting for reservoir operation was the National Weather Service (NWS) California-Nevada River Forecast Center (RFC), which created an operational technique called Ensemble Streamflow Prediction (ESP, which originally stood for Extended Streamflow Prediction; Day 1985; Pica 1997) (An interesting blog post was published in the HEPEX Portal on 26 April 2016, entitled “Tracing the origins of ESP,” by Andy Wood, Tom Pagano, and Maury Roos, with special thanks to Mike Anderson, which can be seen at: <https://hepex.irstea.fr/tracing-the-origins-of-esp/> (last seen on 16/10/2016).) for applications in water supply management. ESP forecasts are obtained by using historical sequences of observed weather (mostly, temperature and precipitation) as forecast input data to a continuous hydrological model, which is run with current initial conditions up to the time of the forecast. The process results in an ensemble that comprises as many members as there are years available in the observation record. Today, there exist many variants of ESP systems (e.g., Gobena and Gan 2010; Wood et al. 2016) for sub-seasonal to seasonal forecasts, and the availability of seasonal weather forecasts and climate outlooks for a wide range of hydrological applications is also expanding.

By recognizing that climate and meteorological uncertainty should be included more explicitly in their forecasting systems, most hydropower producers have also been investing toward enhancing operational systems with state-of-the-art meteorological forecasts from numerical weather prediction (NWP) models. One example of this approach can be found in the history of developments at the French electric power producer EDF in the past decades. At EDF, the use of probabilities in long-term forecasting is present since the 1950s (Desaint et al. 2009). At this time, forecasts for several (2–3) months ahead were produced using simple statistical methods (linear and nonlinear regressions) and were specifically used for the long-term forecasting of inflows to dam reservoirs. Aware of the intrinsic uncertainties of these forecasts, estimates of future inflows were produced and displayed with the help of confidence intervals. Multi-model GCM’s seasonal precipitation forecasts were later explored to improve reservoir management at some months ahead. Garcia-Morales and Dubus (2007) reported that such an “ensemble forecast approach provides useful information for EDF catchments, even with quite low skill, and that a deterministic approach, using only the ensemble mean of the forecasts, is not better than a forecast based on climatology.” Operational forecasters at EDF also acknowledge that a traditional operational practice of probabilistic-based long-term forecasting has facilitated, although much later, the establishment of the EDF 7-day

medium-range streamflow ensemble forecasting system (*personal communication*). In the 1980s–1990s, streamflow forecasts at EDF were based on discharge propagation (hydraulic-based forecasting) in larger river basins and on hydrological (rainfall-runoff) models in smaller catchments. The latter approach used analog-based precipitations as input (i.e., an ensemble of future scenarios created using historic observations of precipitation that were associated with a geopotential field analogous to the forecast one) or deterministic-based numerical weather predictions. In 2008, high-resolution numerical weather predictions and ensemble prediction systems started to be applied in hydrometeorological forecasting (Desaint et al. 2009) (See also the blog post published in the HEPEX Portal on 28 February 2014, entitled “Operational Highlight: use of ensemble hydrometeorological forecasts at EDF (French producer of energy),” contributed by Matthieu Le Lay at: <https://hepex.irstea.fr/operational-use-of-ensemble-hydrometeorological-forecasts-at-edf-french-producer-of-energy/> (last seen on 16/10/2016)).

The use of ensemble streamflow forecasts based on medium-range weather ensemble forecasts or analog approaches for short- to medium-range forecasting at daily time steps has progressed fast in the last decade within hydropower systems. For instance, since 2005, the CNR, a historic producer of hydroelectricity on the Rhone (France) river basin in France, runs a precipitation forecasting system based on an adaptation of model outputs through an analog sorting technique operationally in several catchments (Ben Daoud et al. 2009, 2011). Ensemble streamflow forecasting has also been recently implemented by CEMIG, a major group in the electric energy segment in Brazil, based on weather ensembles from multiple sources and large-scale distributed hydrological modeling (Schwanenberg et al. 2015; Fan et al. 2016).

Despite the examples given above, deterministic hydrologic forecasting is still common in short-term operational practice, while ensemble forecasting is more common at longer lead times in reservoir operations for hydropower (e.g., Weber et al. 2006, 2011; Simard 2011; Crobeddu 2014). When only deterministic forecasts are available, these are often post-processed with bias correction and dressing techniques, allowing to take into account uncertainties from modeling or historical information on possible weather scenarios. Figure 4 illustrates a possible framework for a typical hydropower company producing ensemble streamflow forecasts based on deterministic meteorological forecasts and a hydrological model. The data processing step consists in scanning the observed data for inconsistencies and missing data. Manual data assimilation refers to the interaction between the model and operational forecasters, who can manipulate the state variables and the observed data until achieving a satisfying fit between simulated and observed streamflow. The statistical dressing of the deterministic meteorological forecasts can be based on the incorporation of an error distribution, estimated based on historical records of observed and forecast archives, to the forecasts. An operational example can be in the context of hydropower production found in Crobeddu (2014). It should be noted that the growing complexity of forecasting systems, with the incorporation of multiple data sources from observational networks and advanced techniques such as ensemble forecasts, prompts to the automation of several operations (which would

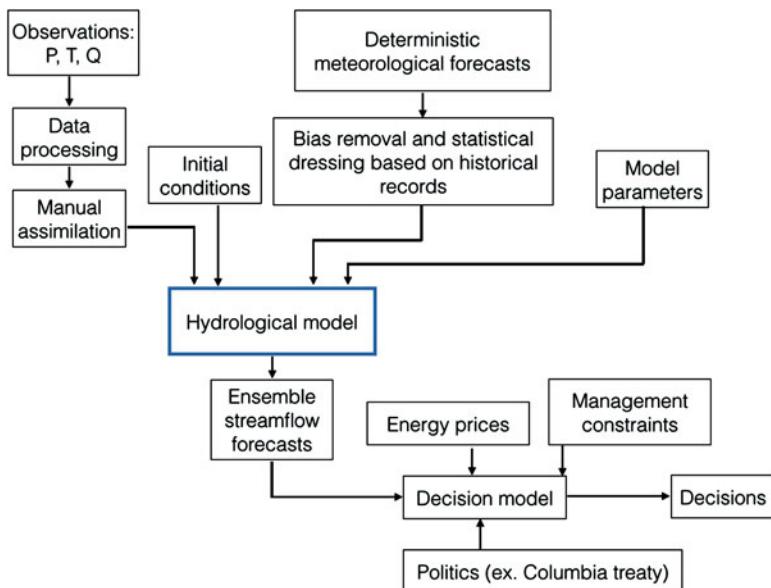


Fig. 4 A typical operational framework for creating streamflow ensembles from deterministic meteorological forecasts in the context of hydropower production

be infeasible in manual-based systems) and the use of computer-aided decision support systems for an increased efficiency of forecasting centers (Pagano et al. 2016).

4 Human Expertise, Quality, and Value of Forecasts

Although forecasters undoubtedly affect the quality of forecasts through data processing, manual data assimilation and any other form of interaction with the hydrological model, their influence can hardly be dissociated from the type of model used and the type of event being forecast. Having in mind this role of human expertise in improving the forecasts, perceptions and definitions of “what is a good forecast?” are crucial. In his 1993 paper, Allan Murphy defines three types of goodness to consider in the overall assessment of forecasts. *Consistency* reflects the agreement between a forecaster’s judgment and the forecast they issue. *Quality* is the agreement between the forecast and the corresponding observation. *Value* relates to the additional benefit to be expected by an end user who would choose to use a specific forecasting system over another. In the context of streamflow forecasting for hydropower systems, it is difficult to isolate one type of goodness, and these three types will be of interest in the evaluation of forecasting systems.

Although forecasters undoubtedly affect the quality of forecasts through data processing, manual data assimilation, and any other form of interaction with the

hydrological model, their influence can hardly be dissociated from the type of model used and the type of event being forecast. Given the high level of complexity and sophistication of forecasting systems for hydropower production, forecaster's judgment is considered an important aspect of the forecasting process. Human expertise plays a crucial role in the forecasting centers of hydropower companies. Usually, forecasters can manually adjust the outputs of the hydrological model and correct inadequacies in the meteorological forcing (both observations and forecasts). Where snow processes dominate the hydrological cycle, and an adequate estimation of the snow water equivalent is key to help dam managers to plan the gradual lowering of the reservoir level before spring melt (e.g., Olsson et al. 2016; Bazile et al. 2017), the forecaster can adjust the simulated snow water equivalent by direct insertion of snow water equivalent measured in situ.

For instance, at BC Hydro, a Canadian electric utility in the province of British Columbia, its forecasting system is qualified as a "manual-interactive" system, meaning that the forecasters can adjust both the inputs and the outputs of the hydrological model, as well as its parameters (Weber et al. 2011). At EDF (France), the notion of "subjective probability" in forecasts (Murphy and Daan 1984), based on expertise applied by the forecaster, was introduced early in their practice of operational forecasting (Garçon et al. 2009). Training and case study analyses were considered as efficient means to help forecasters to calibrate their subjective probabilities (or quantiles), with care to avoid issues of overconfidence, i.e., underestimation of total uncertainties (see Ramos et al. 2010 for an example from EDF forecasts and Mannes and Moore 2013 for an interesting explanation of overconfidence, including a quiz). As noted by Garçon et al. (2009), in order to contribute to make human expert forecasts reliable and to facilitate the production of such forecasts in a routine way, it is essential to provide forecasters with probability forecasts that are also objective, i.e., forecasts that are automatically produced by a probabilistic forecasting system. By expressing forecasts through quantiles or scenarios, forecasters are encouraged to give a more formal indication of forecast uncertainty in the forecasts they issue. Training can play an important role, as well as exchange with forecasters from other organizations or hydrological forecasting services. Simulations and games (e.g., Ramos et al. 2013; Crochemore et al. 2016; Arnal et al. 2016), where the forecaster is confronted with typical decision-making problems as well as the forecast of extreme situations that are rarely seen in daily operational forecasting, can also be a good way to open discussions and enhance human expertise.

The quality of a forecast is routinely evaluated against observations by hydropower producers. There exist many definitions of what is a good forecast in the literature. According to Gneiting and Raftery (2007), a good probabilistic forecast maximizes sharpness subject to calibration. Sharpness represents the extent to which the predictive distribution concentrates around a certain value and is a property of the forecasts only. Calibration, in the context of probabilistic forecasting, refers to the statistical consistency between the probabilistic forecasts and the corresponding observations. It is a joint property of the forecasts and the observations. Hydropower forecasting systems will usually care about issuing forecasts that are accurate, sharp,

and statistically reliable. The main goals of forecast performance assessment are usually the following:

- Compare the performance of forecasts made by different hydrological models.
- Evaluate the skill of a seasonal forecasting system and its ability to provide better forecasts comparatively to a naïve reference such as climatology or the ESP approach.
- Ensure that all main sources of uncertainty are correctly represented in the forecasting system.
- Help the forecasters to adjust their manual interventions on the raw forecasts or to set up post-processing techniques adapted to their needs and system's configuration.
- Identify the components of the forecasting system that most need improvements or additional human and financial efforts.

Forecast quality assessment is usually carried out through the evaluation of statistical scores (e.g., Brier Score, CRPS, reliability diagram, rank histogram, ensemble spread, MAE, correlation, bias, etc.) over a long archive of pairs of forecast and observation, as well as qualitatively, particularly when dealing with selected events, through the visual inspection of forecast and observed hydrographs. Plots with the ensemble streamflow forecasts traces are usually visualized, together with preselected confidence intervals (e.g., the percentiles 10% and 90%) and the median scenario for forecast communication (Ramos et al. 2010).

Finally, the evaluation of the value of a forecast requires that a decision model (or a reservoir management model) be integrated to the hydrometeorological forecasting system in order to evaluate how valuable (in terms of economic benefits) a good forecasting system can be for the business of a hydropower company. For hydropower systems, the economic value of a forecasting system usually depends on observed losses of water (hence, energy that could be produced and sold in the market) and on the economic values of potential power production. The value of the forecasts can be expressed in different ways: for instance, in terms of revenue brought by good decisions when selling energy in the market or in terms of water loss by spilling (and thus not used to produce energy) as a result of a bad forecast or a non-optimal reservoir management. (For some examples, see the blog posts published in the HEPEX Portal: i) on 31 January 2014, entitled “On the economic value of hydrological ensemble forecasts,” contributed by Marie-Amélie Boucher, Maria-Helena Ramos, and Ioanna Zalachori at <https://hepex.irstea.fr/economic-value-of-hydrological-ensemble-forecasts/> and ii) on 17 May 2016, entitled “A never-ending struggle – Improving spring melt runoff forecast via snow information,” contributed by David Gustafsson at <https://hepex.irstea.fr/a-never-ending-struggle-improving-spring-melt-runoff-forecast-via-snow-information/> (last seen on 16/10/2016)). We note that the evaluation of “good” or “bad” decisions is not as straightforward as it appears, even for decision-makers, and measuring the economic consequences of “good” or “bad” forecasts is also a complex exercise for forecasters, managers, and decision-makers. Examples of the assessment of the

value of forecasts in the context of hydropower production can be found in the works of Alemu et al. (2011), Boucher et al. (2012), and Anghileri et al. (2016).

5 Key Issues and Future Challenges

In summary, forecasting for hydropower production is a process involving the forecasting of weather and hydrologic variables at a wide range of space and time scales. Forecasting systems are usually integrated with reservoir management models and decision support tools for electricity production. The planning of operations requires short- (several days) to long-term (several months ahead) streamflow forecasts. The use of ensemble streamflow forecasts in the hydropower sector is growing fast, bringing new challenges and opportunities, mainly in terms of integration of state-of-the-art weather services, data assimilation, forecast quality and value assessment, and risk-based decision-making.

- Enhancing the use of meteorological ensemble predictions and weather services

While most hydropower producers recognize the added benefits of assessing meteorological forecast uncertainty through ensemble prediction systems, many still rely on deterministic meteorological forecasts for short-term streamflow forecasts, and on ESP or analog approaches, based on archived observations and strong assumptions of stationarity, for medium- to long-term forecasts. Meteorological centers throughout the world are however increasingly improving their ensemble prediction systems, and many are now routinely also issuing ensemble monthly and seasonal forecasts. The hydropower sector can benefit from including new products in their forecasting systems, specifically in views of setting up seamless forecasting systems that offer consistent forecasts across space and time scales. In existing sophisticated systems, the need for changing traditional practices and validated techniques can however hamper the introduction of new weather products, as it may require running “old” and “new” systems in parallel until comparisons are established and confidence gained. Challenges remain in comparing competing methods to produce better ensemble predictions and defining optimal ways to better explore the information conveyed by state-of-the-art weather services and risk outlooks.

- Operational implementation of systematic data assimilation methods

In most current operational settings of hydrologic forecasting systems, data assimilation is rather rudimentary and performed manually by human forecasters, based on their expertise. Direct insertion is a common practice, which consists in modifying the value of observed meteorological variables (mostly precipitation and temperature) through a trial and error process until the simulated and the observed streamflow values match. These modifications affect the internal state variables of the hydrological model used to issue the forecast. The modification of inputs is

justified by the uncertainty related to data acquisition (e.g., measurement errors or insufficient network density). However, there exist many different types of systematic data assimilation schemes that could be used to enhance forecasting systems (Liu et al. 2012), and some of them are especially appropriate for ensemble forecasts. The ensemble Kalman filter (e.g., Clark et al. 2008; Trudel et al. 2014; Thiboult et al. 2016) and the particle filter (e.g., Weerts and El Serafy 2006; Leisenring and Moradkhani 2011; Noh et al. 2014) have both been found useful to improve the accuracy of uncertainty estimation, especially for short lead times, and represent open opportunities for ensemble-based hydropower forecasting.

- Further exploring the link between forecast quality and value

Until now, performance assessment of forecasting systems has mostly focused on forecast quality, which evaluates the correspondence between forecasts and observations. However, hydropower production provides the perfect framework for assessing the performance of forecasts also in terms of their economic value. A gain in forecast quality does not always translate into higher forecast value (e.g., Boucher et al. 2012; Anghileri et al. 2016; Côté and Leconte 2016). However, understanding the link between forecast quality and value is important as it can contribute to channelize investments into specific ameliorations of the forecasting system that will impact also its economic value. Forecast value is highly dependent on the case study configuration and the targeted aims of the application. It is therefore important to promote the development of more case studies that explore the complex relationship between quality and value of forecasts in a variety of applications in the hydropower sector.

- Improving the decision-making process and fostering the participation of end-users in streamlining future activities

As outlined in Maier et al. (2014), although there is a growing body of literature regarding the benefit of optimization methods for decision-making, many water resources managers remain reluctant to apply them in their operational practice. The authors also mention that

none of the existing approaches offer a generic and holistic solution for effective and efficient uncertainty propagation during the optimization process. Regarding the decision variables used in optimization-based water resources management approaches; these are almost exclusively modelled as deterministic. However, this often results in rigid, precautionary strategies that may not be sufficiently flexible to adapt to uncertain future changes.

Deterministic forecasts have been progressively replaced by ensemble forecasts as the latter have shown to improve the quality of a forecasting system. There is a call today to also explore the use of ensemble forecasts in optimization models and risk-based decision-making, in order to enhance the value of forecasts for hydropower production.

The challenge here is twofold. Firstly, optimization models need to be more widely used to support risk-based decision-making at all pertinent space and time scales of hydropower production and planning. This involves operationalizing the paradigm shift that has already started about a decade ago and leave the deterministic framework for a probabilistic, uncertain one. Human is by nature uncomfortable with uncertainty (e.g., Kahneman et al. 1982), and, therefore, achieving this paradigm shift is not at all a trivial task. The active involvement of operational forecasters and end-users in the process is essential in this regard (e.g., Wetterhall et al. 2013). Secondly, the choice of a particular optimization model has to be such that the model is able to explicitly handle uncertainty. This uncertainty comes from the future ensemble streamflow forecasts or inflows to reservoirs, but also from market prices of energy and demand forecasts. In addition, future reservoir optimization tools should account more explicitly for the multipurpose vocation of many hydropower dams (e.g., Tilmant et al. 2008). Today, the water-energy nexus is an opportunity for the hydropower sector to promote a more integrated and cost-effective way to approach water releases and allocation in a cross-sectorial planning and strategic decision-making.

References

- N. Addor, S. Jaun, F. Fundel, M. Zappa, An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.* **15**, 2327–2347 (2011). <https://doi.org/10.5194/hess-15-2327-2011>
- M. Akabari, A. Afshar, S.J. Mousavi, Multi-objective reservoir operation under emergency condition: Abbaspour reservoir case study with non-functional spillways. *J. Flood Risk Manage.* **7**, 374–384 (2014)
- E.T. Alemu, R.N. Palmer, A. Polebitski, B. Meaker, Decision support system for optimizing reservoir operations using ensemble streamflow prediction. *J. Water Resour. Plan. Manag.* **137**(1), 72–82 (2011)
- D. Anderson, H. Moggridge, P. Warren, J. Shucksmith, The impacts of ‘run-of-river’ hydropower on the physical and ecological condition of rivers. *Water Environ. J.* **29**, 268–276 (2015)
- D. Anghileri, A. Castelletti, F. Pianosi, R. Soncini-Sessa, E. Weber, Optimizing watershed management by coordinated operation of storing facilities. *J. Water Resour. Plan. Manag.* **139**, 492–500 (2013). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000313](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000313)
- D. Anghileri, N. Voisin, A. Castelletti, F. Pianosi, B. Nijssen, D.P. Lettenmaier, Value of long-term streamflow forecasts to reservoir operations for water supply in snow-dominated river catchments. *Water Resour. Res.* **52**(6), 4209–4225 (2016). <https://doi.org/10.1002/2015WR017864>
- L. Arnal, M.-H. Ramos, E. Coughlan, H.L. Cloke, E. Stephens, F. Wetterhall, S.J. van Andel, F. Pappenberger, Willingness-to-pay for a probabilistic flood forecast: A risk-based decision-making game. *Hydrol. Earth Syst. Sci.* **20**, 3109–3128 (2016). <https://doi.org/10.5194/hess-20-3109-2016>
- R. Bazile, M.-A. Boucher, L. Perreault, R. Leconte, Verification of ECMWF System 4 for seasonal hydrological forecasting in a northern climate. *Hydrol. Earth Syst. Sci.* **20**, 5747–5762 (2017)
- A. Ben Daoud, E. Sauquet, M. Lang, C. Obled, G. Bontron, La prévision des précipitations par recherche d’analogues: état de l’art et perspectives. *La Houille Blanche* **6**, 60–65 (2009). <https://doi.org/10.1051/lhb/2009079> [in French]
- A. Ben Daoud, E. Sauquet, M. Lang, M.-H. Ramos, Peut-on étendre l’échéance de prévision des crues en optimisant la prévision de pluies par recherche d’analogues? Application au bassin de

- la Seine à Paris. *La Houille Blanche* **1**, 37–43 (2011) <https://doi.org/10.1051/lhb/2011004>, [in French]
- A. Björnsen Gurung, A. Borsdorf, L. Füreder, F. Kienast, P. Matt, C. Scheidegger, L. Schmocker, M. Zappa, K. Volkart, Rethinking pumped storage hydropower in the European Alps. *Mt. Res. Dev.* **36**, 222–232 (2016)
- M.-A. Boucher, R. Leconte, Changements climatiques et production hydroélectrique canadienne: où en sommes-nous? *Can. Water Res. J./Revue canadienne des ressources hydriques* **38**(3), 196–209 (2013), [in French]
- M.-A. Boucher, D. Tremblay, L. Delorme, L. Perreault, F. Anctil, Hydro-economic assessment of hydrological forecasting systems. *J. Hydrol.* **416–417**, 133–144 (2012)
- M.P. Clark, D.E. Rupp, R.A. Woods, X. Zheng, R. Ibbitt, A. Slater, J. Schmidt, M. Uddstrom, Hydrological data assimilation with the ensemble Kalman filter: Use of streamflow observations to update states in a distributed hydrological model. *Adv. Water Resour.* **31**, 1309–1324 (2008)
- P. Côté, R. Leconte, Comparison of stochastic optimization algorithms for hydropower reservoir operation with ensemble streamflow prediction. *J. Water Resour. Plan. Manag.* **142**(2), 04015046 (2016). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000575](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000575)
- Crobeddu, Prévisions d'apports probabilistes à Hydro-Québec production – Problématiques, orientations et réalisations, Workshop of the Canadian Water Resources Association in Des prévisions hydrologiques opérationnelles vers une optimisation de la gestion des réservoirs, Québec, 17–19 Sep 2014. Available at http://acrhta2014.ouranos.ca/pdf/Session4_EricCrobeddu.pdf (2014)
- L. Crochemore, M.-H. Ramos, F. Pappenberger, S. van Andel, A. Wood, An experiment on risk-based decision-making in water management using monthly probabilistic forecasts. *Bull. Am. Meteorol. Soc.* **97**(4), 541–551 (2016). <https://doi.org/10.1175/BAMS-D-14-00270.1>
- G. Day, Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan. Manag.* **111**(2), 157–170 (1985). [https://doi.org/10.1061/\(ASCE\)0733-9496\(1985\)111:2\(157\)](https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157))
- B. Desaint, P. Nogues, C. Perret, R. Garçon, La prévision hydrométéorologique opérationnelle: l'expérience d'Électricité de France. *La Houille Blanche* **5**, 39–46 (2009), [in French]
- F.M. Fan, D. Schwanenberg, R. Alvarado, A.A. dos Reis, W. Collischonn, S. Naumann, Performance of Deterministic and probabilistic hydrological forecasts for the short-term optimization of a tropical hydropower reservoir. *Water Resour. Manag.* **30**, 3609–3625 (2016a)
- F.M. Fan, R.C.D. Paiva, W. Collischonn, Chapter 2: Hydrological forecasting practices in Brazil, in *Flood Forecasting – A Global Perspective*, ed. by T. E. Adams III, T. C. Pagano, (Academic, 2016), 433 p. isbn:978-0-12-801884-2
- B. François, M. Borga, S. Anquetin, J.D. Creutin, K. Engeland, A.C. Favre, B. Hingray, M.H. Ramos, D. Raynaud, B. Renard, E. Sauquet, J.F. Sauterleute, J.P. Vidal, G. Warland, Integrating hydropower and intermittent climate-related renewable energies: A call for hydrology. *Hydrol. Process.* **28**(21), 5465–5468 (2014). <https://doi.org/10.1002/hyp.10274>
- B. François, B. Hingray, D. Raynaud, M. Borga, J.-D. Creutin, Increasing climate-related-energy penetration by integrating run-of-the river hydropower to wind/solar mix. *Renew. Energy* **87**, 686–696 (2016)
- M.B. Garcia-Morales, L. Dubus, Forecasting precipitation for hydroelectric power management: How to exploit GCM's seasonal ensemble forecasts. *Int. J. Climatol.* **27**, 1691–1705 (2007). <https://doi.org/10.1002/joc.1608>
- R. Garçon, B. Houdant, F. Garavaglia, T. Mathevot, E. Paquet, J. Gailhard, Expertise humaine des prévisions hydrométéorologiques et communication de leurs incertitudes dans un contexte décisionnel. *La Houille Blanche* **5**, 71–80 (2009), [in French]
- T. Gneiting, A. Raftery, Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007)
- A.K. Gobena, T.Y. Gan, Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. *J. Hydrol.* **385**(1–4), 336–352 (2010)
- C.B. Harto, Y.E. Yan, Y.K. Demisse, D. Elcock, V.C. Tidwell, K. Hallet, J. Macknick, M.S. Wigmosta, T.K. Tesfa, Analysis of drought impacts on electricity production in the Western

- and Texas interconnections of the United States. Technical report, U.S. Department of Energy, Office of Electricity Delivery and Energy Reliability, (2011), 161 pp. <http://energy.sandia.gov/wp-content/gallery/uploads/Drought-Analysis-Report-Final.pdf>
- F. Hendrickx, E. Sauquet, Impact of warming climate on water management for the Ariege River basin (France). *Hydrol. Sci. J.* **59**(5), 976–993 (2013)
- D. Kahneman, P. Slovic, A. Tversky, *Judgement Under Uncertainty: Heuristics and Biases* (Cambridge University Press, Cambridge, 1982), 544 p
- B. Kiani, A. Rowe, P. Wild, L. Pitt, A. Sopinka, T.F. Pedersen, Optimal electricity system planning in a large hydro jurisdiction: Will British Columbia soon become a major importer of electricity? *Energ Policy* **54**, 311–319 (2013)
- A. Kumar, T. Schei, A. Ahenkorah, R. Caceres Rodriguez, J.-M. Devernay, M. Freitas, D. Hall, Å. Killingtveit, Z. Liu, Hydropower, in *IPCC Special Report on Renewable Energy Sources and Climate Change Mitigation*, ed. by O. Edenhofer, R. Pichs-Madruga, Y. Sokona, K. Seyboth, P. Matschoss, S. Kadner, T. Zwickel, P. Eickemeier, G. Hansen, S. Schlömer, C. von Stechow, (Cambridge University Press, Cambridge, UK/New York, 2011). Available at: http://srren.ipcc-wg3.de/report/IPCC_SRREN_Ch05.pdf
- M. Leisenring, H. Moradkhani, Snow water equivalent prediction using Bayesian data assimilation methods. *Stoch. Env. Res. Risk A.* **25**, 253–270 (2011)
- Y. Liu, A.H. Weerts, M. Clark, H.-J. Hendricks Franssen, S. Kumar, H. Moradkhani, D.-J. Seo, D. Schwanenberg, P.J. Smith, A.I.J.M. van Dijk, N. van Velzen, M. He, H. Lee, S.J. Noh, O. Rakovec, P. Restrepo, Advancing data assimilation in operational hydrologic forecasting: Progresses, challenges, and emerging opportunities. *Hydrol. Earth Syst. Sci.* **16**, 3863–3887 (2012). <https://doi.org/10.5194/hess-16-3863-2012>
- M. Lu, U. Lall, A.W. Robertson, E. Cook, Optimizing multiple reliable forward contracts for reservoir allocation using multitime scale streamflow forecasts. *Water Resour. Res.* **53**, 2035–2050 (2017)
- M.H. Maier, Z. Kapelan, J. Kasprzyk, J. Kollat, L.S. Matott, M.C. Cunha, G.C. Dandy, M.S. Gibbs, E. Keedwell, A. Marchi, A. Ostfeld, D. Savic, D.P. Solomatine, J.A. Vrugt, A.C. Zecchin, B.S. Minsker, E.J. Barbour, G. Kuczera, F. Pasha, A. Castelletti, M. Giuliani, M.P. Reed, Evolutionary algorithms and other metaheuristics in water resources: Current status, research challenges and future directions. *Environ. Model. Softw.* **62**, 271–299 (2014)
- A. Mannes, D. Moore, I know I'm right – A behavioural view of overconfidence. *Significance* **10**(4), 10–14 (2013). <https://doi.org/10.1111/j.1740-9713.2013.00674.x>
- A.H. Murphy, What is a good forecast – An essay on the nature of goodness in weather forecasting. *Weather Forecast.* **8**(2), 281–293 (1993)
- A.H. Murphy, H. Daan, Impacts of feedback and experience on the quality of subjective probability forecast: Comparison of results from the first and second years of the Zierikzee experiment. *Mon. Weather Rev.* **112**, 413–442 (1984)
- S.J. Noh, O. Rakovec, A. Weerts, Y. Tachikawa, On noise specification in data assimilation schemes for improved flood forecasting using distributed hydrological models. *J. Hydrol.* **519**, 2707–2721 (2014)
- J. Olsson, C.B. Uvo, K. Foster, W. Yang, Technical Note: Initial assessment of a multi-method approach to spring-flood forecasting in Sweden, *Hydrol. Earth Syst. Sci.* **20**, 659–667 (2016)
- T.C. Pagano, F. Pappenberger, A.W. Wood, M.-H. Ramos, A. Persson, B. Anderson, Automation and human expertise in operational river forecasting. *WIREs Water* **3**(5), 692–705 (2016). <https://doi.org/10.1002/wat2.1163>
- J. Pica, Review of extended streamflow prediction of the National Weather Service NWSRFS ESP, in *CE505 Conference Course, Civil Engineering*, Portland State University, 1 July 1997 (1997)
- M.H. Ramos, T. Mathevret, J. Thielen, F. Pappenberger, Communicating uncertainty in hydro-meteorological forecasts: Mission impossible? *Meteorol. Appl.* **17**, 223–235 (2010)
- M.H. Ramos, S.J. Van Andel, F. Pappenberger, Do probabilistic forecasts lead to better decisions? *Hydrol. Earth Syst. Sci.* **17**(6), 2219–2232 (2013). <https://doi.org/10.5194/hess-17-2219-2013>

- B. Schaeefli, Projecting hydropower production under future climates: A guide for decision-makers and modelers to interpret and design climate change impact assessments. *WIREs Water* **2**, 271–289 (2015). <https://doi.org/10.1002/wat2.1083>
- D. Schwanenberg, F.M. Fan, S. Naumann, J.I. Kuwajima, R.A. Montero, A. Assis dos Reis, Short-term reservoir optimization for flood mitigation under meteorological and hydrological forecast uncertainty - application to the Três Marias Reservoir in Brazil. *Water Resour. Manag.* **29**(5), 1635–1651 (2015)
- S. Séguin, C. Audet, P. Côté, Scenario-Tree modeling for stochastic short-term hydropower operation planning. *J. Water Resour. Plan. Manag.* **143** (2017). [https://doi.org/10.1061/\(ASCE\)WR.1943-5452.0000854](https://doi.org/10.1061/(ASCE)WR.1943-5452.0000854)
- M.-C. Simard, Inflow forecast at Hydro-Québec Production, 2011 CSHS operational forecasting workshop, Vancouver, 6 Oct 2011. Available at http://www.cwra.org/images/BranchesAfiliates/CSHS/mcsimard_presentation_vfinale.pdf (2011)
- G. Tang, H. Zhou, N. Li, F. Wang, Y. Wang, J. Deping, Value of medium-range precipitation forecasts in inflow prediction and hydropower optimization. *Water Resour. Manag.* **24**, 2721–2742 (2010)
- A. Thiboult, F. Anctil, M.-A. Boucher, Accounting for three sources of uncertainty in ensemble hydrological forecasting. *Hydrol. Earth Syst. Sci.* **20**, 1809–1825 (2016)
- A. Tilmant, D. Pinte, Q. Goor, Assessing marginal water values in multipurpose multireservoir systems via stochastic programming. *Water Resour. Res.* **44**, W12431 (2008). <https://doi.org/10.1029/2008WR007024>
- M. Trudel, R. Leconte, C. Paniconi, Analysis of the hydrological response of a distribute physically-based model using post-assimilation (EnKF) diagnostics of streamflow and in situ soil moisture observations. *J. Hydrol.* **514**, 192–201 (2014)
- U.S. Department of Energy, *The Water-Energy Nexus: Challenges and Opportunities Overview and Summary*, 12p. Available at <http://energy.gov/sites/prod/files/2014/07/f17/Water%20Energy%20Nexus%20Executive%20Summary%20July%202014.pdf> (2014)
- M.T.H. Van Vliet, J. Sheffield, D. Wiberg, E.F. Wood, Impact of recent drought and warm years on water resources and electricity supply worldwide. *Environ. Res. Lett.* **11** (2016). <https://doi.org/10.1088/1748-9326/11/12/124021>
- F. Weber, L. Perreault, V. Fortin, Measuring the performance of hydrological forecasts for hydropower production at BC Hydro and Hydro-Québec, in *Proceeding of the 18th Conference on Climate Variability and Change*, AMS, Atlanta, 30 Jan–2 Feb AMS, Boston, 8.5 (2006)
- F. Weber, D. Omikunkle, S. Weston, A. Gobena, The ensemble river forecasting system: Towards gaining certainty in uncertainty, CSHS operational forecasting workshop (2011)
- A.H. Weerts, G.Y. El Serafy, Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models. *Water Resour. Res.* **42**(9), W09403 (2006). <https://doi.org/10.1029/2005WR004093>
- F. Wetterhall, F. Pappenberger, L. Alfieri, H.L. Cloke, J. Thielen-del Pozo, S. Balabanova, J. Daňhelka, A. Vogelbacher, P. Salamon, I. Carrasco, A.J. Cabrera-Tordera, M. Corzo-Toscano, M. García-Padilla, R.J. García-Sánchez, C. Ardilouze, S. Jurela, B. Terek, A. Csik, J. Casey, G. Stankūnavičius, V. Ceres, E. Sprokkereef, J. Stam, E. Anghel, D. Vladikovic, C. Alionte Eklund, N. Hjerdt, H. Djerv, F. Holmberg, J. Nilsson, K. Nyström, M. Sušnik, M. Hazlinger, M. Holubecka, HESS opinions “Forecaster priorities for improving probabilistic flood forecasts”. *Hydrol. Earth Syst. Sci.* **17**, 4389–4399 (2013)
- A.W. Wood, T. Hopson, A. Newman, J. Arnold, L. Brekke, M. Clark, Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeorol.* **17**(2), 651–668 (2016). <https://doi.org/10.1175/JHM-D-14-0213.1>



Hydropower Forecasting in Brazil

Carlos E. M. Tucci, Walter Collischonn, Fernando Mainardi Fan,
and Dirk Schwanenberg

Contents

1	Introduction	1308
2	Methodology	1309
2.1	Observation Network	1310
2.2	The Global Numerical Weather Prediction Model and Ensemble Characteristics	1311
2.3	Hydrological Modeling Framework	1312
2.4	Experiment Setup and Forecast Verification	1313
3	Results	1313
4	Conclusion	1325
	References	1326

Abstract

Most of the electric power in Brazil comes from hydropower, and the short-term power production at each of the major power plants in Brazil is optimized using streamflow forecasts of lead times up to 14 days. These forecasts were usually obtained using stochastic models based only on the last observed streamflow values. During the last few years, rainfall–runoff models that use predicted rainfall as the main input start to replace the stochastic models. However, this new model generation still uses deterministic precipitation forecasts and does not take advantage of the ensemble precipitation forecasts that are already available in Brazil from regional and global meteorological models. Based on recent

C. E. M. Tucci (✉) · W. Collischonn · F. M. Fan

Institute of Hydraulic Research, Federal University of Rio Grande do Sul, Porto Alegre-RS, Brazil
e-mail: rhamaca@gmail.com; collischonn@uol.com.br; fernando.fan@ufrgs.br

D. Schwanenberg

Institute of Hydraulic Engineering and Water Resources Management, Universität Duisburg-Essen, Essen, Germany
e-mail: dirk.schwanenberg@uni-due.de

research results, it is likely that ensemble streamflow forecasts outperform deterministic forecasts in application to short-term reservoir management for objectives such as energy generation and flood mitigation. This chapter presents an assessment of 4 years of ensemble inflow forecasts to a major hydropower reservoir in Brazil, the Três Marias dam, on the São Francisco River. A 14 member ensemble obtained from a global numerical weather prediction model of the Brazilian Center for Weather Prediction is used, and results are evaluated in terms of ensemble applicability for a period between 2008 and 2012. Results are encouraging, and due to this it is believed that ensemble inflow forecasts to major reservoirs in Brazil will be used in a near future as input to the optimization of the national electric power producing system.

Keywords

Ensemble inflow forecasts · Reservoir operation · Brazil

1 Introduction

Hydropower is the most important source of electricity in Brazil. During recent years, this source accounted for 80–90% of total electric power supply (EPE 2013). A national-wide transmission network allows the integrated management of the energy production of hydropower plants and other sources, giving relative flexibility in prioritizing energy production at power plants that show lower production costs. This management is done by a central organization called *Operador Nacional do Sistema* (ONS) that has the objective of optimizing electric energy production. Operational costs of hydropower production are lower than for thermoelectric plants; therefore there is a strong economic reason for maximizing the proportion of energy generated from hydropower (Hamlet et al. 2002). On the other hand, hydropower is dependent on climate, which has its natural variability, leading to risks of power production shortage in the future, which have to be avoided.

In such a complex system, management benefits of prior knowledge of reservoir inflows include the following: (1) spillage can be minimized; (2) reservoirs can operate with greater head of water for longer periods; (3) more energy can be generated at times when energy prices are higher; and (4) more energy can be produced at hydropower plants that have the higher inflow forecasts (Faber and Stedinger 2001; Yeh et al. 1982; Hamlet et al. 2002; Maurer 2002). Forecast errors influence decision-making, leading to suboptimal operation when water is released unnecessarily from reservoirs or when thermal power plants are activated needlessly.

ONS uses a chain of optimization models for the management of this system (Maceira et al. 2002; Pereira and Pinto 1991). One optimization model used by ONS is applied for making operational decisions up to 14 days in advance, and it uses forecasts of inflow to more than 200 reservoirs. Until less than a decade ago, all the inflow forecasts were provided by periodic auto-regressive moving average (PARMA) models (Costa et al. 2014; Maceira and Damázio 2005). In 2005, ONS

started to test inflow forecasting based on rainfall–runoff models with the input of quantitative precipitation forecasts. The forecasting models tested were of different types of rainfall–runoff models that use precipitation forecasts generated by the Brazilian Weather Forecasting Center CPTEC (*Centro de Previsão de Tempo e Estudos Climáticos*), primarily the regional Eta model (Cataldi et al. 2007; Chou et al. 2002). An assessment of results by Guilhon et al. (2007) shows that the new forecasting methods that include quantitative precipitation forecasts outperform auto-regressive forecasting models based on inflow time series only. Following these results, forecasting models and methods that use quantitative precipitation forecasts as input information are gradually replacing PARMA models in the Brazilian electric energy optimization (Guilhon et al. 2007). However, until now all medium-range (14 days) forecasts are deterministic.

In Brazil, the use of ensemble streamflow forecasts, both for floods and for reservoir inflow predictions, is still starting. Tucci et al. (2003) and Tucci et al. (2008), who show results of hindcasting experiments of seasonal streamflow forecasts for the rivers Uruguay and Grande, made some of the first efforts in this direction. More recently, Calvetti (2011), Collischonn et al. (2012), Collischonn et al. (2013), and Meller (2013) describe experiments of short- to medium-range ensemble flood forecasting; however none of the proposed systems are being used operationally. Only Fan et al. (2014, 2015a) describe operational systems.

In the Brazilian scenario, ensemble forecasts could be integrated with the currently used optimization methods within ONS, using a tree structure to transfer ensemble forecasts to multistage stochastic programming optimization methods, like suggested by Raso et al. (2013) and Schwanenbergh et al. (2015). Furthermore, ensemble forecasts will be probably more useful in predicting extreme events, which normally do not have a significant influence on the daily optimization problem, but pose a challenge on the dam safety and people living upstream and downstream along the river.

This chapter describes an experimental use of ensemble forecasts to predict inflow to a major reservoir on the river São Francisco, in the Southeast Brazil. Inflow forecasts of 15 days of lead time were generated twice a day using a 14 member ensemble obtained from the global numerical weather prediction run by the Brazilian Weather Forecasting Center (CPTEC) and a large-scale hydrological model. Hindcasting results from 2008 to 2012 were evaluated by qualitative and quantitative assessment.

2 Methodology

The presented case study deals with the São Francisco river basin, upstream of the Três Marias dam. It has a drainage area of approximately 50,000 km². The watershed is located in the range of 18–22 degrees South, in Southeast Brazil, in a region with a wet season that occurs during the austral summer (November to April) and a dry season that occurs during austral winter (May to October). The basin receives around

1400 mm of annual rainfall in the southern parts and around 1000 mm in the northern parts. Concentration time in the watershed is around 2 days.

The Três Marias dam was built during the 1950s. Its reservoir has a total capacity of $19.5 \times 10^9 \text{ m}^3$ and is used to regulate streamflow of the river São Francisco to improve power generation at the Três Marias power plant and four other major hydropower plants located downstream. The city of Pirapora is located nearly 120 km downstream of the dam and suffers occasionally with floods. Therefore, it is required that the operation should not amplify naturally occurring floods in the city of Pirapora. To support decision-making, forecasts are needed both for the reservoir inflow and for streamflow of tributaries located downstream, as shown by Fan et al. (2014).

Ensemble streamflow forecasts of inflow to the Três Marias reservoir were obtained retrospectively, in hindcasting mode, for a period extending from July 2008 until July 2012, using a large-scale rainfall–runoff model with input data from observation network and from quantitative precipitation forecasts obtained from a global ensemble atmospheric model of the Brazilian Weather Forecasting Center (BWFC). Forecasts were obtained in hourly time steps ranging to 360 h (15 days) in advance and were compared with observed inflow.

2.1 Observation Network

For Brazilian standards, the São Francisco river basin upstream of the Três Marias dam has a relatively good coverage of rainfall and streamflow gauges (Fig. 1). Observed rainfall data from seven gauging stations are used as input data for the hydrological model. In the developed forecasting system, only the gauging stations with hourly data and with automatic transmission of data are used, as would be the case in operational streamflow forecasting.

The streamflow observation network in the basin consists of the same seven gauging stations, with hourly data collection and automatic transmission. The stations are located on the main rivers flowing into the reservoir, as can be seen in Fig. 1. Drainage area information for each gauging station is given in Table 1.

Inflow to the reservoir is estimated indirectly by water balance, based on measurements of outflow through turbines and spillway, and measurements of forebay elevation of the reservoir. This method, when applied on hourly time steps, usually results in noisy time series. The observed inflow hydrograph from 2008 to 2012 is shown in Fig. 2. The average inflow is approximately $600 \text{ m}^3 \text{s}^{-1}$. The hydrograph shows peak inflow values as high as $6000 \text{ m}^3 \text{s}^{-1}$, and dry season flows as low as $100 \text{ m}^3 \text{s}^{-1}$. High flows occur usually from December to March, while the lowest flows occur normally in August or September.

Figure 2 also shows that the estimated inflow in hourly time step is highly uncertain, due to the noise resulting from the estimation method by reservoir water balance. A 12 h moving average filter was used to reduce the noise in the inflow data, but this was not sufficient to remove it, as can be seen in Fig. 2, which actually shows filtered data.

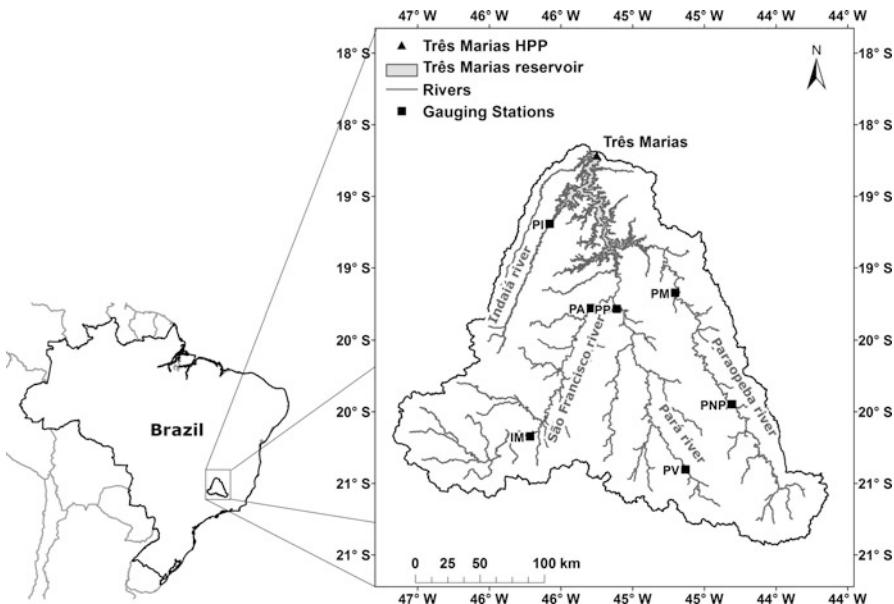


Fig. 1 Map of the São Francisco river basin upstream of the Três Marias reservoir, showing the location of the reservoir, the most important rivers, and the gauging stations used in the ensemble inflow forecasts

Table 1 Summary of results of the hydrological model during the calibration period at seven river gauging stations (from December 2006 to June 2011)

Gauging station	Drainage area (km^2)	NS	NSlog	ΔV
PNP	5784	0.77	0.89	-9.9
PM	10450	0.81	0.93	-3.6
PV	2619	0.80	0.86	8.0
PP	11358	0.56	0.63	-14.9
IM	5485	0.88	0.90	-1.8
PA	14244	0.86	0.90	-6.6
PI	2208	0.50	0.79	6.1

In the study presented in this chapter, this reservoir inflow information was used to compare the ensemble forecasting results. Therefore, it has to be noted that observed inflow has also uncertainties.

2.2 The Global Numerical Weather Prediction Model and Ensemble Characteristics

In this study, quantitative precipitation forecasts (QPFs) from the atmospheric global circulation model (AGCM) (Mendonça and Bonatti 2009; et al. 2008) provided by

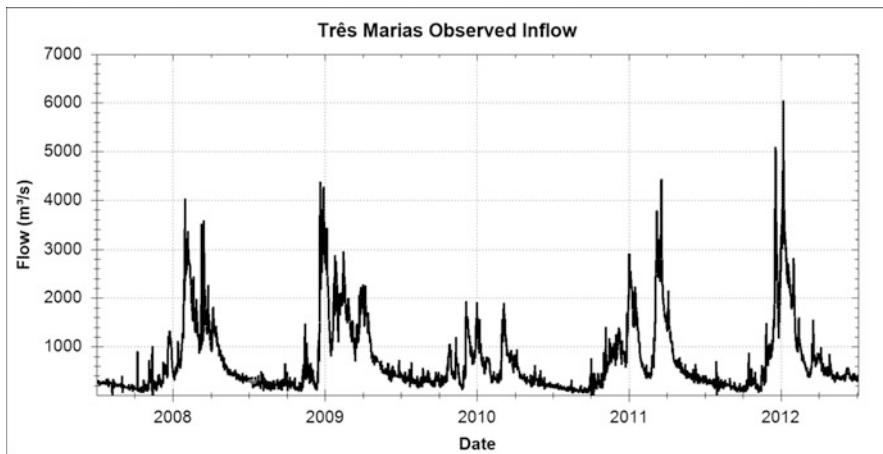


Fig. 2 Time series of estimated inflow to the Três Marias reservoir after filtering using a 24 h moving average

the Brazilian CPTEC, with 15 days of lead time, are used as meteorological inputs in the forecasting system. The data used in the assessment was obtained from The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE) project portals (Bougeault et al. 2010). The precipitation forecasts consist of 14 members of perturbed precipitation produced by CPTEC for the whole globe, with horizontal grid resolution of approximately 0.9375° , available twice a day (00:00:00 UTM and 12:00:00 UTM). The time step of the available QPF is 6 h, which were disaggregated equally to hourly to be used as hydrological model input. The information was also spatially downscaled to the watershed domain using a Thiessen polygon approach.

2.3 Hydrological Modeling Framework

The MGB-IPH (*Modelo de Grandes Bacias – Instituto de Pesquisas Hidráulicas*) hydrological model (Collischonn et al. 2007a) was used to conduct the generation of ensemble inflow forecasts for the Três Marias reservoir. MGB-IPH is a large-scale distributed hydrological model that calculates streamflow from precipitation data. The model has been applied in several different South American river basins, like the Amazon (Paiva et al. 2013), the Paraná (Fan et al. 2012b), the Grande (Tucci et al. 2008; Bravo et al. 2009), and the Uruguay (Collischonn et al. 2005). The model is also currently being used operationally for streamflow forecasts in the Paranaíba river basin (Collischonn et al. 2007b), the Pelotas river basin (Fan et al. 2012a), and the Tocantins river basin (Fan et al. 2015a).

For the application in the São Francisco river basin described here, the model was calibrated considering hourly time steps using the rainfall and streamflow data of the observation network. The selected period for calibration was December 2006 to June

2011. Table 1 shows the Nash–Sutcliffe (NS) model efficiency coefficient, the Nash–Sutcliffe model efficiency coefficient for logarithms of streamflow (NSlog), and the volume error of the model after calibration for the stations located in the Três Marias basin.

Table 1 indicates that better results in terms of the Nash–Sutcliffe efficiency are usually obtained at places with larger drainage area. The exception is Porto Pará (PP) gauging station, which is located downstream of a small reservoir, for which the release data is not taken into account in the model. The lowest Nash–Sutcliffe efficiency is found at the Porto Indaiá (PI) gauging station. The low value can be explained by the poor rainfall gauging in this part of the basin, where the only gauge is located at the same place as the river gauge (the basin outlet).

In the context of real-time forecasting, the model employs a technique of model updating. This technique is a feedback process in which the most recent streamflow observations are used to correct the initial conditions of river streamflow and groundwater storage (Paz et al. 2007; Meller et al. 2012).

Reservoir inflow data is not assimilated into the hydrological model. To obtain inflow forecasts to the reservoir, an output correction of the model results is therefore applied based on an auto-regressive (AR) model, which corrects the forecasts values based in the last observed values prior to the forecast start.

2.4 Experiment Setup and Forecast Verification

The experiment conducted and presented here consists of the use of QPF from the global meteorological model to run the MGB-IPH model calibrated to the Três Marias basin for a period of 4 years from July 2008 to July 2012.

Streamflow ensemble results are generated twice a day, during the whole 4 years of period. Results of reservoir inflows are evaluated by comparing every 15 day (360 h) forecast to the corresponding observed inflow data, by qualitative and quantitative assessment, using metrics that are commonly used for ensemble assessments (Brown et al. 2010; Bradley and Schwartz 2011; Hersbach 2000; Jolliffe and Stephenson 2012; Stanski et al. 1989; Wilks 2006).

3 Results

Five different forecasts are discussed here in more detail using graphs of spaghetti plots of individual ensemble member forecasts. The first visual analysis is for the forecast issued on 18 December 2008 at 00 UTC (Fig. 3). At this moment, the inflow to the Três Marias reservoir was around $1000 \text{ m}^3\text{s}^{-1}$ and declining after a slightly higher peak that occurred 1 day before. Forecasts based on all the 14 members show a substantial increase of inflow in the following days, peaking at a range between 3200 and $4000 \text{ m}^3\text{s}^{-1}$. The ensemble members also show a second peak, 8 days after the first one, with a much wider peak range between 2000 and $10,000 \text{ m}^3\text{s}^{-1}$.

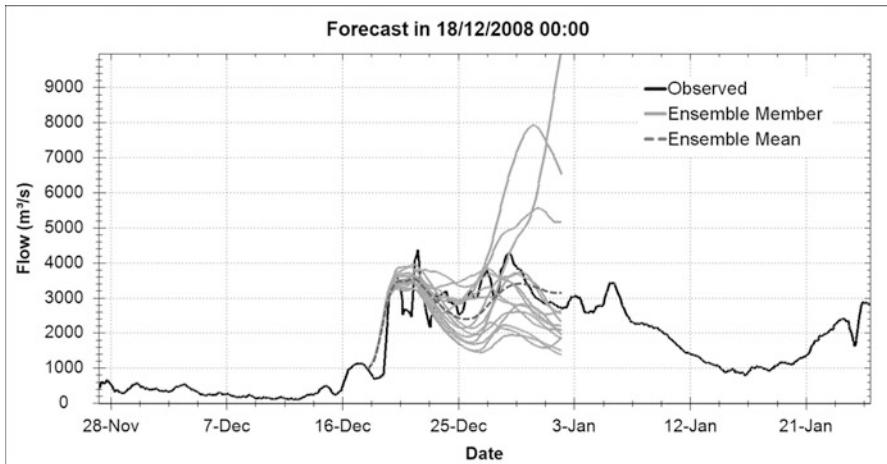


Fig. 3 Forecast of the major flood of the rainy season in 2008/2009, starting at 00 UTC 18 December 2008

Observations are well encompassed by the ensemble spread, and the ensemble mean presents a relatively good forecast.

Figure 3 shows that the increase of inflow from around 1000 to 3000–4000 m^3s^{-1} in about 2 days was correctly predicted by all members. The relatively narrow spread of the ensemble may be because part of the rainfall that caused the inflow increase had already fallen before the start of the forecast. Another interesting feature of the forecast example shown in Fig. 3 is that a second inflow peak actually occurred around 30 December, and this peak was captured by most of the members, with greater values for some of them and smaller for others.

Considering the potential use of ensemble inflow forecast for the optimization of the Brazilian power grid, the forecast shown in Fig. 3 can be considered as very successful, since the total volume flowing into Três Marias reservoir in the next 2 weeks was well forecasted.

Figure 4 shows the case of the forecast issued at 01 December 2009 at 00 UTC. At the time of the forecast, inflow to the Três Marias reservoir was near to $500 \text{ m}^3\text{s}^{-1}$, while 5 days later inflow peaked at just less than $2000 \text{ m}^3\text{s}^{-1}$. In this case, the peak was mainly formed by rainfall that occurred after the forecast was issued, i.e., after 01 December, and the figure shows that the ensemble members agreed on the timing and magnitude of the inflow increase until 06 December, when the observed peak occurred.

Figure 5 shows the case of the forecast issued on 02 March 2011 at 00 UTC. At the time of the forecast, inflow to the Três Marias reservoir was around $700 \text{ m}^3\text{s}^{-1}$, while 9 days later the same inflow peaked at almost $3800 \text{ m}^3\text{s}^{-1}$. The observed hydrograph was totally covered by the ensemble members, with most members indicating inflow rising until 09 March.

As in the previous case, the occurrence of an inflow increase was consistent between members, especially until 07 March, when inflow increase actually

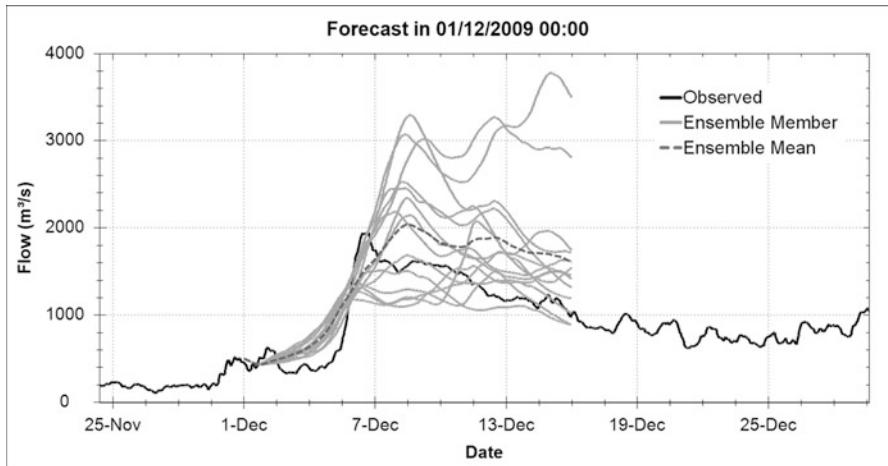


Fig. 4 Forecast of the major flood December 2009, starting at 00 UTC 01 December 2009

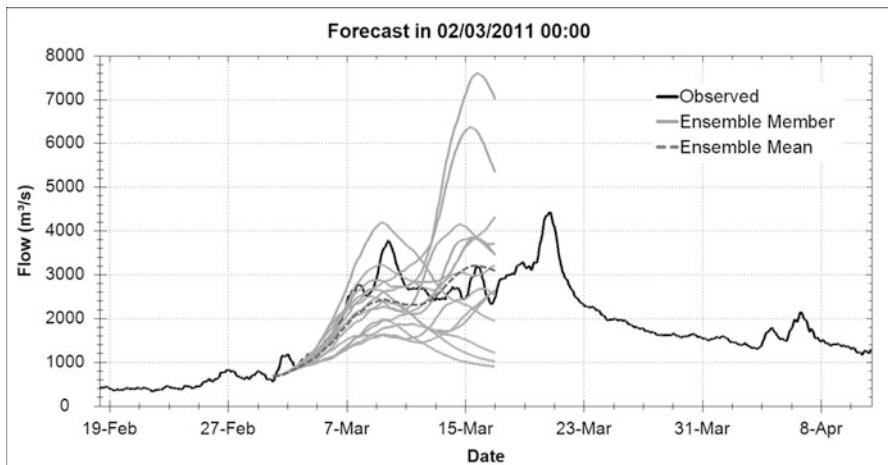


Fig. 5 Ensemble flow forecasting for the major flood of the rainy season 2010/2011, starting at 00 UTC 02 March 2011

occurred. Two members of the forecast indicate a very high peak inflow, over $6000 \text{ m}^3 \text{s}^{-1}$ on 16 March 2011, almost at the end of the forecast. Those two members of the ensemble did not capture the first peak in the hydrograph, just the second one (that did not occur). At the same time, most of the members that correctly captured the first peak indicated a decrease or just a smaller peak in the flow after it. Possibly this scenario happened because the numerical weather model was predicting a wet atmosphere in the upcoming days, with uncertainties about when the precipitation (that unloads the water budget in the atmosphere) would occur.

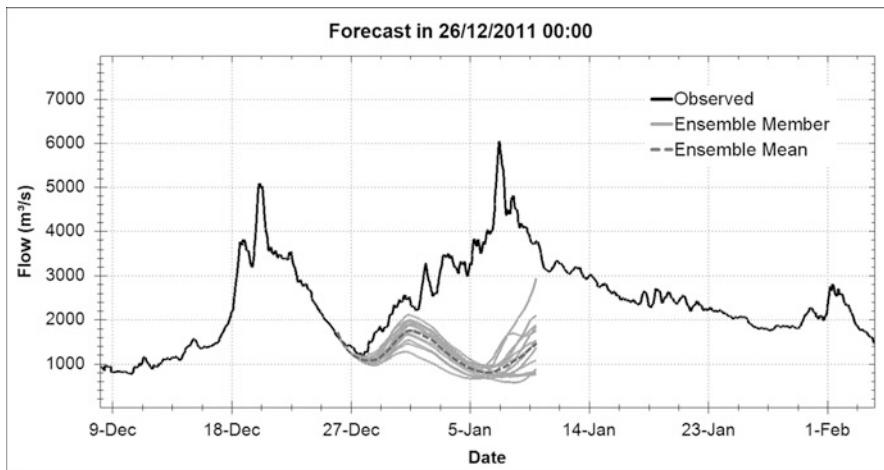


Fig. 6 Forecast of the major flood of the rainy season in 2011/2012, starting at 00 UTC 26 December 2011

Figures 6 and 7 show two forecasts issued for the most important inflow peak of the 2012 summer, during which maximum inflow was estimated to be $6000 \text{ m}^3 \text{s}^{-1}$, on 07 January. Figure 6 shows a forecast issued at 26 December 2011, during a recession of the inflow hydrograph, when estimated inflow was about $1400 \text{ m}^3 \text{s}^{-1}$. The spread of the ensemble is relatively low, with all the members predicting a continuing receding hydrograph for around 2 days, followed by an increase averaging just less than $2000 \text{ m}^3 \text{s}^{-1}$, and then decreasing again to $1000 \text{ m}^3 \text{s}^{-1}$ just after 05 January. Comparing this forecast to the observed inflow to the reservoir, it can be seen that the forecast was more or less correct up to the fourth or fifth day, with a small under-forecasting bias. After those days, the inflow hydrograph continued to rise, while all the members of the ensemble predicted a decrease. In terms of total volume inflow during the 2 weeks of forecast, it can be easily seen that there was an underestimation during the first week of the forecast and an underestimation higher than 100% during the second week of the forecast.

Also, the spread of ensemble members in this case was very low. This fact can be problematic for flood control and dam safety. For example, if one uses a threshold of $2500 \text{ m}^3/\text{s}$ to operate the reservoir, but all ensemble members indicate values below this limit, a wrong decision can be taken under assumption of a reliable range of the ensemble members.

Figure 7 shows a forecast issued just 3 days and 12 h after, on 29 December 2011 at 12 UTC. In this case, inflow increase during the first and second days was underestimated, but the ensemble spread started to grow suddenly, suggesting that a major peak could be expected. Four ensemble members overestimated the inflow peak, while ten underestimated it, and all members suggested a flow peak 2 days earlier than the real one. Assessing the forecast in terms of total volume during the following 2 weeks, it can be considered as relatively successful.

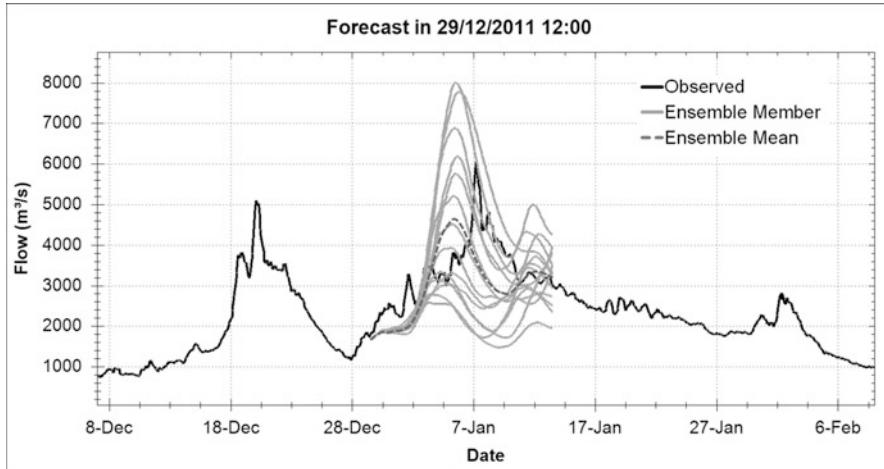


Fig. 7 Forecast of the major flood of the rainy season in 2011/2012, starting at 12 UTC 29 December 2011

The examples shown in Figs. 3, 4, 5, 6, and 7 shows that in general some small floods are not forecasted during the first 24 or 48 h. Even in the examples of what is considered as a successful forecast (Figs. 3, 4, 5, and 7), differences between forecasts and inflow observations are relatively large during the early lead times. Possibly, these errors result from different causes: imperfections of the hydrological model and in the data assimilation method. However, it is possible that the most important cause in the case of the Três Marias forecasts is the quality of the observed inflow hydrograph, as mentioned before, and the quality of observed precipitation data, due to the low gauge density.

Following the visual evaluation of the forecasts, results were analyzed using summarizing statistics and performance metrics. The ensemble mean was evaluated through the mean absolute error (MAE) and the correlation coefficient (R). The ensemble forecast quality and spread were evaluated using, respectively, the mean continuous ranked probability score (CRPS) and the rank histogram. These assessments were done first for all the available data and second considering only observed inflow above the upper 10% percentile ($1400 \text{ m}^3/\text{s}$).

Threshold exceedance verification was also done using the relative operating characteristic (ROC) curves, the Brier score (BS), and the forecast convergence score (FCS). The calibration of forecasts was evaluated by reliability diagrams. Ensemble mean was used as a reference deterministic forecast compared to the ensemble. In all cases the same limit of $1400 \text{ m}^3/\text{s}$ was used as a reference threshold.

All metric analyses presented here were executed using the Ensemble Verification System (EVS), a software package for ensemble verification developed by the NWS Office of Hydrologic Development (OHD), and presented by Brown et al. (2010).

The obtained results for the metric mean absolute error (MAE) and mean continuous ranked probability score (CRPS) are shown in Fig. 8. The CRPS is equal to the

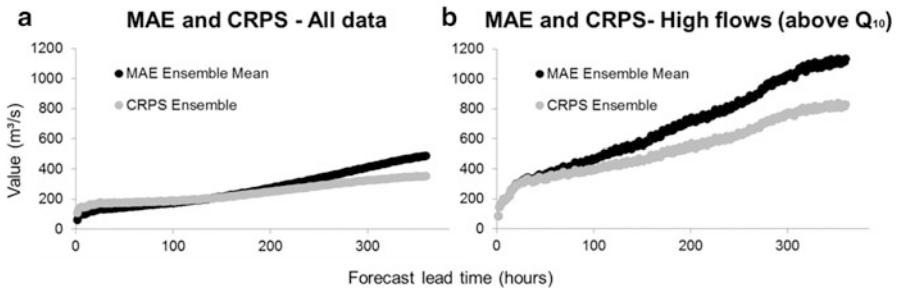


Fig. 8 MAE and CRPS assessment for Três Marias inflow: **(a)** using all data and **(b)** only for high flows, using a threshold of Q_{10} ($1400 \text{ m}^3/\text{s}$)

MAE, if used to compute the performance of a single deterministic forecast, and this allows the comparison of the quality from the ensemble forecast to the absolute error of the ensemble mean.

Results shown by Fig. 8 indicate that the skill of the forecasts decreases with increasing lead time, as is normally expected. In Fig. 8a (assessment for all data), the CRPS and the MAE are close to $200 \text{ m}^3/\text{s}^{-1}$ on early lead times and approximately $400 \text{ m}^3/\text{s}$ for longer lead times; these values are below the average inflow of Três Marias (ca. $600 \text{ m}^3/\text{s}$).

When considering only high flows, using the $1400 \text{ m}^3/\text{s}^{-1}$ threshold, both the CRPS and MAE statistics show larger values, as can be seen in Fig. 8b. CRPS, in this case, is close to $800 \text{ m}^3/\text{s}$ for the longer lead times, while MAE has values close to $1200 \text{ m}^3/\text{s}$. These values are relatively high in comparison to the average inflow and to the flow with 10% exceedance probability ($1400 \text{ m}^3/\text{s}$). However, for earlier lead times up to 10 days (240 h), the metrics indicate values of approximately $600 \text{ m}^3/\text{s}$, which are around the average inflow and lower than the threshold.

The comparison between MAE and CRPS shows that for lead times up to 72 h, the values are very similar. This happens because of very limited spread in the first time steps of the forecast. All members of the streamflow ensemble are similar, because the streamflow forecast primarily depends on observed rainfall. It is also noteworthy that for these early lead times, the errors are not zero. As discussed in the visual assessments, this is related to the hydrological model errors, but also to lack of observed precipitation and to uncertainties in the inflow mass balance back-calculated data.

Figure 8b also shows that for lead times above 72 h, CRPS values are lower than MAE values, indicating benefits in the use of ensemble forecasts, especially for high flows.

The correlation coefficient evaluation for the ensemble mean compared to the observations is presented in Fig. 9. This analysis is only focused on the ensemble mean and thus does not explore the benefits of using ensemble uncertainty. However, the correlation coefficient results are important to verify that major issues do not affect the results, for example, if the rainy period is not displaced or shifted.

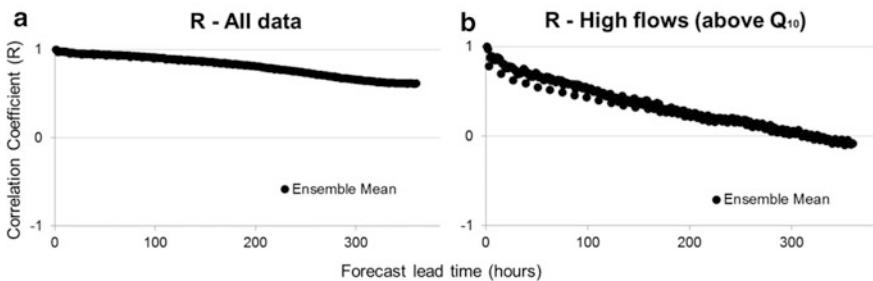


Fig. 9 Correlation coefficient between forecast and observed reservoir inflow versus lead time: (a) the complete inflow time series and (b) only for high flows ($Q > 1400 \text{ m}^3 \cdot \text{s}^{-1}$)

In this sense results obtained for the Três Marias inflow are suitable. The correlation coefficient using all data is always above 0.5, and the correlation coefficient for high flows (only values above Q_{10}) is above 0.5 for lead times for up to 4 days and above zero until 13 days of lead time.

Figure 10 shows the rank histograms for different lead times, from 48 to 360 h. Figure 10a shows the rank histograms for the whole data set, while Fig. 10b shows the rank histogram just for the cases when observed reservoir inflow exceeded $1400 \text{ m}^3 \text{s}^{-1}$.

The U-shaped form of the histograms suggests that in most of the cases, one of the following two outcomes happened: all the 14 members of the ensemble predicted a lower inflow than what was actually observed, or all the members predicted a higher inflow than was actually observed. This indicates a lack of spread of the ensemble at all lead times.

A rank histogram with a U shape during the early lead times can be considered normal in the case of ensemble forecasting systems that ignore the uncertainty in initial conditions of the hydrological model and of the observed data and only consider the meteorological uncertainty of weather forecasts. As a result, the ensemble shows a very low spread during the early time steps of the forecast, almost acting as a deterministic forecast.

It is possible that the U-shaped form of the rank histogram, even for long lead times shown in Fig. 10a, is probably related to the kind of forecasting errors that occur during the long dry season of the São Francisco River. During the dry season, forecast rainfall is normally close to zero in all members of an ensemble, and, therefore, all members result in the same inflow forecast. In this case observed inflow will usually be above (or below) all the members, because all the members are equal.

Figure 10b shows results where the influence of the dry season has been removed by preparing the rank histogram using only the cases when observed inflow was higher than $1400 \text{ m}^3 \text{s}^{-1}$. It can be seen that the form of the histogram did not change for short lead times. However, for lead times of about 96 h (4 days) to 360 h (15 days), the form of the rank histogram showed a tendency of more uniform distribution.

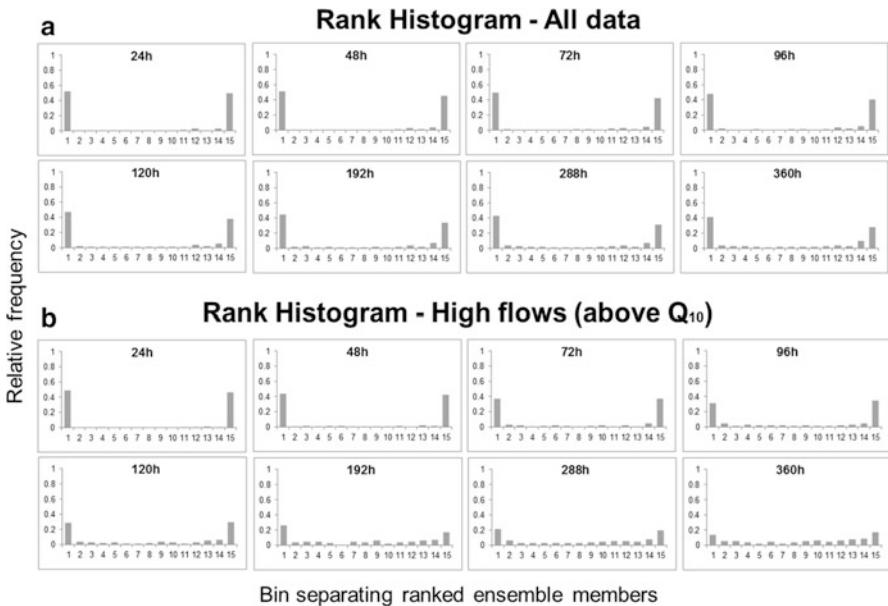


Fig. 10 Rank histogram assessment for Três Marias inflow: (a) using all data and (b) only for high flows, using a threshold of Q_{10} ($1400 \text{ m}^3/\text{s}$)

An assessment of $1400 \text{ m}^3/\text{s}$ threshold exceedance using ROC curves to the Três Marias inflow is presented in Fig. 11 for lead times of 48, 120, 240, and 336 h. The point representing the ensemble mean is also shown in Fig. 11.

Results for the first lead time (48 h) are very similar for all the ensemble percentiles and for the ensemble mean because at this lead time, the spread in the forecast is usually very small. At this lead time, the probability of detection (POD) of the threshold is around 0.9 for a probability of false detection (POFD) near to zero.

Results from the lead times 120 h (5 days) and 240 h (10 days) suggest that it is possible to take decisions related to the $1400 \text{ m}^3/\text{s}$ threshold exceedance with a higher true alarm rate by using the ensemble forecast higher percentiles than using the ensemble mean. With the upper bounds of the ensemble forecast, it is possible to have a true alarm rate of approximately 0.9 with a false alarm rate between 0.02 and 0.12. Using the ensemble mean, the true alarm rate would be around 0.8 for false alarm rates near 0.09.

In the case of the 360 h lead time ROC curve, there is a trade-off in the use of ensembles or ensemble means. With the higher ensemble percentiles, it is possible to have a higher probability of true alarm (between 0.85 and 0.9), but also a higher false alarm rate (0.18–0.2). And the use of the ensemble mean provides a smaller false alarm rate (0.13), although also a smaller true alarm rate (0.83) than ensemble higher bounds.

Figure 12 shows the results of Brier score (BS) for detection of $1400 \text{ m}^3/\text{s}$ floods in the Três Marias inflow. Lower BS values indicate better forecasting performance

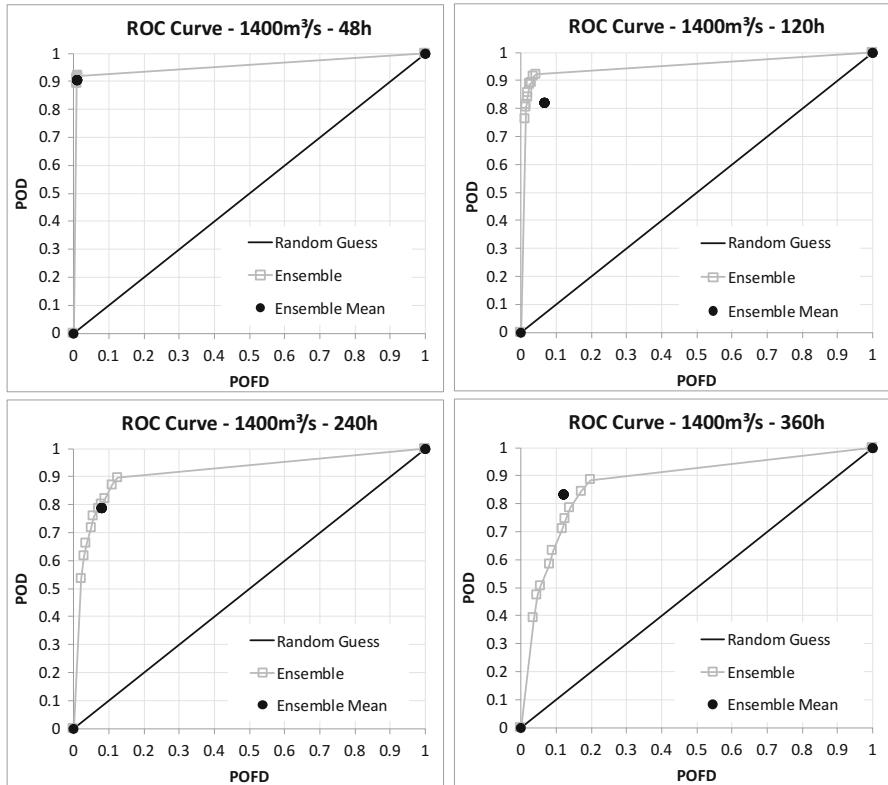


Fig. 11 ROC curves for four lead times for the Três Marias inflow using a threshold of $1400 \text{ m}^3/\text{s}$

in terms of threshold detection. The main information given by Fig. 12 is that BS is generally lower for the ensemble than for the ensemble mean, which means that threshold exceedance is usually better detected by the full ensemble forecast, if compared to the use of the ensemble mean as a consensus deterministic forecast. The errors of squared probabilities vary from 0.02 (lead time 48 h) to 0.10 (lead time 360 h) using the full ensemble, while using the ensemble mean errors vary from 0.02 (lead time 48 h) to 0.12 (lead time 360 h). And in the early lead times until 48 h, results are very similar between both forecasts due to the greater dependence of the system to observed conditions and lower uncertainty in the QPF.

Figure 13 shows the forecast convergence score (FCS) results for detection of $1400 \text{ m}^3/\text{s}$ floods in the Três Marias inflow. This metric describes the consistency of two sequential forecasts in terms of threshold detection (Pappenberger et al. 2011). Values near to zero indicate more consistent decisions. The FCS, as discussed by Pappenberger et al. (2011), is therefore not a performance metric, as it does not use comparisons with observations. But it is a descriptive metric that addresses a useful characteristic of forecasts, which is the capability of issuing consistent consecutive forecasts.

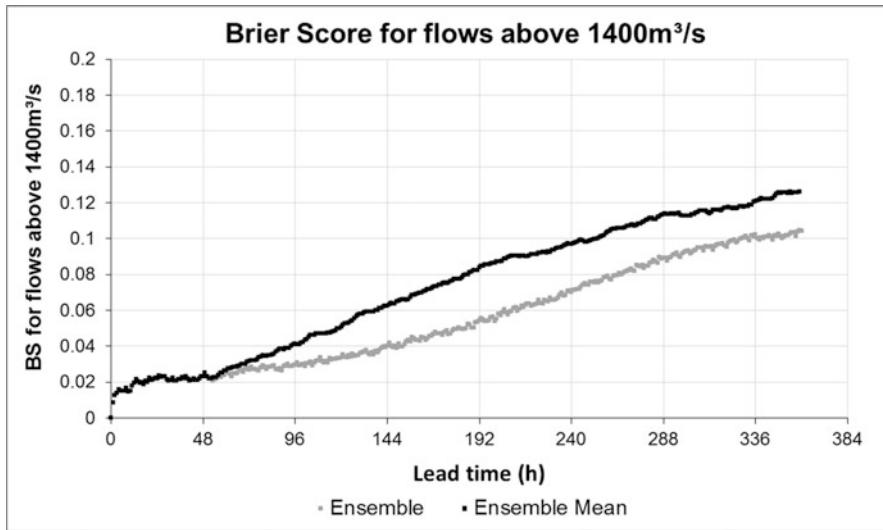


Fig. 12 Brier score analysis to the Três Marias inflow using a threshold of 1400 m³/s

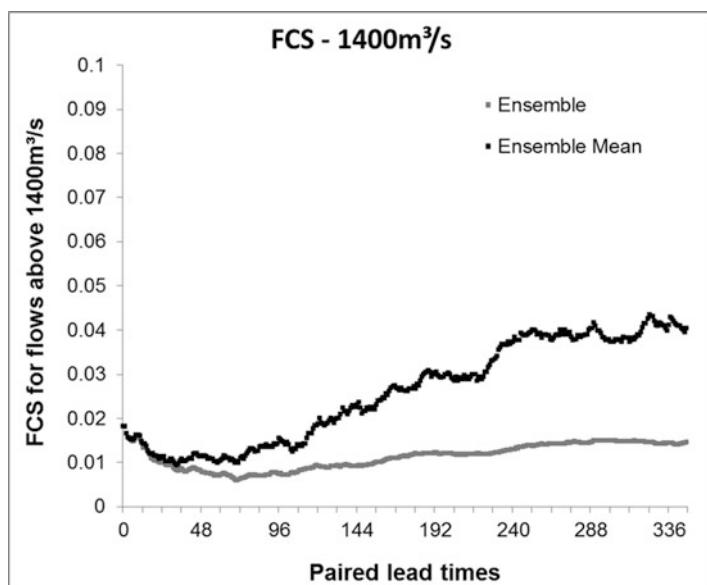


Fig. 13 Forecast convergence score analysis to the Três Marias inflow using a threshold of 1400 m³/s

This characteristic is especially important in reservoir short-term operations, when the reservoir is also addressed for flood control, which is the case of Três Marias reservoir. More consistent forecasts could lead to less inconstant decision-making concerning water releases or storage for preventing damages and maximizing power production.

Results obtained for the FCS indicate that for early lead times, between 1 to 2 days, results are very similar for the full ensemble and the mean. From this lead time on the FCS of the ensemble is always smaller than the FCS of the ensemble mean. This means that the full ensembles are more consistent in their indications than the deterministic forecasts.

It is important to mention that, in terms of BS and FCS, ensemble forecasts can be issued for intermediate probabilities of occurrences, while deterministic forecasts can only be issued for binary decisions (occurrence or not occurrence). Despite results already suggesting that full ensembles have better performance and are more consistent, there is also one additional value in such probability-based results.

Given the availability of probabilities in the future, a hydropower reservoir operation strategy could make use of this information for setting the decisions for the early operation lead times considering a scenario described by multiple probable futures. Putting it another way, it is possible to make a weighted decision that would allow a satisfactory operation considering all possible futures.

A deterministic forecast, on the other hand, generally leads to a single decision, and if, given all future uncertainties, something deviates from the predicted future, abrupt maneuvers of spillway gates may be necessary to offset the impact of the previously incorrect decisions.

Figure 14 shows the reliability diagram for detection of 1400 m³/s floods in the Três Marias inflow. The reliability diagram shows the conditional bias of the forecasts, comparing the forecasted probability ($P(F)$) in the x-axis with the conditional observed frequency ($P(O|F)$) in the y-axis. A sample count (sharpness) histogram for forecasted probabilities partitioned in five classes is also shown below in each diagram. In the sample count diagrams, the y-axis is plotted in a logarithmic scale, also done by Hamil et al. (2008), for better visualization of the number of samples in each class.

For the lead time of 48 h, the reliability diagram indicates a good calibration, although also a small number of samples were issued in the intermediate classes of probabilities. This may influence the results, but can be considered normal for early lead times, where the spread of the members is small.

For the lead time of 120 h, the forecasts also show a good calibration, with the issued line near to the perfect 45° line. For example, when forecasts were issued for the threshold with a probability of 0.7, the event was observed with a frequency near to 0.6. In terms of sample count, small probabilities had a greater number of occurrences than others. But all the classes had more than 30 samples.

For the two longer lead times, results show a behavior generally related to a positive conditional bias in the forecasts. This means that the forecasted probabilities

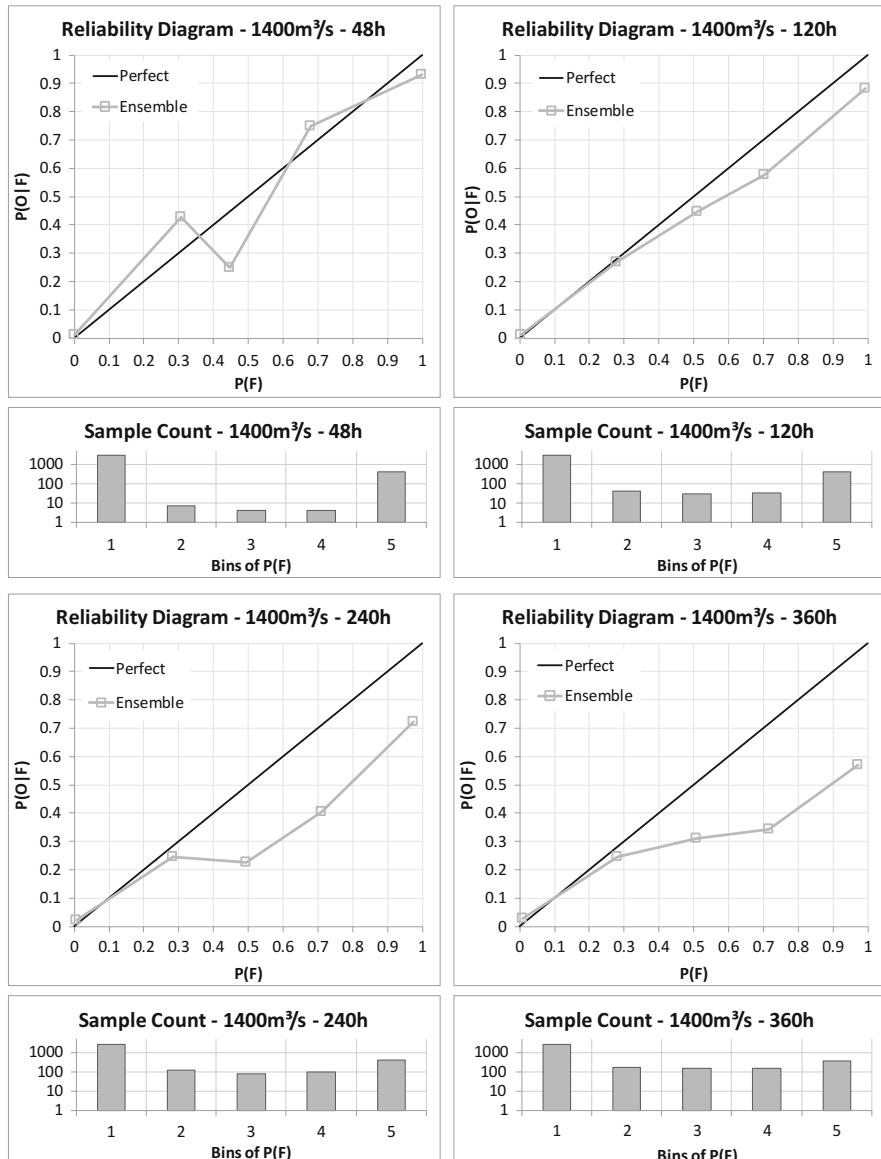


Fig. 14 Reliability diagram analysis to the Três Marias inflow using a threshold of $1400 \text{ m}^3/\text{s}$

were higher than the conditional observed frequencies. For example, in the 240 h lead time, when forecasts were issued with a probability of 0.7, the event was observed with a frequency near to 0.4. And in the 360 h lead time, when forecasts were issued with a probability of 0.7, the event was observed with a frequency near to 0.35. This indicates that for these longer lead times, the system overestimates the

probability of event occurrence, requiring caution when the spread of the members is used as an uncertainty measurement. In these cases, the sharpness histograms always indicated more than hundred samples for all classes.

Overall, metric assessment results indicate that the evaluated hydrological forecasting system has results similar to observed reality. However, increasing uncertainty is expected with lead time, and the results with lead times longer than 10 days may have a quality that is not desirable for the system. However, shorter lead times' results are useful from a reservoir operation point of view, since errors are in an acceptable range and the true alarm–false alarm trade-off given by the ROC curves has acceptable rates.

In terms of comparison between deterministic forecasts given by the ensemble mean as a consensus, the ensemble forecasts indicate benefits in terms of Brier score and in terms of consistency given by the forecast convergence score (FCS). And in terms of reliability, the system presents a good calibration until 5 days of lead time. Beyond this lead time, more caution should be taken in the use of probabilities for threshold exceedance.

More spread on early lead times of the forecasting system (measured by the rank histograms) could be obtained, if the system also made use of a hydrological ensemble, not only a meteorological one. However, it can be said that the most important problems identified on early lead times are the ones related to lack of observed information and to low quality of inflow observations obtained by reservoir water budget. Work to minimize the importance of these errors can be related to including more data sources in the forecasting system, such as radar or satellite rainfall information, and also trying other methods of data assimilation that considers uncertainties in the input data.

Finally, it is worth noting that the hydrological model calibration period and forecasting test period partially overlap. This potentially leads to relatively higher-quality results that could be obtained in operational applications.

4 Conclusion

This case study presents an assessment of a short-term ensemble forecasting system to predict inflows to a major Brazilian hydropower reservoir. Hindcasting experiments were conducted for the period from 2008 to 2012 using data from a global numerical weather prediction model of the Brazilian meteorological center CPTEC, in combination with a large-scale hydrological model largely used in Brazil for hydrological forecasting.

Results for the São Francisco River are encouraging in terms of ensemble applicability. It is believed that ensemble inflow forecasts to major reservoirs in Brazil will be used in the near future as input to the chain of models used for the optimization of the national electric power producing system, due to results shown here and in other references (Fan et al. 2014, 2015a, b; Schwanenberg et al. 2015). The forecasting system framework described here is currently being tested in other

river basins in Brazil, including the Tocantins River, located in the Northern region of Brazil, and the Uruguay River, located in the South.

Results also show that forecasts can be improved by adopting a scheme to consider the uncertainty in initial conditions of the hydrological model and in the observed data. Another improvement could be obtained by increasing the number of gauging stations with hourly data and telemetric transmission. One meteorological radar that was recently installed in the Southeast of the basin will probably lead to improvements in short lead time forecasts, and the use of satellite information can be interesting for improving the observed rainfall generalization.

A benefit from ensemble forecasts that was discussed in this text is the ability to issue for intermediate probabilities of occurrences, while deterministic forecasts can only issue for binary decisions (occurrence or nonoccurrence). And, given the availability of these probabilities in the future, a hydropower reservoir operation strategy could make use of this information for setting the decisions for the early operation lead times considering a scenario described by multiple probable futures. Putting it another way, using ensemble forecasts it should be possible to make a weighted decision that would allow a satisfactory operation considering all possible future uncertainties.

This kind of assessment for short-term reservoir operation, where the ensemble forecast probabilistic scenarios are used, can be conducted by multistage stochastic optimization approaches. Such a method was tested by Schwanenberg et al. (2015) for the Três Marias reservoir seeking the objectives of flood control and energy generation in the São Francisco River, with encouraging results.

References

- P. Bougeault, Z. Toth, C. Bishop, B. Brown, D. Burridge, D.H. Chen, B. Ebert, M. Fuentes, T.M. Hamill, K. Mylne, J. Nicolau, T. Paccagnella, Y. Park, D. Parsons, B. Raoult, D. Schuster, P.S. Dias, R. Swinbank, Y. Takeuchi, W. Tennant, The THORPEX interactive grand global ensemble. *Bull. Am. Meteorol. Soc.* **91**(8), 1059–1072 (2010). 14p
- A.A. Bradley, S.S. Schwartz, Summary verification measures and their interpretation for ensemble forecasts. *Mon. Weather Rev.* **139**(9), 3075–3089 (2011)
- J.M. Bravo, A.R. Paz, W. Collischonn, C.B. Uvo, O.C. Pedrollo, S. Chou, Incorporating forecasts of rainfall in two hydrologic models used for medium-range streamflow forecasting. *J. Hydrol. Eng.* **14**, 435–445 (2009)
- J.D. Brown, J. Demargne, D.-J. Seo, Y. Liu, The ensemble verification system (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations. *Environ. Model. Softw.* **25**(7), 854–872 (2010)
- L. Calvetti, Hydrometeorologic ensemble forecasts in the Upper Iguassu river basin using WRF and TopModel. Ph.D. thesis (in portuguese). Instituto de Astronomia, Geofísica e Ciências Atmosféricas. Universidade Federal de São Paulo, 2011. 141p
- M. Cataldi, C.O. Machado, L.G.F. Guilhon, S.C. Chou, J.L. Gomes, J.F. Bustamante, Análise de Previsões de Precipitação obtidas com a utilização do modelo ETA como insumo para modelos de previsão semanal de vazão natural. *Rev. Bras. Recursos Hídricos.* **12**, 5–12 (2007)
- S.C. Chou, C.A. Tanajura, Y. Xue, C.A. Nobre, Validation of the coupled ETA/SSiB model over South America. *J. Geophys. Res.* **107**(D20), 8088 (2002)

- W. Collischonn, C.E.M. Tucci, R. Haas, I. Andreoli, Forecasting river Uruguay flow using rainfall forecasts from a regional weather-prediction model. *J. Hydrol.* **305**, 87–98 (2005)
- W. Collischonn, D.G. Allasia, B.C. Silva, C.E.M. Tucci, The MGB-IPH model for large scale rainfall-runoff modeling. *Hydrol. Sci. J.* **52**, 878–895 (2007a)
- W. Collischonn, C.E.M. Tucci, R.T. Clarke, S.C. Chou, L.G. Guilhon, M. Cataldi, D.G. Allasia, Medium-range reservoir inflow predictions based on quantitative precipitation forecasts. *J. Hydrol.* **344**, 112–122 (2007b)
- W. Collischonn, A. Meller, P.L. Silva Dias, D.S. Moreira, Hindcast experiments of ensemble streamflow forecasting for the Paraopeba river. *Geophys. Res. Abstr.* **14**, 3596 (2012)
- W. Collischonn, A. Meller, F. Fan, D.S. Moreira, P.L. Silva Dias, D. Buarque, J.M. Bravo, Short-term ensemble flood forecasting experiments in Brazil. *Geophys. Res. Abstr.* **15**, 11910 (2013)
- F.S. Costa, I.P. Raupp, J.M. Damazio, P.D. Oliveira, L.G.F. Guilhon, The methodologies for the flood control planning using hydropower reservoirs in Brazil, in *6th International Conference on Flood Management, IC MF6*, São Paulo, 2014 (ABRH, São Paulo, 2014)
- EPE – EMPRESA DE PESQUISA ENERGÉTICA, Anuário estatístico de energia elétrica 2013. Rio de Janeiro: EPE, 2013. 253 p
- B.A. Faber, J.R. Stedinger, Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts. *J. Hydrol.* **249**, 113–133 (2001)
- F. M. Fan, V. A. Siqueira, P. R. M. Pontes, W. Collischonn, Avaliação do Uso de Dados de Precipitação do Merge para Simulação Hidrológica da Bacia do Rio Paraná. In: XVII Congresso Brasileiro de Meteorologia, 2012, Gramado – RS. AVALIAÇÃO DO USO DE DADOS DE PRECIPITAÇÃO DO MERGE PARA SIMULAÇÃO HIDROLÓGICA DA BACIA DO RIO PARANÁ, 2012a
- F. M. Fan, P. R. M. Pontes, W. Collischonn, L. F. S. Beltrame, Sistema de Previsão de Vazões para as Bacias dos Rios Taquari-Antas e Pelotas, in XI Simpósio de Recursos Hídricos do Nordeste, 2012, João Pessoa PB. Sistema de Previsão de Vazões para as Bacias dos Rios Taquari-Antas e Pelotas, 2012b
- F.M. Fan, W. Collischonn, A. Meller, L.C.M. Botelho, Ensemble streamflow forecasting experiments in a tropical basin: the São Francisco river case study. *J. Hydrol.* **519**, 2906–2919 (2014)
- F. M. Fan, W. Collischonn, K. Quiroz, M. V. Sorribas, D. C. Buarque, V. A. Siqueira, Flood forecasting on the Tocantins River using ensemble rainfall forecasts and real-time satellite rainfall estimates. *J. Flood Risk Manag. v. SI* (2015a). doi:10.1111/jfr3.12177
- F. M. Fan, D. Schwanenberg, W. Collischonn, A. Weerts, Verification of inflow into hydropower reservoirs using ensemble forecasts of the TIGGE database for large scale basins in Brazil. *J. Hydrol.: Reg. Stud.* (2015b) (in press). doi:10.1016/j.ejrh.2015.05.012
- L.G.F. Guilhon, V.F. Rocha, J.C. Moreira, Comparison of forecasting methods for natural inflows in hydropower developments. *Braz. J. Water Res.* **12**(3), 13–20 (2007) (in Portuguese)
- T. Hamill, R. Hagedorn, J. Whitaker, Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. part II: precipitation. *Mon. Weather Rev.* **136**(7), 2620–2632 (2008)
- A.F. Hamlet, D. Huppert, D.P. Lettenmaier, Economic value of longlead streamflow forecasts for Columbia river hydropower. *J. Water Resour. Plann. Manage.* **128**(2), 91–101 (2002)
- H. Hersbach, Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast.* **15**, 559–570 (2000)
- I.T. Jolliffe, D.B. Stephenson (eds.), *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd edn. (Wiley, Chichester, 2012)
- M.E.P. Maceira, J.M. Damázio, Periodic auto-regressive streamflow models applied to operation planning for the Brazilian hydroelectric system, in *Regional Hydrological Impacts of Climatic Change – Impact Assessment and Decision Making*. IAHS Publ, vol. 295 (IAHS Press, Wallingford, 2005)
- M.E.P. Maceira, L.A. Terry, F.S. Costa, J.M. DAMÁZIO, A.C.G. e MELO, Chain of optimization models for setting the energy dispatch and spot price in the Brazilian system, in *Anais do XIV Power Systems Computation Conference*, session 43, paper 1, Sevilla, 2002

- E.P. Maurer, *Predictability of Runoff in the Mississippi River Basin*. Water Resour. Ser. Tech. Rep. vol. 172 (University of Washington, Seattle, 2002)
- A. Meller, Short-range ensemble flood forecasting. Ph.D thesis. (In Portuguese) Instituto de Pesquisas Hidráulicas. Universidade Federal do Rio Grande do Sul, 2013. 224p
- A. Meller, J.M. Bravo, W. Collischonn, Assimilação de Dados de Vazão na Previsão de Cheias em Tempo-Real com o Modelo Hidrológico MGB-IPH. Rev. Bras. Recur. Hidr. **17**, 209–224 (2012)
- A.M. Mendonça, J.P. Bonatti, Experiments with EOF-based perturbation methods and their impact on the CPTEC/INPE ensemble prediction system. Mon. Weather Rev. **137**, 1438–1459 (2009)
- R.C.D. Paiva, D.C. Buarque, W. Collischonn, M.-P. Bonnet, F. Frappart, S. Calmant, C.A. Bulhões Mendes, Large-scale hydrologic and hydrodynamic modeling of the Amazon River basin. Water Resour. Res. **49**, 1226–1243 (2013)
- F. Pappenberger, K. Bogner, F. Wetterhall, Y. He, H.L. Cloke, J. Thielen, Forecast convergence score: a forecaster's approach to analysing hydro-meteorological forecast systems. Adv. Geosci. **29**, 27–32 (2011). doi:10.5194/adgeo-29-27-2011
- A. R. Paz, W. Collischonn, C. E. M. Tucci, R. T. Clarke, D. G. Allasia, Data assimilation in a large-scale distributed hydrological model for medium-range flow forecasts, in Symposium of the IUGG, 2007, Perugia. Quantification and Reduction of predictive uncertainty for sustainable water resources management (Proceedings of Symposium HS2004 at IUGG2007), (IAHS Press, Wallingford, 2007), p. 471–478
- M. Pereira, L. Pinto, Multi-stage stochastic optimization applied to energy planning. Math. Program. **52**(1–3), 359–375 (1991)
- L. Raso, N. Van De Giesen, P. Stive, D. Schwanenberg, P.J. Van Overloop, Tree structure generation from ensemble forecasts for real time control. Hydrol. Process. **27**, 75–82 (2013)
- D. Schwanenberg, F.M. Fan, S. Naumann, J.I. Kuwajima, R.A. Montero, A. Assis Dos Reis, Short-term reservoir optimization for flood mitigation under meteorological and hydrological forecast uncertainty. Water Resour. Manag. **920–4741**, 1–17 (2015). doi:10.1007/s11269-014-0899-1
- H. Stanski, L. Wilson, W. Burrows, *Survey of Common Verification Methods in Meteorology* (World Meteorological Organization, Geneva, 1989)
- C.E.M. Tucci, P.L.S. Dias, R.T. Clarke, G.O. Sampaio, W. Collischonn, Long-term flow forecasts based on climate and hydrologic modeling: Uruguay river basin. Water Resour. Res. **39**(7), 1–2 (2003)
- C.E.M. Tucci, W. Collischonn, R.T. Clarke, A.R. Paz, D. Allasia, Short- and long-term flow forecasting in the Rio Grande watershed (Brazil). Atmos. Sci. Lett. **9**, 53–56 (2008)
- D. Wilks, *Statistical Methods in the Atmospheric Sciences* (Academic, Amsterdam/Boston, 2006)
- W.W.-G. Yeh, L. Becker, R. Zettlemoyer, Worth of inflow forecast for reservoir operation. J. Water Resour. Plann. Manage. **108**(WR3), 257–269 (1982)



New York City's Operations Support Tool: Utilizing Hydrologic Forecasts for Water Supply Management

James Porter, Gerald Day, John C. Schaake, and Lucien Wang

Contents

1	Introduction	1330
1.1	History of the NYC Water Supply System	1331
1.2	Key Operational Objectives	1340
1.3	Operations Support Tool Overview	1343
2	Ensemble Forecasts for NYC OST	1347
2.1	Forecast Types	1347
2.2	Ensemble Postprocessor (EPP)	1351
3	Applying Ensemble Forecasts in Operations	1353
3.1	Transition to Probabilistic Analysis and Decision Making	1353
3.2	Spill Mitigation	1355
3.3	Water Quality	1359
3.4	Conservation Releases	1362
4	Conclusion	1369
	References	1369

J. Porter (✉)

New York City Department of Environmental Protection, Bureau of Water Supply,
New York, NY, USA

e-mail: jporter@dep.nyc.gov

G. Day

RTI International, Ft. Collins, CO, USA

e-mail: gday@rti.org

J. C. Schaake

U.S. National Weather Service (retired), Annapolis, MD, USA

e-mail: jcschaake@comcast.net

L. Wang

Hazen and Sawyer, San Francisco, CA, USA

e-mail: lwang@hazenandsawyer.com

Abstract

The New York City Department of Environmental Protection (DEP) supplies over one billion gallons per day (BGD) of water to more than nine million people in the New York City metropolitan area, making it one of the largest suppliers of surface water in the United States. DEP's water supply system is as complex as it is large; it draws water from three distinct watersheds and features a number of interconnections and redundancies allowing for a large number of potential operating conditions. The system has a wide range of objectives – from supplying clean, reliable water for municipal demand to meeting environmental flow requirements for downstream stakeholders. Combined with the existing system complexity, these disparate objectives can make operational decision making a challenge.

In 2013, DEP launched the Operations Support Tool (OST) – a state-of-the-art model built to assist the utility in water supply operation decisions. OST consists of a system model (OASIS) to simulate water supply operation decisions and a linked hydrodynamic two-dimensional water quality model (CE-QUAL-W2). The model is initialized using current system conditions (e.g., reservoir elevations, water quality conditions) and is driven forward in time using ensemble hydrologic forecasts. This setup gives DEP the ability to simulate a wide variety of operational strategies in near real-time, allowing for objective alternative analysis prior to making operational decisions. Ensemble hydrologic forecasts are a critical part of the success of this approach, as they enable DEP to evaluate decisions probabilistically by explicitly considering hydrologic uncertainty.

This chapter provides an overview of the New York City water supply system, details the hydrologic forecasts used in OST, and reviews a handful of real operational applications of OST and the hydrologic forecast system.

Keywords

Water resource management · Applications of hydrologic ensemble forecasts · Reservoir operations · Decision support · Water quality management · Turbidity · Probabilistic risk · System modeling · Supply reliability · Conservation releases · Real-time modeling · Data visualization · New York City water supply

1 Introduction

The New York City Department of Environmental Protection (DEP) operates the New York City water supply system, which delivers over one billion gallons per day (BGD) to more than nine million people. DEP's water supply system is as complex as it is large; it draws water from three distinct watersheds and features a number of interconnections and redundancies allowing for a large number of potential operating conditions. The system has a wide range of objectives – from supplying clean, reliable water for municipal demand to meeting environmental flow requirements for downstream stakeholders. Combined with the existing system complexity, these disparate objectives can make operational decision making a challenge.

In 2013, DEP launched the Operations Support Tool (OST) – a state-of-the-art model built to assist the utility in water supply operation decisions. OST consists of a system model (OASIS) to simulate water supply operation decisions and a linked hydrodynamic two-dimensional water quality model (CE-QUAL-W2). The model is initialized using current system conditions (e.g., reservoir elevations, water quality conditions) and is driven forward in time using ensemble hydrologic forecasts. This setup gives DEP the ability to simulate a wide variety of operational strategies in near real-time, allowing for objective alternatives analysis prior to making operational decisions. Ensemble hydrologic forecasts are a critical part of the success of this approach, as they enable DEP to evaluate decisions probabilistically by explicitly considering hydrologic uncertainty.

This chapter begins with a history of the NYC Water Supply System, an overview of the system's operational objectives, and an introduction to the City's OST. The second section describes the different types of ensemble forecasts that are available as options within the OST, discusses the expected forecast skill in these statistical and physically based options, and presents an ensemble postprocessor that is used with the physically based ensemble forecasts to insure the forecasts are unbiased and reliable. The final section discusses the consideration of risk in the decision-making process and presents several examples of how probabilistic forecasts are used for specific decisions.

1.1 History of the NYC Water Supply System

Through the early 1800s, municipal water was supplied in New York City via a 48-acre pond in Lower Manhattan and a system of various wells. As the population increased, the supply became inadequate in meeting demand, and the water became increasingly polluted due to municipal and industrial discharges. In 1830, the City began construction of the Croton Water Supply System in Westchester County. By 1842, construction was completed on the Old Croton Reservoir and its corresponding Aqueduct, which delivered water to two distribution reservoirs in Manhattan. By the late 1800s, capacity issues necessitated development of additional reservoirs in the system, development of a second aqueduct (New Croton Aqueduct), and expansion of the Old Croton Reservoir (New York City Department of Environmental Protection 2016).

In the early 1900s, the City made plans to further expand the water supply system into the Catskill Mountain Region. By 1915, the City completed development of the Catskill Aqueduct and Ashokan Reservoir which impounds Esopus Creek. By 1928, additional storage was added with the completion of Schoharie Reservoir and the Shandaken Tunnel, which allows water to be diverted from Schoharie to Ashokan (New York City Department of Environmental Protection 2016).

In 1928, the City announced plans to develop new reservoirs near the headwaters on the Delaware River. Shortly thereafter, the State of New Jersey brought an action to prevent New York State and City from using any water in the Delaware Basin. In 1931, the US Supreme Court granted approval for the City to divert up to 440 MGD

from two headwater tributaries to the Delaware River (Office of the Delaware River Master 2015). Under the initial Supreme Court decision, the City completed development of the Delaware Aqueduct (1944), Rondout Reservoir (1950), Neversink Reservoir (1954), and Pepacton Reservoir (1955). In 1952, the City petitioned the Supreme Court to increase its allocation to the Delaware Basin. In 1954, the Supreme Court issued a decree – consented to by the states bordering the Delaware River (New York, New Jersey, Pennsylvania, and Delaware) and New York City – increasing the City's allocation to 800 MGD. The amended decree also required the City to release water from the Delaware Reservoirs to maintain a 1750 CFS flow objective at Montague, NJ (Office of the Delaware River Master 2015). Following the amended decree, the City finished development of Cannonsville Reservoir in 1964, marking the completion of the City's upstate reservoir system (New York City Department of Environmental Protection 2016).

The City's current supply system, managed by DEP, consists of three water source systems (Catskill, Croton, and Delaware) with 19 reservoirs and 3 controlled lakes and encompasses over 2000 mi² of watershed area. The system has approximately 568 BG of total storage capacity and supplies over 1 BGD in daily average demand to over nine million customers. The majority of the system (Catskill and Delaware source systems) is unfiltered given a Filtration Avoidance Determination (FAD) issued by the US Environmental Protection Agency (EPA). The Catskill and Delaware Systems are treated by an ultraviolet (UV) disinfection system and chlorine. The Croton System has been filtered since the completion of the Croton Water Filtration Plant (WFP) in 2015. The following sections provide more detailed information regarding the individual systems as well as interconnections offering operational flexibility.

1.1.1 Croton System

The Croton System is the oldest and smallest system in the City's water supply. The System is located east of the Hudson River and consists of 12 reservoirs and 3 controlled lakes totaling about 89 BG in storage. A map of the System is presented in Fig. 1.

Reservoirs in the Croton System are arranged largely in series. Upstream reservoirs, such as Middle Branch and East Branch, release (or spill) water to their respective tributaries, which then drain to the next downstream reservoir. These operations continue until New Croton Reservoir, which is the terminal reservoir in the system. At New Croton, operations staff may either divert water to the New Croton Aqueduct or make releases to the Croton River. Historically the Croton System supplies about 10% of the City's daily demand, though it can supply close to 30% during times of drought or water quality concern.

1.1.2 Catskill System

The Catskill System includes Schoharie Reservoir, the Shandaken Tunnel, Ashokan Reservoir, and the Catskill Aqueduct. Combined, Schoharie and Ashokan have about 132.5 BG in capacity and historically have met 30–40% of the City's annual average demand. Additionally, the system supplies a number of smaller communities

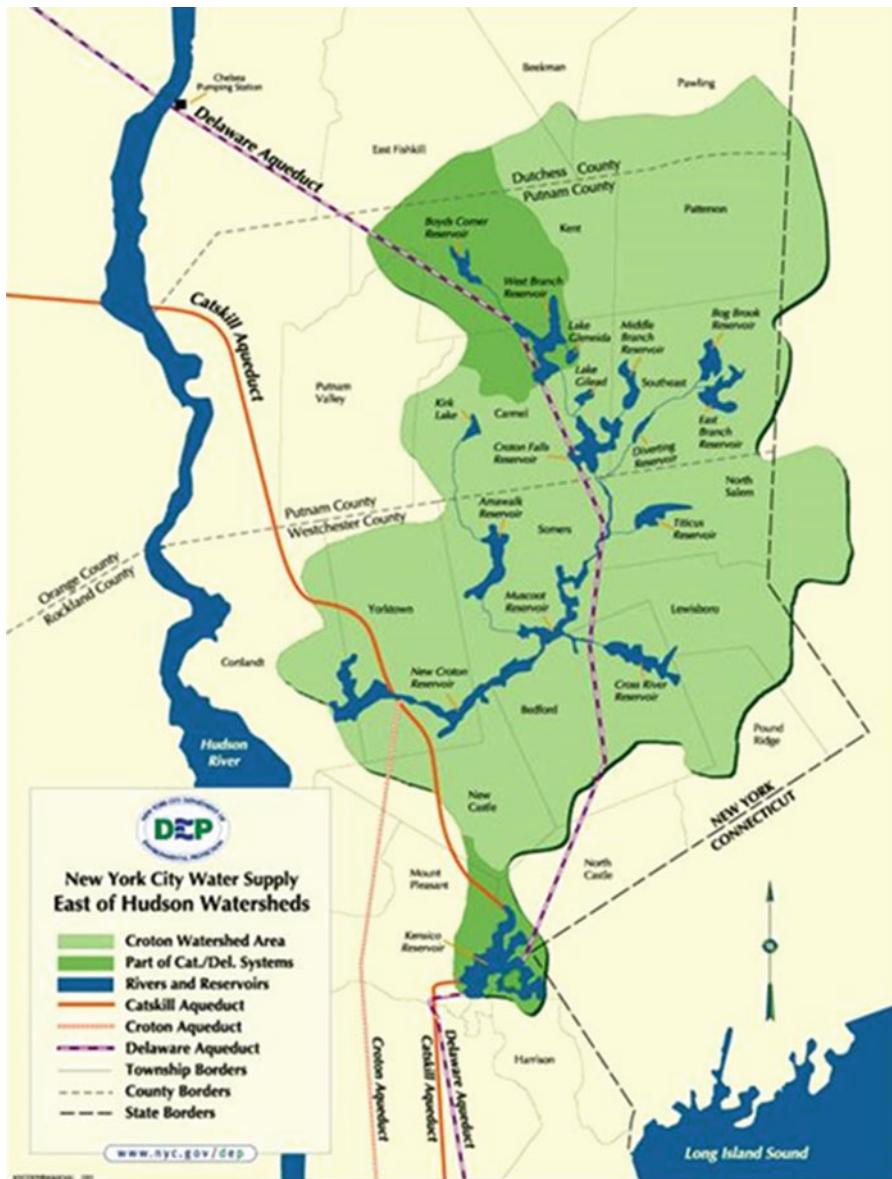


Fig. 1 Map of the Croton System

along the Catskill Aqueduct in Ulster, Orange, Putnam, and Westchester Counties. Water from the Catskill System is unfiltered, but is disinfected with chlorine, followed by secondary disinfection by ultraviolet light at the Catskill-Delaware Ultraviolet Disinfection Facility (Cat-Del UV) in Westchester County. Water from the Catskill System is generally of high quality; however the watershed undergoes



Fig. 2 Map of Catskill System

intermittent turbidity events that threaten water quality following large storms (refer to Sect. 1.2.2 for more information). A map of the Catskill System is presented in Fig. 2.

Schoharie Reservoir impounds approximately 315 mi^2 area of Schoharie Creek and lies in portions of Schoharie, Delaware, and Greene Counties. Schoharie Reservoir is connected to Ashokan Reservoir via the Shandaken Tunnel which travels over 18 miles southeast across Greene County and discharges to Esopus Creek before eventually draining to Ashokan's West Basin (Weiss et al. 2013). Diversions from the Shandaken Tunnel are regulated for turbidity and temperature by a State Pollutant Discharge Elimination System (SPDES) permit, which can



Fig. 3 Ashokan Reservoir schematic

constrain operations depending on background water quality conditions in Esopus Creek and Schoharie Reservoir (NYSDEC 2008; New York Codes, Rules, and Regulations 1991). The Shandaken Tunnel is also required to supplement flow in Esopus Creek for the purpose of maintaining a minimum combined flow at the confluence of the Tunnel and the Creek. Esopus Creek conveys the combined native Esopus Creek flow and Shandaken Tunnel discharge roughly 12 miles southeast into Ashokan Reservoir.

Ashokan Reservoir (Fig. 3) impounds Esopus Creek with Olive Bridge Dam and a series of six man-made dikes. It is separated into two basins (West and East) by the Dividing Weir Dike. Flow moves into the West Basin via Esopus Creek, over the dividing weir, and eventually into the East Basin. Water can also be moved between the Basins using gates located within the Dividing Weir Dike. Inflow to the East Basin is relatively small, as its drainage area is 18 mi² (compared to 239 mi² in the West Basin).

Water can be drawn from both the West and East Basins into the Catskill Aqueduct, which delivers water to Kensico Reservoir and New York City, and the Ashokan Release Channel (ARC) which discharges water to Lower Esopus Creek. Under the Ashokan Interim Release Protocol (IRP), DEP is required to release up to 15 MGD from the ARC for community and conservation purposes (New York State Department of Environmental Conservation/New York City Department of Environmental Protection (DEC/DEP), 2011). DEP is also required to make spill mitigation releases corresponding with a Conditional Seasonal Storage Objective (CSSO). The CSSO, shown in Fig. 4, shows the reservoir guide curve relating usable storage in Ashokan to the time of year. DEP is expected to make releases from the

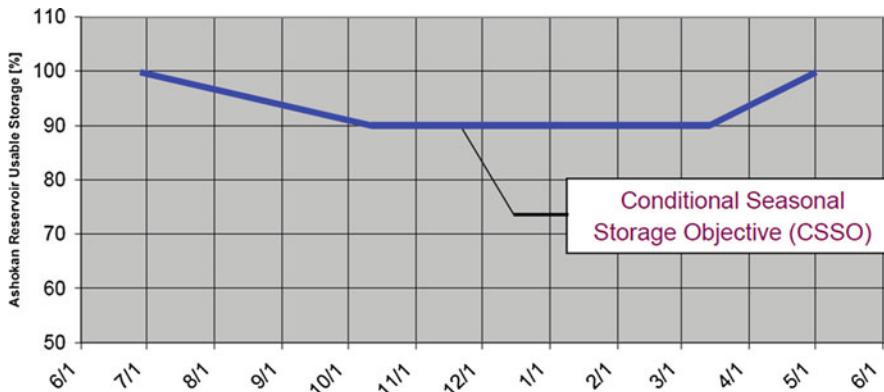


Fig. 4 Ashokan CSSO guide curve

ARC that prevent or mitigate storage exceeding the CSSO based on forecasted inflow and operations (see Sect. 3.2 for more information).

1.1.3 Delaware System

DEP's Delaware Water Supply System is comprised of four reservoirs – Pepacton, Cannonsville, Neversink, and Rondout – and historically has supplied about 50–60% of in-City water demand. Like the Catskill System, the Delaware is unfiltered but disinfected with chlorine and ultraviolet light. Historically, water from the Delaware is the best quality in the NYC System. Though the Delaware System lacks the turbidity challenges present in the Catskill System, its operations affect a much larger number of downstream stakeholders, including the States of New Jersey, Pennsylvania, and Delaware. As a result, reservoir release policy for Cannonsville, Pepacton, and Neversink is significantly more complex (see Sect. 1.2.3 for more information). Figure 5 displays a map of the Delaware System.

Cannonsville, Pepacton, and Neversink dam separate tributaries of the Upper Delaware River and divert water to Rondout Reservoir via separate tunnels. Rondout Reservoir impounds Rondout Creek near Grahamsville, NY, and is the terminal reservoir in the Delaware System. Rondout diverts water to West Branch and Kensico Reservoirs via the Delaware Aqueduct. DEP is restricted to diverting no more than 800 MGD on an annual rolling average basis by Supreme Court Decree.

1.1.4 Terminal Reservoirs, Treatment, and Distribution

New York City has two terminal reservoirs prior to treatment and distribution: New Croton Reservoir and Kensico Reservoir.

The 1989 Surface Water Treatment Rule required water from the Croton System to undergo filtration. Between 1993 and 2015, the System was taken offline during the siting, design, and construction of the plant. Completed in May 2015, Croton WFP is an underground stacked dissolved air flotation (DAF)/filtration process combined with UV and chlorine disinfection. It is designed to treat up to

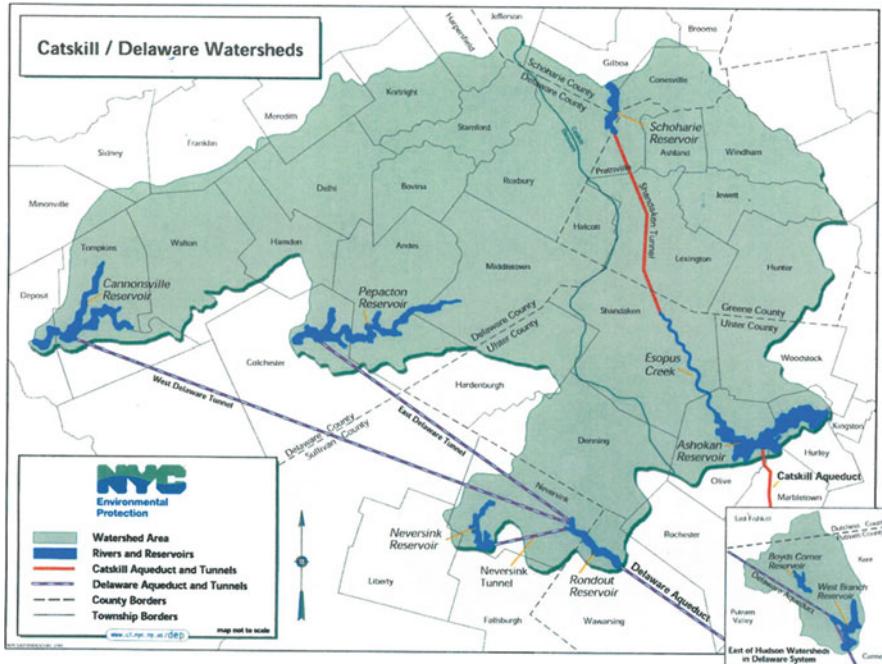


Fig. 5 Map of the Delaware System

290 MGD, making it one of the largest DAF facilities in the world. Water from New Croton Reservoir may be diverted directly to the filtration plant, or it can be diverted to Jerome Park Reservoir, a small reservoir in the Bronx that acts as a surge basin for meeting diurnal demand.

Water from the Catskill and Delaware Systems is mixed in Kensico Reservoir, located near Valhalla in Westchester County. Figure 6 shows Kensico Reservoir and its associated infrastructure.

Water from the Catskill System enters Kensico directly from the Catskill Aqueduct near the northwest end of the Reservoir. Prior to entering the Reservoir, a coagulant – aluminum sulfate (alum) – may be added to the Catskill Aqueduct at the Catskill Alum Plant. Alum is only added to Catskill water during extreme turbidity events.

Water from the Delaware System enters Kensico via the Delaware Aqueduct near the northeast side of the Reservoir. Prior to entering Kensico, water from the Delaware Aqueduct can pass through West Branch Reservoir in the northwest end of the Croton System; however DEP operations can choose to bypass West Branch entirely and divert directly to Kensico.

Mixed Catskill/Delaware water is diverted from Kensico to the Cat-Del UV Disinfection Facility via the continuation of the Delaware Aqueduct. Cat-Del UV is the largest disinfection facility in the world, with a capacity to treat over 2 BGD. After disinfection, mixed Catskill/Delaware water is sent to Hillview Reservoir and is distributed to the five boroughs.



Fig. 6 Kensico Reservoir and its associated infrastructure

1.1.5 System Interconnections

New York City's water supply system has a number of interconnections that increase operational flexibility. Two of the more important include the Shaft 4 Interconnection and the Croton Falls/Cross River Pump Stations.

The Shaft 4 Interconnection is a new piece of infrastructure that connects the Catskill and Delaware Aqueducts where they cross near Gardiner, New York (Fig. 7). The Shaft 4 Interconnection will allow DEP operators to take the Catskill System offline during turbidity events and fill the Catskill Aqueduct with water from Rondout to satisfy outside community demand. Additionally, the Interconnection provides an increase in transmission through the Delaware Aqueduct, which may be useful during drought or infrastructure outage scenarios.

The Croton Falls and Cross River Pump Stations allow water from Croton Falls Reservoir and Cross River Reservoir (Fig. 1) to be pumped into the Delaware



Fig. 7 Location of the Shaft 4 Interconnection

Aqueduct. Pumping out of the Croton System into the Delaware Aqueduct bypasses the New Croton Aqueduct in favor of Kensico Reservoir and the Delaware Aqueduct – which has a larger capacity for distribution. In certain drought and infrastructure outage scenarios, this flexibility can allow DEP to prolong diversions from Kensico and increase water supply reliability.

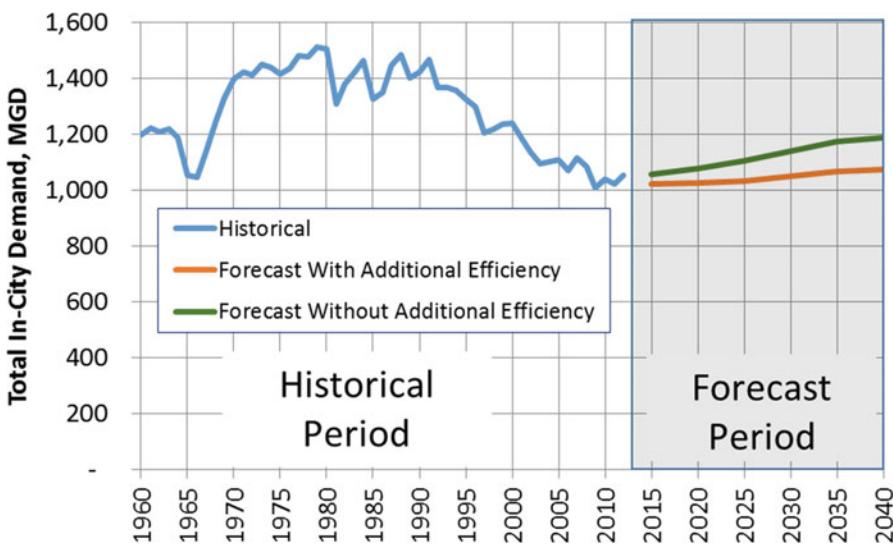


Fig. 8 Historical and projected annual average NYC demand

1.2 Key Operational Objectives

As a water supply utility DEP's main operational objectives are to provide reliable, clean water to their customers. However, with such a large watershed area in the highly populated US Northeast, additional objectives such as fisheries habitat, recreation, and spill mitigation must be considered. This section provides an overview of the different operating objectives of the NYC Water Supply System.

1.2.1 Water Supply Reliability

The main objective of any water supply utility is to reliably meet consumer demand. DEP is the primary water supplier for over nine million people in the greater New York City area for which demand is currently just over 1 BGD on an annual average basis and is projected to increase over the next 30 years (Fig. 8) (New York City Department of Environmental Protection 2014). Demand in New York City varies seasonally, with the months of July and August experiencing the highest values (Fig. 9). In response to this, DEP operates the system to refill each reservoir at or around June 1 each year and then draw down the reservoirs at the same time to keep the system balanced. These operations can require dynamic adjustments, for which hydrologic forecasts are important in helping DEP meet these goals.

Drought risk is also a concern for the NYC Water Supply System. Since the development of the NYC Water Supply System, the largest observed drought occurred between 1962 and 1966 (Seager et al. 2012). During this period, the NYC System was greatly stressed between 1963 and 1965, and a conservation campaign was launched to preserve water supply reliability (NYCDEP 2016). Since the 1960's drought, the City has established a drought plan for managing

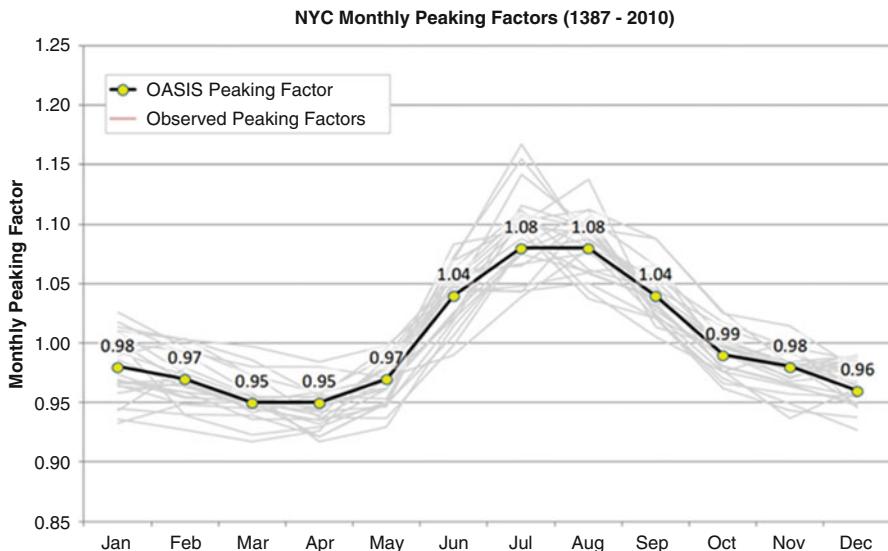


Fig. 9 Monthly peaking factors for NYC demand

water supply and demand during periods of drought. Depending on the severity of drought (Watch, Warning, or Emergency) mitigating actions include:

- Drought awareness and voluntary water use reductions
- Maximization of Croton WFP
- Utilization of Croton Falls and Cross River Pump Stations
- Activation of alternate water supply sources, including Queens groundwater and the Chelsea pump station
- Enhanced leak detection and repairs

1.2.2 Water Quality

Source water quality is a major operational objective for the NYC Water Supply System because the City has a FAD for its Catskill and Delaware Systems. Water quality in both systems is generally pristine, as the City maintains a comprehensive and successful watershed protection program. However, elevated turbidity is occasionally a concern in the Catskill System. Turbidity in the Catskill System reservoirs is typically less than 5 NTU. However, infrequent, extreme storm events can erode naturally occurring silt and clay deposits in stream banks and channels in the Schoharie and Ashokan watersheds, which can lead to elevated turbidity levels in Schoharie Reservoir, Shandaken Tunnel diversion, Esopus Creek, Ashokan Reservoir, and, occasionally, in Catskill Aqueduct diversions to Kensico Reservoir. Sustained periods of elevated turbidity in Ashokan Reservoir and in the Catskill Aqueduct may require treatment with alum to ensure safe drinking water supply and

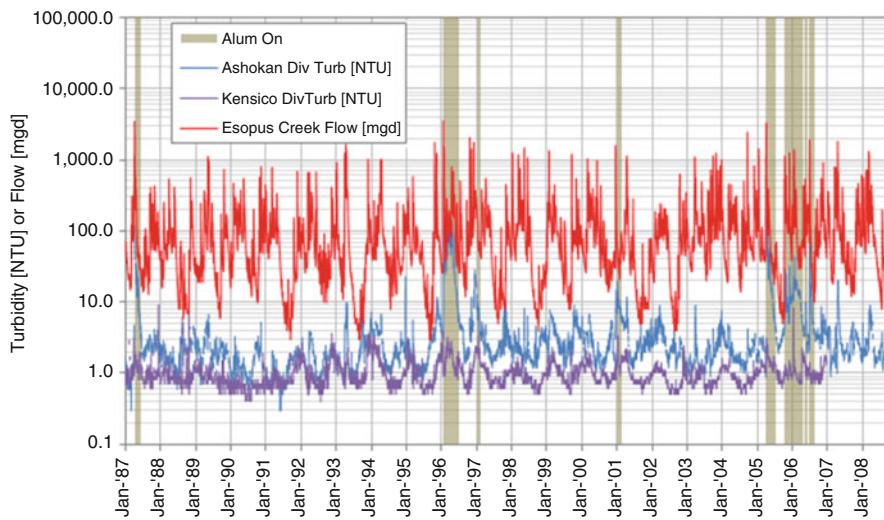


Fig. 10 Esopus Creek flow, Ashokan diversion turbidity, Kensico diversion turbidity, and periods of alum application (tan shading) (1/1/1987–9/30/2008)

maintain compliance with the Surface Water Treatment Rule (SWTR) 5 NTU raw (source) water turbidity limit for diversions from Kensico Reservoir (Fig. 10).

The USEPA has discouraged the practice of alum treatment in public water supply reservoirs since 1990. Further, as expressed in the 2007 FAD, USEPA and NYSDOH do not believe that DEP should rely extensively or exclusively on alum treatment to control turbidity and expect significant improvement in DEP's ability to prevent, manage, and control turbidity in the Catskill System in order to maintain long-term filtration avoidance. Infrastructure upgrades, including Shaft 4, and historical modeling performed under the Catskill Turbidity Control Studies showed that dynamic reservoir operations could be used to successfully control turbidity and limit the use of alum (Weiss et al. 2013). This work formed the basis for the Operations Support Tool (OST), which uses forecasts to assist operations staff in dynamically modifying reservoir operations.

1.2.3 Environmental Flow Requirements

New York City is required to meet a number of environmental flow requirements in each of their major systems. Environmental flow requirements, or conservation releases, serve the purpose of protecting downstream users and conserving fish habitat and stream ecosystems. Conservation releases in the Catskill and Croton System are relatively simple – by and large the requirements are constant or subscribe to a simple seasonal pattern (see Sects. 1.1.1 and 1.1.2). Release rules in the Delaware Basin are significantly more complex.

Releases from the Delaware Reservoirs (Cannonsville, Pepacton, and Neversink) are mandated by Supreme Court Decree in order to meet flow objectives at Montague, NJ. Since the passage of the Supreme Court Decree, a number of supplemental

release policies have been agreed to and implemented by the Decree Parties (New York, New Jersey, Pennsylvania, and Delaware) – the most recent being the Flexible Flow Management Program (FFMP). Flow rates prescribed by the FFMP were designed to mimic a more natural flow regime in the Delaware River and attempt to make more efficient use of excess unallocated water in Cannonsville, Pepacton, and Neversink. The FFMP consists of a number of different release tables or schedules, which relate reservoir elevation, time of year, and release rate. In 2011, the Decree Parties agreed to use the City's OST and hydrologic forecasts to support the selection of release schedules. More information about this strategy is presented in Sect. 3.4.

1.2.4 Other Objectives

Beyond water supply reliability, water quality, and environmental flow requirements, the NYC Water Supply System supports a number of secondary objectives including spill mitigation, downstream river recreation, and generation of hydropower.

1.3 Operations Support Tool Overview

New York City's Operations Support Tool (OST) uses near real-time data, hydrologic forecasting, and predictive modeling to guide reservoir system operations, identify future water stress conditions, and take preventive actions to maintain supply reliability, water quality, and environmental performance. OST consists of four major components: the data acquisition system, the OASIS water supply system model, the linked CE-QUAL-W2 reservoir water quality models, and the dashboard data visualization tool (Fig. 11).

1.3.1 Data Acquisition System

OST is initialized using near real-time system conditions and is driven forward using hydrologic forecasts. In order to accomplish this efficiently, OST was designed with an automated data acquisition system that collects, aggregates, and cleans data from multiple different sources, including (but not limited to):

- Reservoir operations data (e.g., storage, diversions, releases, spills, water quality) from the City's SCADA system
- Hydrologic forecasts from the NWS
- Observed streamflow from the USGS

Data is stored, cleaned, and served to the OASIS system model from a centralized database.

1.3.2 OASIS Water Supply System Model

OASIS is a generalized water supply system model developed by HydroLogics. OASIS represents a water supply system using a network of nodes (reservoirs, junctions) and arcs (aqueducts, streams) to simulate routing water through the

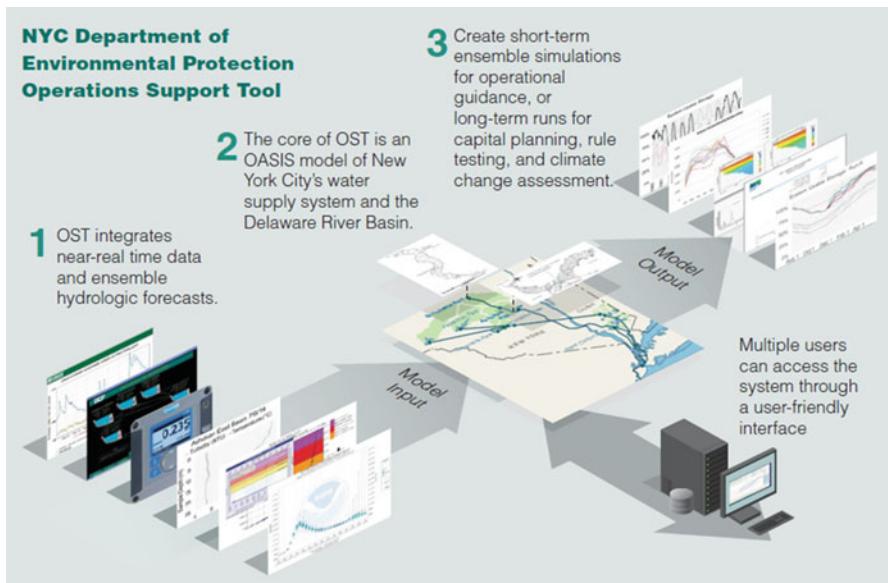


Fig. 11 Operations Support Tool Schematic

system (e.g., reservoir releases or diversions). OASIS is not a hydraulic model; it operates at a daily time step using average flow rate and/or cumulative daily volumes. The model uses linear programming optimization to determine the “best” way to route water subject to both human operating rules and physical constraints. Though the model can be constrained to reproduce historical operations and flows in the system, it is more powerful when used as a tool for scenario/alternatives analysis. OASIS enables the user to easily modify operating rules and execute multiple alternative runs, allowing for efficient “what-if” analyses prior to altering actual operations.

The OASIS model of the City’s reservoir system simulates daily operations throughout the entire system and Delaware River Basin. Major components of the OASIS model include inflows, linked hydrodynamic water quality model, demands, system physical data, and operating rules. These components are described in the following sections.

OASIS Inflows

All nodes in the model require an inflow source in order to route water downstream. At each node, inflows need to be the total unimpaired watershed flow to that location. This is an important concept, because the point of the model is to test and analyze alternate operating scenarios. In order for this to be performed accurately, any historical operational influences must be removed prior to application in the model. For example, at gauges below an impairment, such as a reservoir,

man-made influences (e.g., consumptive demand, releases, or spills) need to be removed from the raw gauge flow.

Inflows into OASIS may either be historical or forecasted. When executed to evaluate long-term performance of the system, inflows should be a long historical record of inflows. When executed in real time to evaluate short- to mid-term operational adjustments, inflows to OST are an ensemble forecast (Sect. 2).

Linked Hydrodynamic Water Quality Model

OST incorporates two-dimensional, hydrodynamic water quality models at Schoharie, Ashokan, Kensico, and Rondout. These models are developed in CE-QUAL-W2 (W2), a dynamic, laterally averaged, two-dimensional (longitudinal-vertical) hydrothermal/transport model developed by the Army Corps of Engineers (Cole and Wells 2006). In addition to the underlying fluid motion and mass transport framework, the W2 models include a three particle size class turbidity sub-model that simulates the fate and transport of turbidity in the reservoirs and accounts for both settling and resuspension processes. Specification of model coefficients and model testing for the Catskill W2 models is supported by process studies and by extensive, temporally and spatially detailed, in-reservoir automated and event-based water quality monitoring (Gelda and Effler 2007a, b, c).

Each of the W2 models is dynamically linked to the OASIS model. The water quality models run in parallel with OASIS, such that for each simulation day, the W2 models simulated 1 day of reservoir water quality before OASIS continues on to the next simulation day. In this way, OASIS-simulated water supply decisions are informed by the simulated water quality and vice versa. Therefore, the linked water supply-water quality model is able to simulate the feedback between reservoir operations (e.g., diversion and release decisions) and water quality for the Catskill System and Kensico Reservoir. Additionally, the model infrastructure has been developed to automatically initialize and “spin up” the water quality profiles in the W2 models, significantly reducing the effort required to set up and execute OST simulations for near real-time operational support.

Demands

Demands in OASIS are representative of water that is removed from the system as consumptive use. Demands may be represented as time series data of historical (or forecasted) consumptive use or repeating monthly/seasonal patterns. When executed to evaluate long-term performance of the system, demands are typically represented as monthly patterns. When executed in real time to evaluate short- to mid-term operational adjustments, OST utilizes forecasts of total diversions from Kensico and New Croton to meet demands from New York City and outside communities serviced by the NYC System.

Physical Data

The OASIS model includes data that represent physical constraints on the flow and storage of water (e.g., spillway rating curves, maximum capacities of aqueducts and release works, elevations of structures, reservoir storage-elevation curves).

Operating Rules

Operating rules in OASIS are coded and controlled through its operations control language (OCL), which is a flexible, script-based interface. In OCL, the user is able to define inputs to the linear program which take the form of targets and constraints. Targets, such as minimum flow requirements and demands, are operating rules that the user would *like* to see met, but ultimately may not be met under all circumstances. For example, in a severe drought, we may want to see a conservation release rate achieved, but it may not be possible given the amount of water in a reservoir. Constraints are operating conditions that cannot be violated under any circumstances and are typically reserved for representing physical limitations of the system. For example, a constraint may represent the maximum flow in an aqueduct (Fig. 12).

The user is able to control operational preferences by setting weights on targets in the linear program. Weights are multipliers that standardize decision variables (controlled flows and volumes) into points; the objective function of the OASIS linear program is to maximize the number of points on each day in the simulation. Consider the following example:

In OST, baseline operating rules have been tested for satisfactory long-term performance over an 82-year hydrologic record. Performance is defined by metrics measuring water supply reliability (e.g., probability of refill, minimum annual storage) and water quality (e.g., diversion turbidity, modeled days of alum addition). The baseline operating rules include:

- Diversion preferences in the Delaware, Catskill, and Croton Systems accounting for overall balance between the three systems, Catskill System turbidity, and forecasted reservoir inflow and associated probabilities of refill.
- Infrastructure configuration, including the operation of interconnections between systems for the purposes of turbidity control, drought preparedness, and spill mitigation.
- Release targets for all NYC reservoirs including:
 - Croton System (per requirements in NYSDEC Part 672-3)

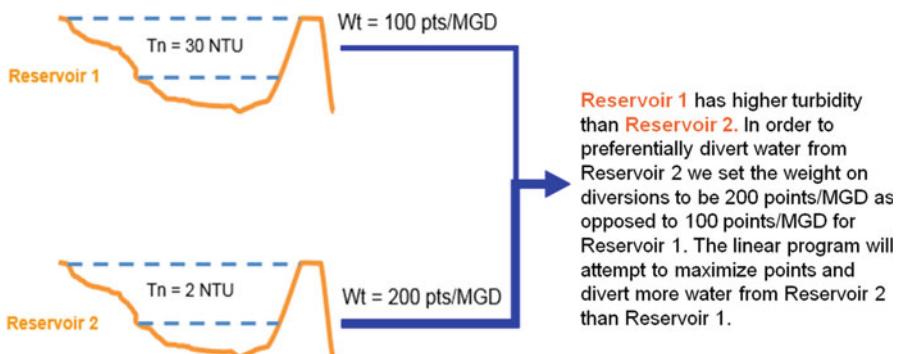


Fig. 12 Simple operating rule example

- Schoharie Reservoir (per requirements in NYSDEC Part 670 and the Shandaken SPDES Permit)
- Ashokan Reservoir (per the current Ashokan Interim Release Protocol (New York State Department of Environmental Conservation/New York City Department of Environmental Protection (DEC/DEP) 2011))
- Rondout Reservoir (per requirements in NYSDEC Part 672–2)
- Cannonsville, Pepacton, and Neversink Reservoirs (per the current FFMP (Decree Parties to the 1954 U.S. Supreme Court 2016)).
- Release, diversion, and demand rules for all lower Delaware Basin reservoirs

A number of operating rules in the system are informed by hydrologic forecasts – these are further discussed in Sect. 3.

1.3.3 Dashboard Data Visualization

A key feature of OST is the ability to automatically view observed data (e.g., historical or current streamflow, precipitation and snowpack levels, reservoir storage elevations) with projected data (e.g., the range of potential future reservoir storage levels derived from ensemble simulations using NYCDEP's OASIS system operations model). This functionality provides decision makers with both current situational awareness and insight into how today's operation decisions affect future water quality and supply reliability. The OST Dashboard allows users to readily access the data relevant to the current system condition, to review the impacts and benefits of alternative operating decisions, and to summarize and disseminate charts, tables, and custom reports that synthesize the current decision-making context (Fig. 13).

2 Ensemble Forecasts for NYC OST

New York City water supply decisions are based on a large collection of historical and real-time water quality and quantity data as well as forecast information and operator experience. The OST provides a framework that supports risk-based decision making through the use of ensemble forecasts of reservoir system inflows and streamflow at selected locations. When the ensemble forecasts are coupled with NYC Water Supply System OASIS model, they provide a means for water managers to explore a range of scenarios of future system inflows, operational alternatives, and future system states (e.g., reservoir water levels, reservoir turbidity). System operating rules have been developed based on the probability of reaching critical thresholds of system states to provide operators with guidance for risk-based decision making.

2.1 Forecast Types

OST was configured to support five different ensemble forecast options that are produced with different methodologies. These options include an approach



Fig. 13 Screenshot of the OST Dashboard

assuming climatology, two statistical methods based on streamflow data, and two methods that use hydrologic models to forecast streamflow based on meteorological forecasts.

2.1.1 Statistical Forecast Methods

The most basic method of producing a forecast ensemble is to simply use the historical flow data. This method assumes that each year of historical data, beginning with the current date, represents a possible future scenario. The method does not account for any knowledge of the current state of the system (i.e., conditions) and captures only the climatological uncertainty in the ensemble. The second ensemble forecast option accounts for the soil moisture state of a watershed through the use of recent streamflow data (Hirsch 1981). Hirsch used monthly historical streamflow data and an autoregressive moving average model to generate ensembles of monthly future streamflow. In OST these data are disaggregated based on ratios between historical daily and monthly flows to daily values for input to the system model.

In an attempt to extract more information from the recent observed daily streamflow, a General Linear Model (GLM) was applied as a third option in OST to directly generate forecast ensembles of daily reservoir inflows. The GLM as described in Zhao et al. (2011) uses the correlation structure in daily streamflow, which has been transformed with a normal quantile transformation, to generate forecasts of daily flows. This OST ensemble option is called eHirsch, for “extended Hirsch,” and is expected to provide better results than the Hirsch option in the near term and comparable results beyond the first week or two.

2.1.2 Physically Based Methods

The fourth ensemble forecast option is the classical National Weather Service (NWS) Ensemble Streamflow Prediction (ESP) methodology (Day 1985). The

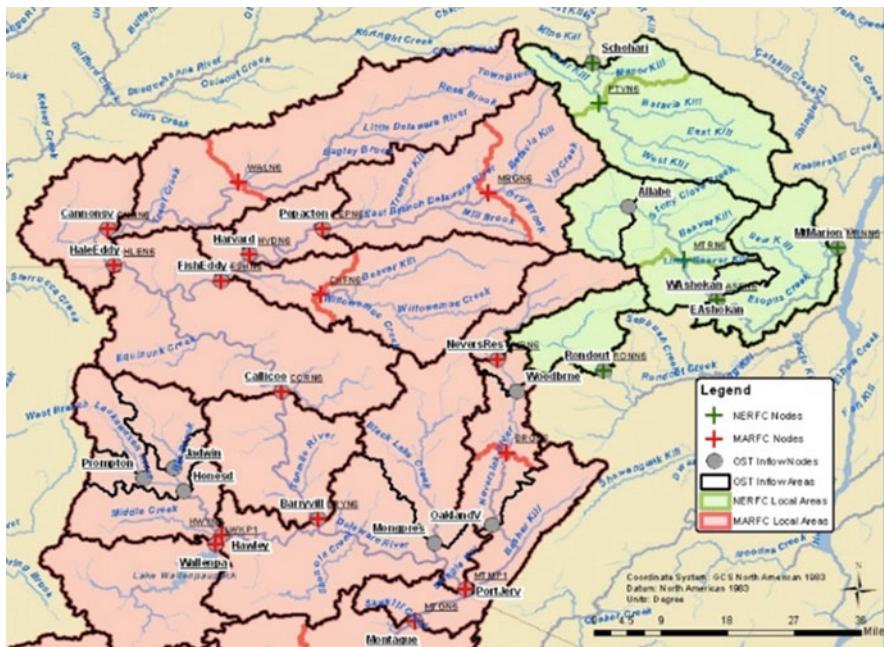


Fig. 14 DEP OASIS model nodes and NWS forecast points

Northeast River Forecast Center (NERFC) forecasts the Catskill and Croton New York City water supply watersheds, and the Middle Atlantic River Forecast Center (MARFC) forecasts the Delaware watershed part of the system. NYCDEP coordinated with NERFC and MARFC to provide ESP forecasts for the forecast points within their system. The NWS forecast points and the nodes for the OST are shown in Fig. 14 for the major part of the New York City water supply watersheds. The lower part of the Delaware watershed and the Croton watershed, which is east of the Hudson, are not shown in the figure, but they are also modeled in OST. In many cases, the NWS forecast points line up with the location of OST nodes; however, there are (1) cases where there are multiple OST nodes above a NWS forecast point, (2) cases where there are multiple NWS forecast points above an OST node, and (3) cases where the locations are different. NWS provides ensemble forecasts for the local areas above each forecast point to facilitate their use in OST. An ensemble postprocessor (EPP) was developed to produce a set of ensemble forecasts for the local area above each OST node given the available ensemble forecasts at the NWS forecast points. The EPP will be discussed in a later section.

ESP has been implemented across the country as part of the Advanced Hydrologic Prediction Services (AHPS) program. The ESP approach accounts for current watershed conditions (e.g., soil moisture and snowpack) through the use of continuous hydrologic models that are operated and maintained by the NWS River Forecast Centers (RFCs). NERFC uses the SNOW-17 temperature index snow

model (Anderson and Crawford 1964) and the Sacramento Soil Moisture Accounting (SACCSMA) model (Burnash et al. 1973). MARFC uses the SNOW-17 model and the Continuous Antecedent Precipitation Index runoff model (Sittner et al. 1969). Ensembles are generated with ESP by assuming that historical years of meteorological data represent samples of possible future occurrences. Some applications of the ESP methodology have attempted to blend short-term meteorological forecasts with the climatological data. Other applications have used a year weighting approach to account for the fact that the historical years may not be equally likely given current atmospheric conditions. As ESP has been implemented for the NYCDEP area, it does not account for any weather forecast or current climate information. In general, ESP assumes that the major source of uncertainty is due to climatology; however, the EPP mentioned above is designed to adjust the ensemble to also reflect hydrologic model uncertainty.

The fifth ensemble forecast option is the latest NWS ensemble forecasting system, the Hydrologic Ensemble Forecast Service (HEFS), which is described in Demargne et al. (2014) and schematically shown in Fig. 15. This latest ensemble system, which NWS has implemented in the Community Hydrologic Prediction System (CHPS) environment, includes a Multi-Ensemble Forecast Processor (MEFP) that processes RFC short-term (typically 1–3 days) meteorological forecasts, 15-day forecasts from the Global Ensemble Forecast System (GEFS), and 9-month forecasts from the Climate Forecast System (CFSv2) to produce unbiased forecast ensembles of precipitation and temperature that reflect the uncertainty of the forecasts. Forecasts beyond the 9-month lead time are based only on climatology. The MEFP removes systematic biases in the RFC, GEFS, and CFSv2 meteorological forecasts. The MEFP also downscale the GEFS and CFSv2 meteorological

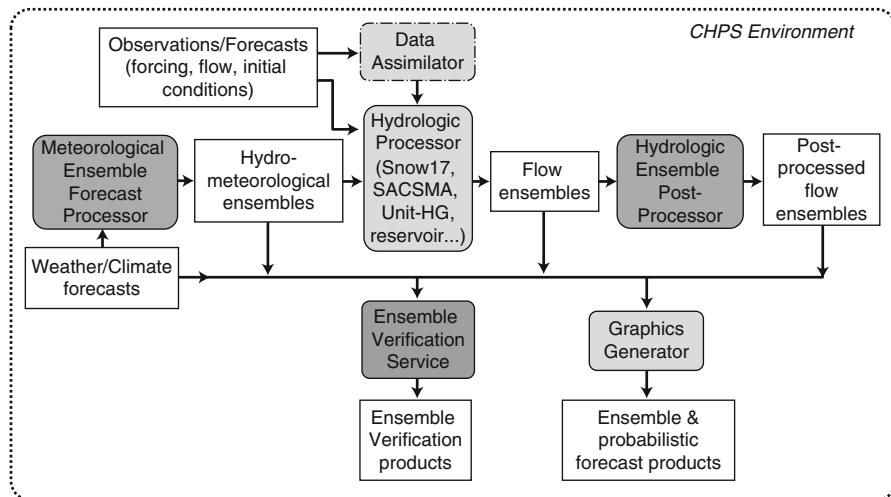


Fig. 15 NWS Hydrologic Ensemble Forecast Service schematic. (Taken from Demargne et al. 2014)

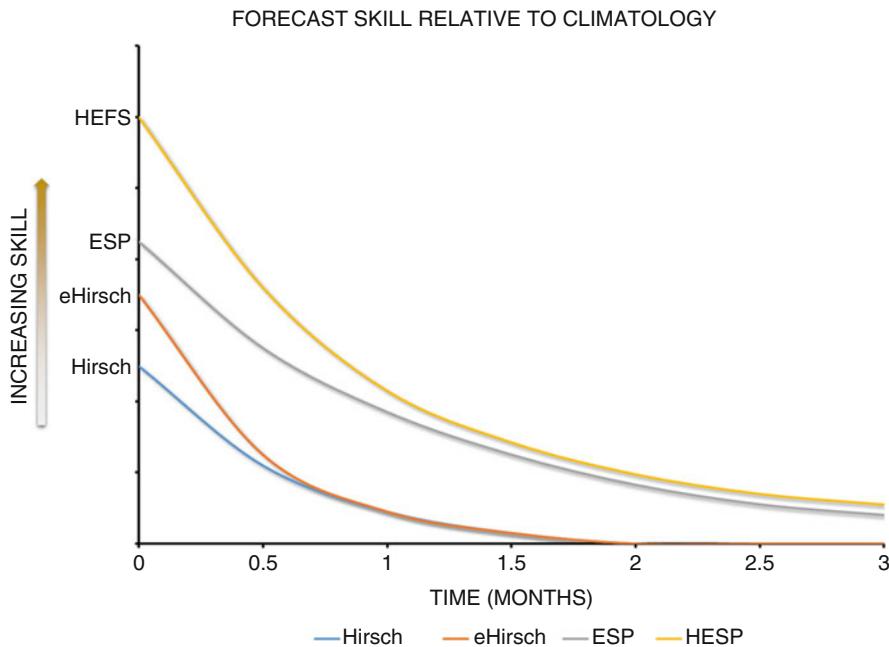


Fig. 16 Conceptual forecast skill relative to climatology

forecasts to the scale of each hydrological sub-basin. And the MEFP generates ensemble precipitation and temperature forecasts for input to the NWS hydrologic forecast models that generate the HEFS ensemble hydrologic forecasts. NYCDEP receives raw HEFS forecasts, i.e., forecasts that have not been postprocessed, to allow the use of the EPP mentioned above.

The NWS RFCs are currently providing NYCDEP with raw forecast ensembles from both the ESP and HEFS approaches. Of the five ensemble forecast options, the HEFS option is typically the first choice because it accounts for the current watershed conditions and incorporates the available weather/climate forecast information. Figure 16 conceptually shows how the forecast skill is expected to vary with lead time for the different forecast options. The OST team has also developed a forecast verification tool to allow quantitative comparison of the forecast options.

2.2 Ensemble Postprocessor (EPP)

OST requires ensemble forecasts of the inflows to a set of OST inflow nodes. Several approaches to creating these ensemble forecasts are available to OST. The role of EPP is to produce ensemble inflow forecasts on the basis of ensemble hydrologic forecasts from the NWS at a set of NWS forecast points. Some of the NWS forecast points are colocated with OST inflow nodes. But many are not colocated.

EPP is a critical component of OST, because it processes the raw ESP and HEFS forecasts for NWS forecast points to create ensemble forecasts for all the OST nodes. A different EPP segment produces the ensemble inflow time series for each OST input node. As part of the processing, EPP deals with a number of issues, including: (1) differences between the locations of the NWS forecast points and the OST nodes, (2) cases where NWS forecasts need to be combined to estimate an OST node, (3) cases where EPP is used to subdivide an NWS forecast to create estimates for multiple OST nodes, and (4) hydrologic model bias and uncertainty. Hydrologic model bias may be the result of the model structure (e.g., inaccurate representation of local processes), a less than optimal set of parameters, and/or a difference between the location of the NWS forecast point and the OST node (e.g., the time series used for NWS model calibration might be different than the one NYCDEP associates with the OST node). The uncertainty is due to an inability of the hydrologic model to adequately capture the watershed processes, processes not accounted for (e.g., diversions), errors in the meteorological forcings, and uncertainty of initial model states.

The EPP design is based on two fundamental assumptions. The first is that the hydrologic ensemble forecasts being processed were produced by a hydrologic forecast system that processes meteorological inputs having the same climatological properties as the meteorological inputs used to create the historical forecast simulations that were used to create the EPP parameter file. This assumption is met by NWS HEFS and AHPS ESP forecasts. The NWS HEFS includes MEFP that is calibrated to process GEFS and CFSv2 meteorological inputs to produce ensemble forcing for input to the NWS hydrologic forecast models. The NWS has shown that the MEFP satisfies this assumption.

The second assumption is that the effects of hydrologic forecast model error can be accounted for by (i) extracting “information” contained in hydrologic forecast model simulations about the flows that are observed (e.g., in USGS gaged flow measurements or OST inflow estimates) and (ii) using this information to generate an ensemble of equally likely simulated traces of the observed flow. The ensemble simulation processor described below was developed to test this assumption. Test results have shown that this assumption is satisfied.

An additional assumption is made that input and output streamflow variables, which have highly skewed probability distributions, can be transformed to variables that have standard normal distributions. Moreover, it is assumed the joint distribution of any two input and output standard normal variables can satisfactorily be approximated by a bivariate normal distribution, and any multiple combinations of these standard normal variables have a multivariate normal distribution. This kind of assumption has been widely used in hydrology.

The EPP includes two basic algorithms. The first algorithm is a GLM (Zhao et al. 2011) that essentially transforms an input vector containing information derived from an input forecast into an ensemble of output vectors that include all of the information in the input vector but that also account for effects of hydrologic model uncertainty. EPP allows these vectors to be partitioned to contain data for multiple inputs and/or outputs corresponding to different input/output nodes/locations. The

scientific foundation for the GLM is based on first and second moment properties of random variables. The second algorithm is an Autoregressive Transfer Function Processor (ARTFP) that uses a small imbedded version of the GLM that operates recursively in conjunction with an autoregressive lag-1 Markov model.

There are two versions of the EPP: an operational version and a reforecast version. The operational processor is used to postprocess real-time NWS ESP or HEFS forecasts. The reforecast version is used to postprocess archived NWS ESP and HEFS forecasts.

Operation of the EPP is controlled by a parameter file. Each EPP segment must have its own parameter file. Each EPP parameter file specifies which input time series and output time series are used for that segment. It specifies exactly how the EPP algorithms are to be used for the segment. It provides information needed to transform input values into standard normal deviates and to transform processed output standard normal deviates into final output values. The same parameter file is used by the operational postprocessor and the reforecast postprocessor.

A third version of the EPP, the ensemble simulation processor, was used to develop and test the actual EPP configurations for each OST inflow node. The ensemble simulation processor is used to generate an ensemble of equally likely traces of estimates of the streamflow that will actually occur for a given historical streamflow trace. The EPP simulation version uses historical NWS forecast model simulations (based on historical observed precipitation and temperature input data) to make ensemble predictions of the historically observed (or estimated) inflow data to corresponding OST inflow node(s). This makes it possible to test if the EPP has satisfied the second fundamental assumption, noted above, that the EPP can adequately account for the bias and uncertainty inherent in hydrologic forecast model error. This processor was used to generate ensemble simulations for each EPP segment. Verification statistics were computed for each of these using an OST forecast verification tool. Results demonstrate that the EPP adequately accounts for effects of hydrologic forecast model error.

3 Applying Ensemble Forecasts in Operations

With the integration of ensemble forecasts into OST, DEP has the ability to test alternative operating strategies in real time and quantitatively examine hydrologic risk in their operational decision making. This section provides a summary for how DEP incorporates forecasts and probabilistic risk into their operational decision-making framework and explores a number of operational examples showcasing this strategy.

3.1 Transition to Probabilistic Analysis and Decision Making

Since the implementation of OST, DEP has modified the way operational decisions are made in the water supply system. OST has allowed DEP to examine operational alternatives in a more iterative, systematic framework (Fig. 17).

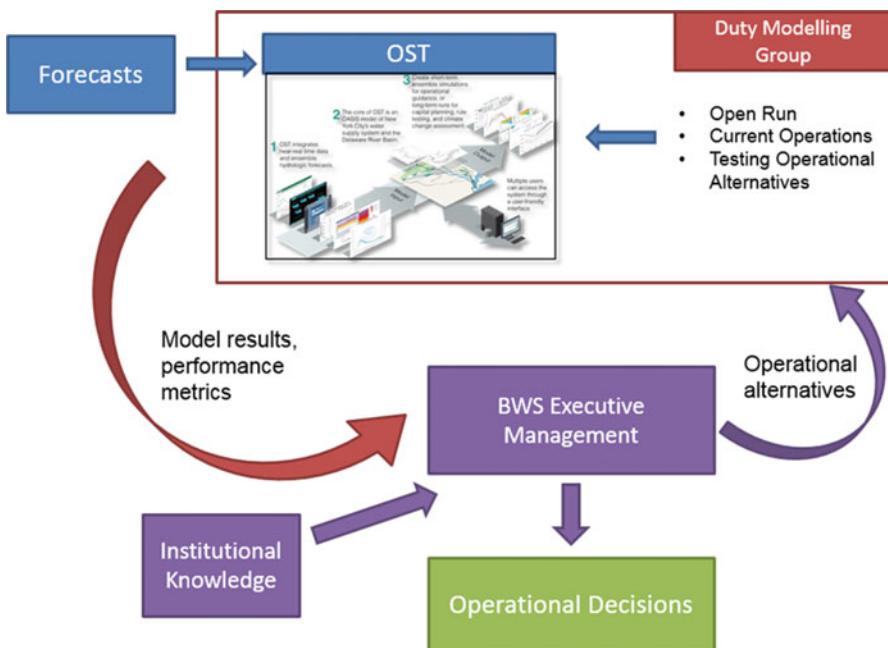


Fig. 17 DEP's operational decision-making framework

DEP has established a “Duty Modeler Group” within their operations staff that is responsible for executing operational support model runs several times a week. All runs are initialized with the most current reservoir surface water elevation conditions and are driven forward using a selected ensemble hydrologic forecast. These runs consist of at least two different alternatives: the “open run,” which executes the current baseline operating rules without any operational modifications or temporary infrastructure constraints, and the “current operations run,” which sets major aqueduct diversions and reservoir releases to the current operational conditions for the first forecasted week of the run. Sometimes, including during turbidity events and infrastructure shutdown, operational conditions can be modeled for a period longer than a week. Once the runs have completed, the duty modeler uses the OST Dashboard to postprocess the run and return a number of standardized plots and performance metrics including:

- Time series plots showing reservoir inflows, elevations, diversions, and releases for recently observed data as well as the modeled projections. The duty modeler typically chooses whether to show the modeled projections as ensemble traces, aggregated percentiles of the ensemble (e.g., 10th, 50th, 90th percentiles), or both.
- Probability distributions of reservoir usable storage aggregated over varying time horizons.

- Probability of reservoir refill at or around June 1.
- Regulatory summaries for anticipated release rates in the Delaware System (see Sect. 3.4).
- During Catskill turbidity events, the duty modeler may return plots of Ashokan and Kensico diversion turbidity and metrics detailing the likelihood of alum addition in Kensico Reservoir.

After the runs are postprocessed, executive operations staff reviews the performance metrics and, using institutional knowledge, propose operational alternatives if necessary. The duty modeler takes this input and makes additional OST runs, called “testing operational alternatives” (TOA), and postprocesses the model results in the same manner described above. This process continues in a loop until the executive operations staff is satisfied that the performance metrics sufficiently meet DEP’s operational objectives.

3.2 Spill Mitigation

Though DEP’s main operational objective is meeting water supply needs, a number of the reservoirs in the system have spill mitigation objectives. New York City’s Delaware Basin reservoirs (Cannonsville, Pepacton, and Neversink) and Ashokan Reservoir in the Catskill System have formalized policies for spill mitigation. Each of these reservoirs has a Conditional Seasonal Storage Objective (CSSO) guide curve (see Fig. 4 for the Ashokan CSSO). The CSSO creates a target 10–15% void space in each reservoir between the fall and early spring to help reduce spills and mitigate downstream flood events. In order to be protective of DEP’s supply reliability objectives, the CSSO ramps up to 100% usable storage in late spring; similarly, the CSSO ramps down to the void target (10 or 15%, varies by reservoir) between July and the fall season. DEP uses hydrologic forecasts and the OST in determining appropriate release volumes to maintain the CSSO and protect water supply reliability. This section presents a case study from examining spill mitigation operations at Ashokan in winter 2014.

By February 2014, a relatively large snowpack had built up in the Catskill Mountains (Figs. 18 and 19). Beginning in mid-February, DEP began examining increased releases from the Ashokan Release Channel (ARC) in an attempt to create void space in the reservoir to manage the incoming spring snowmelt. Using the decision-making framework outlined in Sect. 3.1, DEP operations staff examined 38-day release rates of 100, 200, 300, and 600 MGD and examined the likelihood of maintaining storage below the CSSO (Fig. 21) as well as impacts on the probability of June 1 refill (Fig. 20).

Figure 21 shows a set of time series plots presenting observed usable storage in Ashokan (black dotted line) along with a number of statistics (minimum, maximum, and quantiles (75th, 50th, and 25th)) summarizing the ensemble modeled storage relative to the CSSO (dotted red line). Note that while the ARC release rate does not appear to have an effect on the ensemble maximum usable storage, the 75th

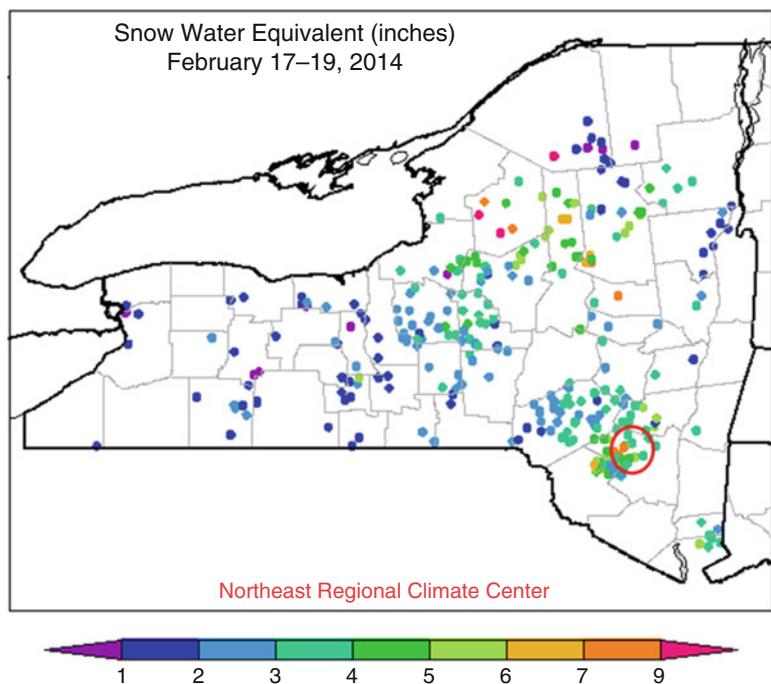


Fig. 18 NWS snow water equivalent surveys in the Ashokan watershed

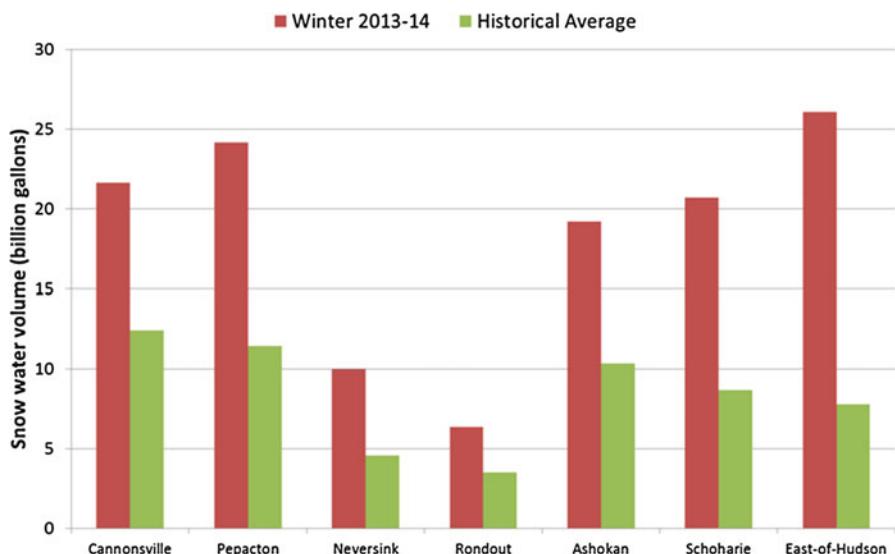


Fig. 19 DEP snow survey data, February 20, 2014

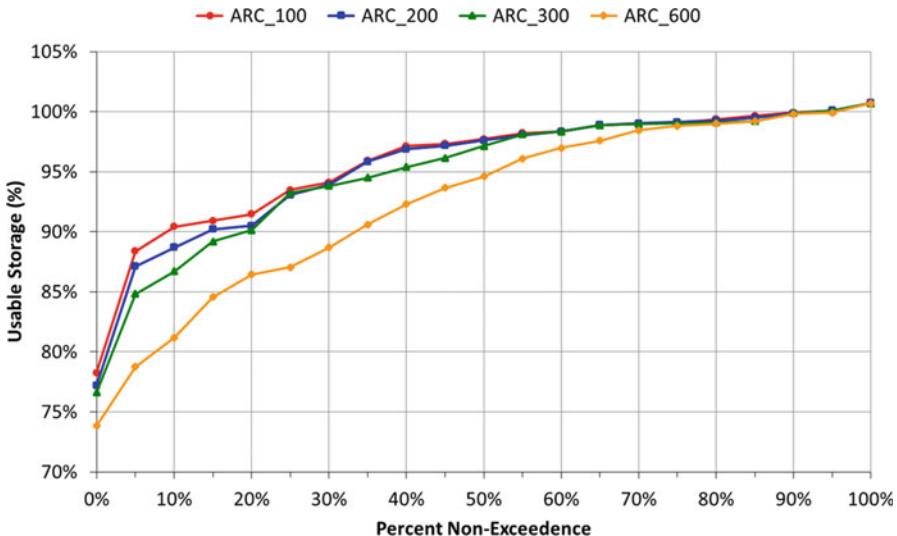


Fig. 20 February 19, 2014 OST simulations showing impacts on Ashokan June 1 refill

percentile drops below the CSSO for the remainder of the simulated period at a release rate between 300 and 600 MGD. This indicates that approximately 75% of the ensemble traces did not exceed the CSSO given a release rate above 300 MGD.

Releases out of the ARC drain to Lower Esopus Creek and are not recaptured by another DEP reservoir, so they represent a loss of water from the NYC System; thus it is important to consider the water supply reliability trade-offs associated with the release. Figure 20 shows probability distributions of modeled Ashokan usable storage on June 1. With the exception of the left-hand tail, the usable storage distributions for runs reflecting releases 100–300 MGD are fairly close together. At release rates between 100 and 300 MGD, there is at least an 80% probability that Ashokan will refill to at least 90% usable storage. Meanwhile, the 600 MGD scenario has about a 65% chance to refill to at least 90%.

Based on these metrics, DEP decided to release from the ARC at 400 MGD in mid-February. Throughout the rest of the month and continuing into the spring, DEP reevaluated the ARC release rate by reproducing the above analysis on a daily basis with updated reservoir conditions and forecasts. After a week of releasing at 400 MGD, DEP ramped releases down to 300 MGD for close to 3 weeks (Fig. 22). By mid-March, DEP developed close to a 20 BG void in the reservoir. Put into a historical context, the drawdown storage in mid-March 2014 was below the 10th percentile of recorded reservoir operations between 1983 and 2013 (Fig. 23). This void space was critical in capturing the spring snowmelt and was successful in preventing/mitigating spring spill events. Furthermore, Fig. 23 shows that the reservoir did indeed refill prior to June 1, indicating that the enhanced releases did not adversely affect refill in this instance.

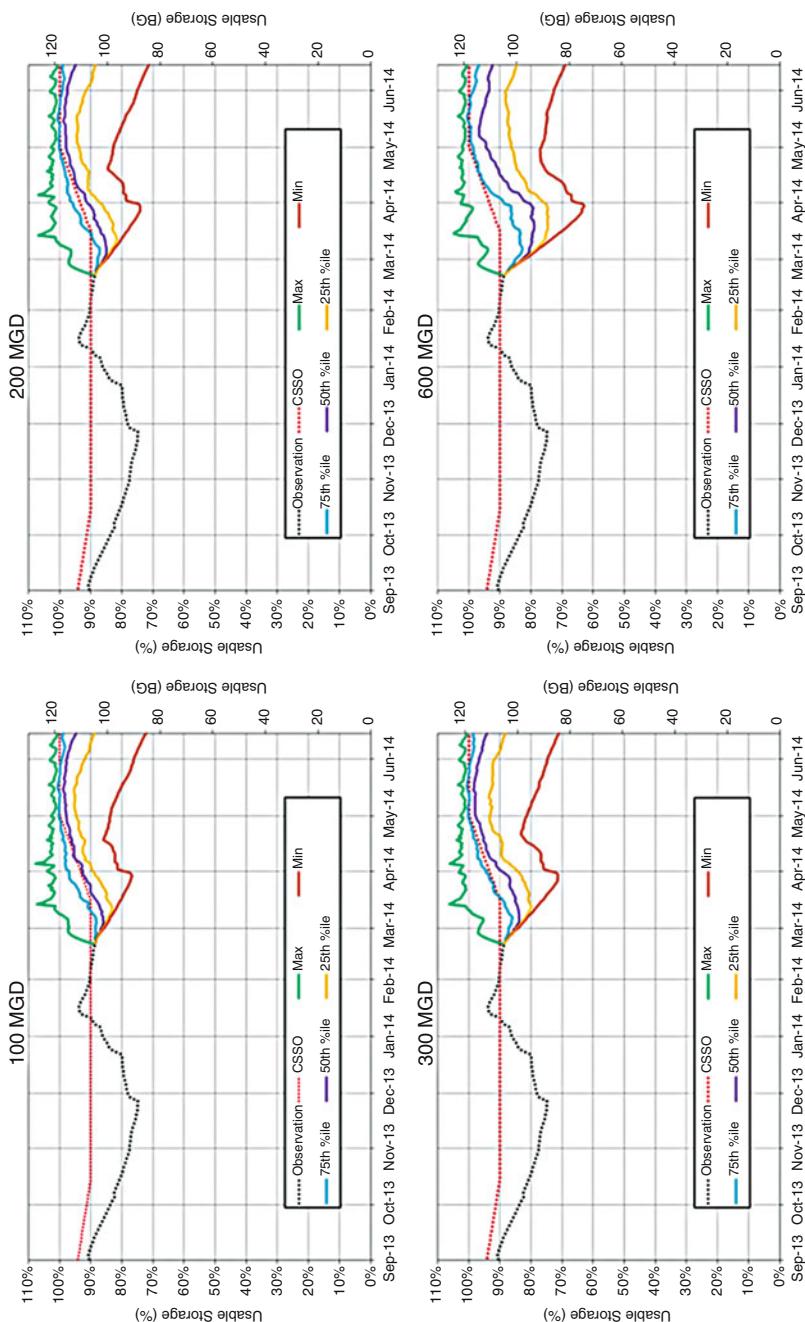


Fig. 21 February 19, 2014 OST simulations showing impacts on Ashokan usable storage given ARC release rates of 100–600 MGD

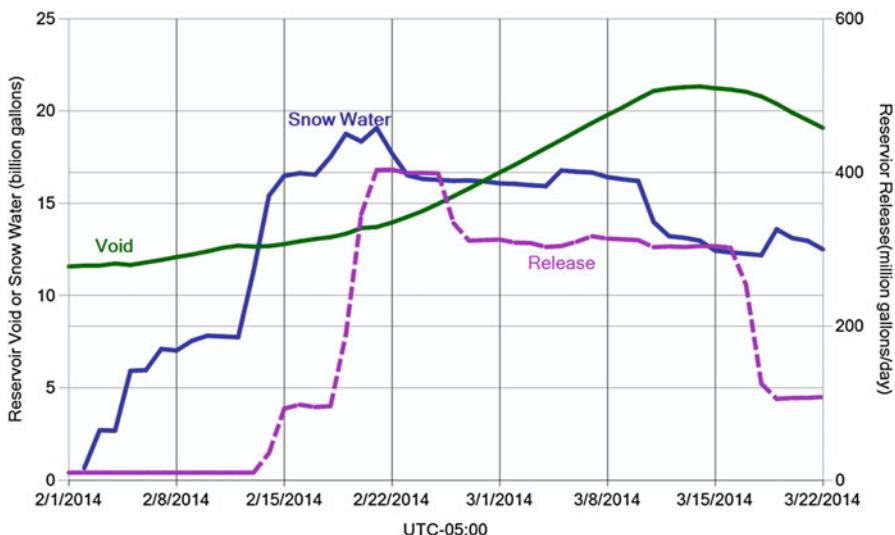


Fig. 22 Actual Ashokan void, release, and observed snow water equivalence, winter-spring 2014

OST, HEFS, and DEP's decision-making framework were critical in this operational case study. OST and the decision framework allowed operations staff to quickly model different ARC release alternatives and evaluate their efficacy and probabilistic risk to water supply reliability. Specifically, the information on range and likelihood provided by the HEFS ensemble forecast gave DEP managers confidence that the reservoir would refill even with sustained releases drawing reservoir storage to a very low level. Without this information (e.g., with only a traditional deterministic forecast), managers would have acted more conservatively and not drawn the reservoir down so far. This type of uncertainty information is a fundamental and paradigm-shifting element of ensemble forecasts for operational decision-making.

3.3 Water Quality

New York City has significant operational flexibility in managing the water supply system during turbidity events. One of the major operational turbidity control strategies is making preferential diversions from Ashokan's East Basin. Recall that Ashokan is physically separated into two basins and that the majority of the flow from Esopus Creek enters Ashokan through its West Basin (Sect. 1.1.2). During turbidity events, if there is sufficient storage in the West Basin, it can be used as a settling pond allowing DEP to divert less turbid water from the East Basin. In order to effectively utilize this strategy, it is important to minimize transfer of turbid water from the West Basin to the East Basin.

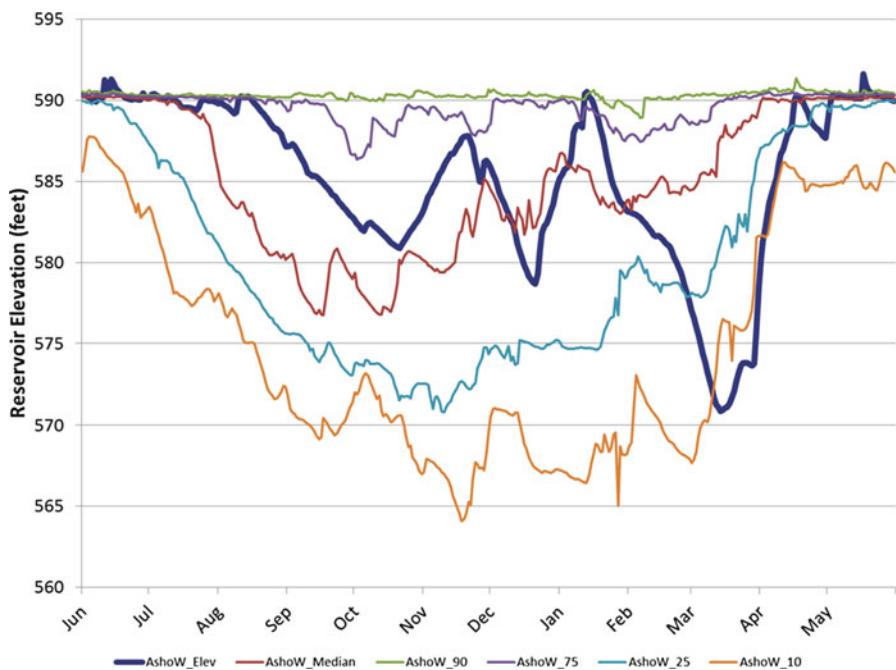


Fig. 23 Ashokan West Basin elevation in 2014 (thick blue line) compared with statistics of historical Ashokan elevations

Turbid water transfers to the East Basin can be minimized by making void space in the West Basin. Ideally, void space in the West Basin can be developed by making preemptive releases prior to a storm event. The operational strategy for preemptive releases is very similar to the spill mitigation example presented in Sect. 3.2; however, in this context the major performance metrics would be the likelihood of spilling water over the dividing weir, Ashokan diversion turbidity, and probability of refill. DEP operations staff would iterate OST runs searching for a release rate and duration that would minimize spill over the dividing weir and Ashokan diversion turbidity, without adversely affecting the refill probability (Fig. 24).

In the event that DEP is unable to make preemptive releases prior to a storm/turbidity event, operations staff has the option of making releases out of the West Basin after the storm has passed. The Ashokan IRP allows for the limited discharge of turbid water under the Operational Release Protocol (New York State Department of Environmental Conservation/New York City Department of Environmental Protection (DEC/DEP) 2011). Turbid water releases are limited based on time of year, discharge turbidity, and inflow turbidity to the reservoir. The following example details an instance where DEP made use of operational releases for turbidity control.

Figure 25 shows a time series plot of inflow into Ashokan Reservoir along with incoming turbidity. A large storm at the end of January filled the West Basin and resulted in elevated turbidity. After the event, DEP ran a number of OST simulations



Fig. 24 Turbid water spilling into the East Basin after Hurricane Sandy

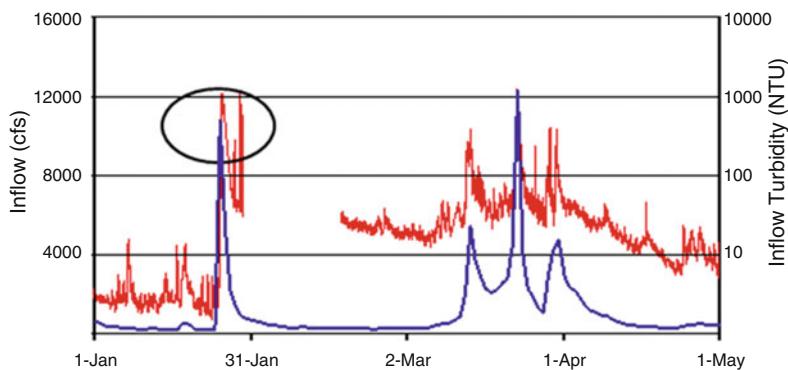


Fig. 25 Observed inflow (in blue) and incoming turbidity (red) into Ashokan. OST runs made after the circled event

examining operational releases from the West Basin. OST was initialized with the current reservoir elevations and water quality profiles and driven forward using ensemble hydrologic forecasts. Figure 26 examines the probability of withdrawal turbidity exceeding 10 NTU for two different scenarios, a constant 350 MGD release from the West Basin compared to a scenario with no releases. Without making releases (red line), there was a high probability that withdrawal turbidity would

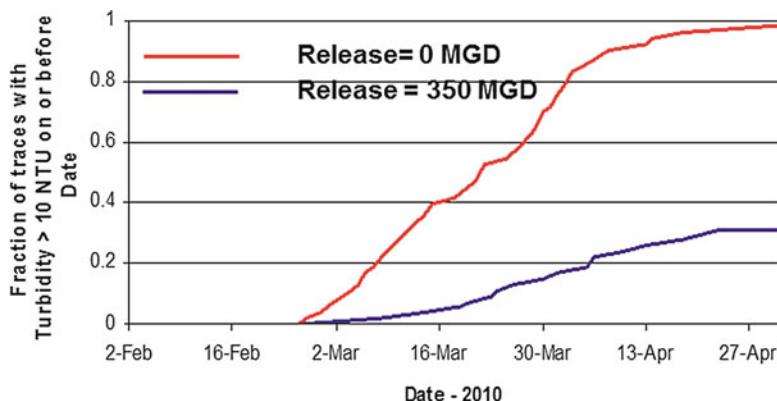


Fig. 26 Probability of withdrawal turbidity greater than 10 NTU

exceed the 10 NTU threshold. By the end of February, the exceedance probably was non-zero and over 80% by the end of March. Comparatively, releases at 350 MGD resulted in a lower probability of exceeding 10 NTU; by the end of March, the probably was still below 40%.

Though releases at 350 MGD clearly reduced the risk of turbid water withdrawals from Ashokan, there was a non-negligible impact on water supply reliability in the Catskill System. Figure 27 shows “rainbow plots” of the modeled ensemble usable storage comparing the no release and 350 MGD release scenarios. Note that the no release scenario had a very high probability of refilling by June 1 (close to 98%). Conversely, the 350 MGD scenario results in greater variability for June 1 refill. By June 1, approximately 80% of the ensemble members refilled the Catskill System to at least 95% usable storage. Furthermore, the lowest 10th percentile showed usable storage refilling to between 65% and 82%. OST and ensemble forecasts allowed DEP to quantitatively examine the efficacy of releasing water for turbidity control and, at the same time, examine the trade-offs of the releases on water supply reliability.

3.4 Conservation Releases

The upper reaches of the Delaware River Basin, from Hancock, NY, to Trenton, NJ, are classified as Special Protection Waters, and most of the areas in the reach are included in the National Wild and Scenic Rivers System (Delaware River Basin Commission 2015). These reaches boast exceptional fisheries habitat as well as populations of the endangered dwarf wedge mussel (Gong et al. 2010). Coldwater releases from New York City’s Delaware Basin reservoirs are beneficial in preserving these habitats (Fig. 28). With the implementation of ensemble hydrologic forecasts and OST, DEP is better able to contribute to this objective while maintaining water supply reliability for New York City.

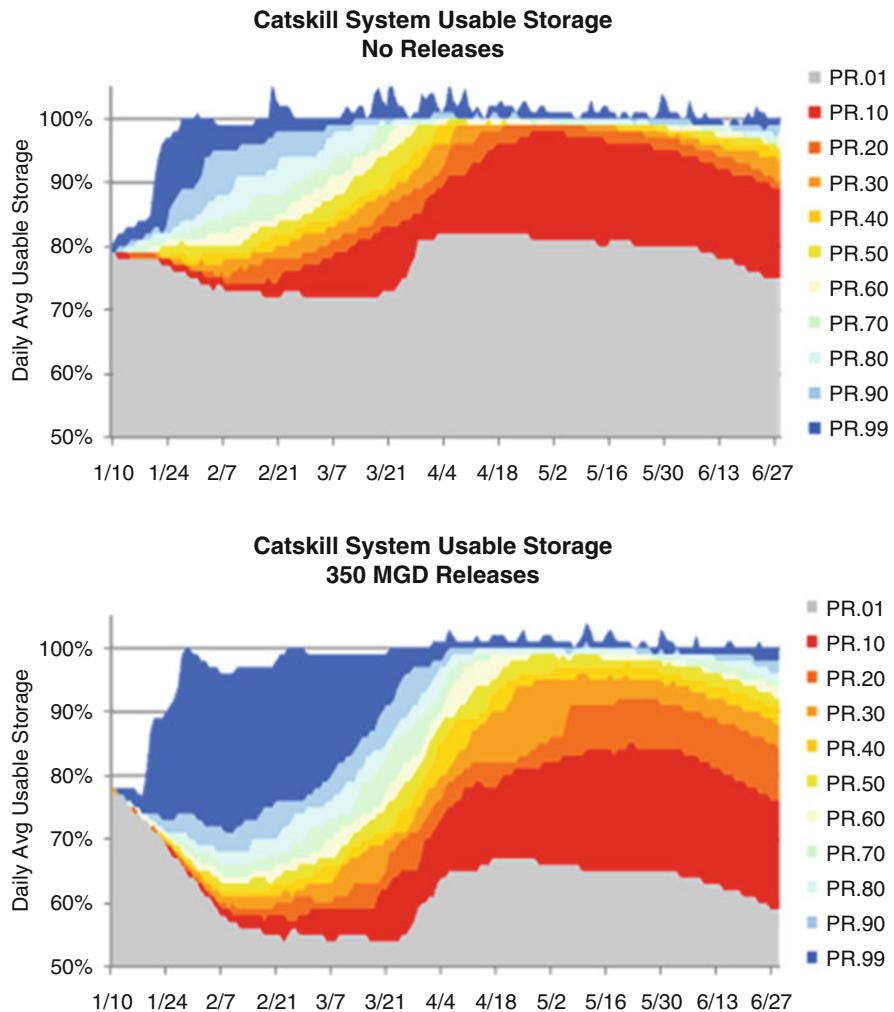


Fig. 27 Impacts of West Basin releases on Catskill System usable storage

Historically, releases from Cannonsville, Pepacton, and Neversink have been based on storage zone and a seasonally varying release schedule. Storage zones in the reservoirs are based on guide curves relating combined usable storage and time of year (Fig. 29). An example release schedule from the 2016 FFMP is presented in Table 1. Under this framework, releases would be determined by using the current storage zone and date to look up the corresponding release rate from the schedule. This release policy is reactive as it is entirely based on current reservoir conditions.

With the implementation of ensemble hydrologic forecasts and OST, a more proactive release policy can be pursued. Beginning with the 2011 revision of the FFMP agreement, New York City agreed to make voluntary augmented releases



Fig. 28 Extent and protection level of the Upper Delaware cold water habitat

using OST-determined forecasted available water (FAW) (Decree Parties to the 1954 U.S. Supreme Court 2016). The revised FFMP expanded on the previous release policy by adding seven release schedules similar in structure to the example

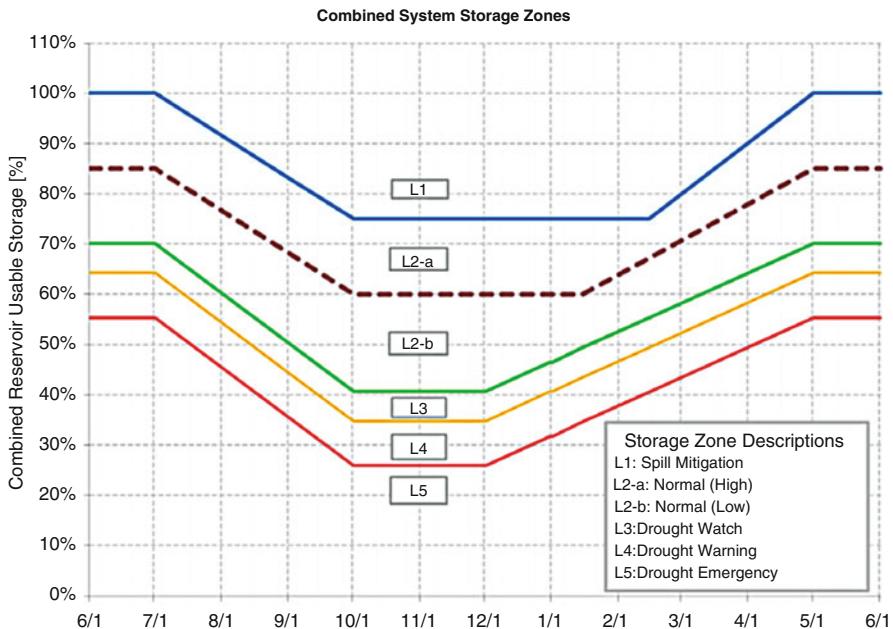


Fig. 29 Guide curves designating storage zone in Cannonsville, Pepacton, and Neversink

displayed in Table 1. Based on a forecast-informed mass balance, DEP selects a schedule with higher or lower releases as appropriate. The release schedule selection process adheres to the following steps:

- Mass balance determining forecast available release volume accumulated to June 1
- Determining the forecast daily average release rate
- Determining the current storage zone for each reservoir
- Matching the forecast average release rate to the closest release schedule

The forecast available water mass balance proceeds as follows (Table 2):

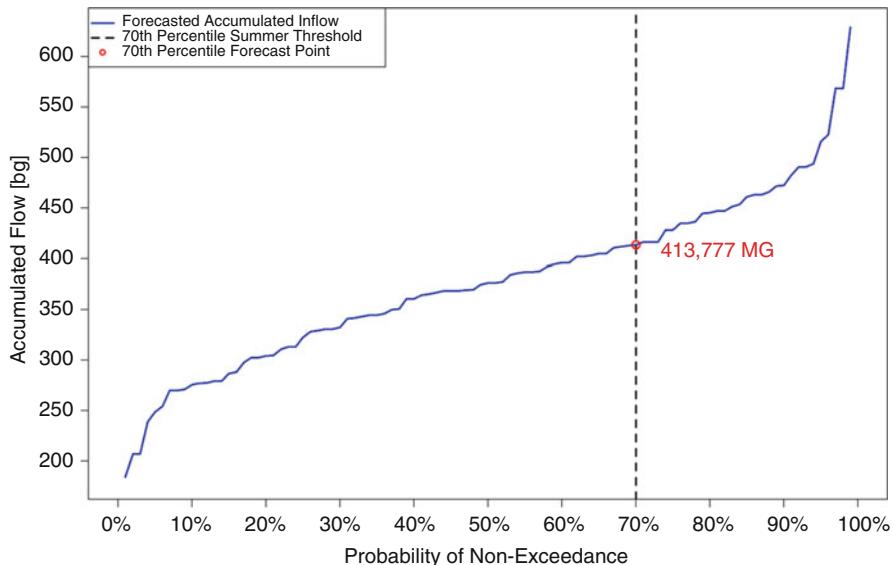
Note that the forecasted cumulative PCN inflow through June 1 is a singular value derived from the HEFS ensemble forecasts. Prior to OST execution, an automated process accumulates each ensemble member for PCN to the next June 1. After the ensemble members are accumulated, percentiles of the distribution are calculated at 5% increments between 5% and 95%. Figure 30 shows a graphical example of the 70th percentile accumulated to June 1 forecast from an August 1 forecast date. The forecast percentiles represent a surrogate for DEP's hydrologic risk. In our example in Fig. 30, there is a 70% probability that accumulated flow to June 1 will be less than 414 BG. Generally, as the forecast non-exceedance percentile increases in the mass balance, releases increase as well.

Table 1 FFMP release schedule 4a. Values in CFS

	Winter			Spring			Summer			Fall		
	Dec 1–31 Mar	Apr 1–30 Apr	May 1–20 May	May 21–31 May	Jun	Jun 1–15 Jun	Jun 16–30 Jun	Jul 1–31 Aug	Sep 1–15 Sep	Sep 16–30 Sep	Sep 1–30 Nov	
Cannonsville strong zone												
L1-a	1500	1500	*	*	*	*	1500	1500	1500	1500	1500	
L1-b	400	400	*	*	*	*	400	400	400	400	400	
L1-c	110	110	200	250	275	275	275	275	175	110	110	
L2-a	75	75	150	200	225	225	225	225	150	75	75	
L2-b	60	60	135	175	190	190	190	190	135	60	60	
Pepacton strong zone												
L1-a	700	700	*	*	*	*	700	700	700	700	700	
L1-b	300	300	*	*	*	*	300	300	300	300	300	
L1-c	85	85	110	130	150	150	150	150	100	85	85	
L2	50	50	75	90	100	100	100	100	60	50	50	
Neversink strong zone												
L1-a	190	190	*	*	*	*	190	190	190	190	190	
L1-b	125	110	*	*	*	*	150	150	150	150	125	
L1-c	65	65	85	100	110	110	110	100	75	65	65	
L2	35	35	55	65	75	75	75	65	50	35	35	

Table 2 Forecast available water mass balance

	Total Pepacton, Cannonsville, Neversink (PCN) storage
+	Forecasted cumulative PCN inflow through June 1
-	Assumed cumulative PCN diversions through June 1
-	June 1 storage target (full)
=	Cumulative PCN release target through June 1

**Fig. 30** Distribution of forecasted accumulated PCN inflow from August 1, 2007 to June 1, 2008

DEP's baseline operating rules use a high forecast percentile in the summer and gradually transition to a more conservative percentile as the year moves toward June 1. These baseline percentiles are shown in Table 3. Depending on the model performance – for example, if probability of June 1 refill is negatively impacted – operations staff may adjust the forecast percentile to compensate.

After the June 1 release target is determined, a daily average release is determined by dividing it by the number of days remaining to June 1. This daily average target is used along with the current storage zone to select the closest FFMP release schedule.

Figure 31 on the following page walks through an example from an OST run on June 1, 2016. Section A shows the FAW mass balance, using the 70th percentile forecast. Section B distributes the total release volume to June 1 (208,916 MG) over 366 days to the next June 1, which yields 571 MGD or 833 CFS. Section C shows the current storage zone in each of NYC's Delaware Basin reservoirs; at 97% or above, each reservoir is in zone L2 (see Fig. 29). Finally, Section D performs the FFMP schedule matching procedure. Note that all of the release rates in the June 1–June 15 period (see Table 1) are returned for each release schedule (4a–4g) at the

Table 3 Baseline forecast percentiles for FAW mass balance

Time of year	Forecast percentile(%)
June– August	70
September–November	50
December–April	30
May	20

**OST-FFMP General Release Summary**

Decision Day: 6/1/2016

General Release Mass Balance	A																																																				
Combined Pepacton, Cannonsville, and Neversink (PCN) Storage: + PCN Inflow Forecast Accumulated to Jun 1: 265,393 MG -Expected PCN Diversion Accumulated to Jun 1: 443,755 MG -Jun 1 Storage Target: 229,363 MG = Available Release Quantity Accumulated to Jun 1: 270,870 MG = Available Release Quantity Accumulated to Jun 1: 208,916 MG																																																					
Available Release Quantity Eventy Distributed to June1																																																					
Available Release Quantity Accumulated to Jun 1: 208,916 MG / Number of Days to Distribute Release Quantity: 366 days Current PCN Release Target: 571 mgd Current PCN Release Target: 883 cfs																																																					
Current Storage Zone for Schedule Selection																																																					
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th style="text-align: left;"></th> <th style="text-align: center;">Usable Storage</th> <th style="text-align: center;">Usable Storage + Snow Storage</th> <th style="text-align: center;">Zone</th> </tr> </thead> <tbody> <tr> <td style="text-align: left;">PCN</td> <td style="text-align: center;">98.0%</td> <td style="text-align: center;">*</td> <td style="text-align: center;">L2</td> </tr> <tr> <td style="text-align: left;">Pepacton</td> <td style="text-align: center;">98.5%</td> <td style="text-align: center;">*</td> <td style="text-align: center;">L2</td> </tr> <tr> <td style="text-align: left;">Cannonsville</td> <td style="text-align: center;">97.3%</td> <td style="text-align: center;">*</td> <td style="text-align: center;">L2-a</td> </tr> <tr> <td style="text-align: left;">Neversink</td> <td style="text-align: center;">97.7%</td> <td style="text-align: center;">*</td> <td style="text-align: center;">L2</td> </tr> </tbody> </table> <p style="text-align: center;">*Not applicable (snow sgtorage is included in the forecast)</p>			Usable Storage	Usable Storage + Snow Storage	Zone	PCN	98.0%	*	L2	Pepacton	98.5%	*	L2	Cannonsville	97.3%	*	L2-a	Neversink	97.7%	*	L2																																
	Usable Storage	Usable Storage + Snow Storage	Zone																																																		
PCN	98.0%	*	L2																																																		
Pepacton	98.5%	*	L2																																																		
Cannonsville	97.3%	*	L2-a																																																		
Neversink	97.7%	*	L2																																																		
Use Release Target and Storage Zone to Select OST-FFMP Release Schedule																																																					
<table border="1" style="width: 100%; border-collapse: collapse;"> <thead> <tr> <th rowspan="2" style="text-align: left; vertical-align: bottom;">OST-FFMP Schedule</th> <th style="text-align: center;">Pepacton</th> <th colspan="3" style="text-align: center;">L2 Storage Zone, Summer Season (cfs)</th> </tr> <tr> <th style="text-align: center;">L2</th> <th style="text-align: center;">L2-a</th> <th style="text-align: center;">Cannonsville</th> <th style="text-align: center;">Neversink</th> <th style="text-align: center;">PCN</th> </tr> </thead> <tbody> <tr> <td style="text-align: left;">Table 4a</td> <td style="text-align: center;">100</td> <td style="text-align: center;">225</td> <td style="text-align: center;">75</td> <td style="text-align: center;">400</td> <td style="text-align: center;">Total</td> </tr> <tr> <td style="text-align: left;">Table 4b</td> <td style="text-align: center;">110</td> <td style="text-align: center;">245</td> <td style="text-align: center;">80</td> <td style="text-align: center;">435</td> <td></td> </tr> <tr> <td style="text-align: left;">Table 4c</td> <td style="text-align: center;">125</td> <td style="text-align: center;">275</td> <td style="text-align: center;">90</td> <td style="text-align: center;">490</td> <td></td> </tr> <tr> <td style="text-align: left;">Table 4d</td> <td style="text-align: center;">140</td> <td style="text-align: center;">325</td> <td style="text-align: center;">100</td> <td style="text-align: center;">565</td> <td></td> </tr> <tr> <td style="text-align: left;">Table 4e</td> <td style="text-align: center;">140</td> <td style="text-align: center;">400</td> <td style="text-align: center;">100</td> <td style="text-align: center;">640</td> <td></td> </tr> <tr> <td style="text-align: left;">Table 4f</td> <td style="text-align: center;">140</td> <td style="text-align: center;">500</td> <td style="text-align: center;">110</td> <td style="text-align: center;">750</td> <td></td> </tr> <tr> <td style="text-align: left;">Table 4g</td> <td style="text-align: center;">140</td> <td style="text-align: center;">500</td> <td style="text-align: center;">110</td> <td style="text-align: center;">750</td> <td></td> </tr> </tbody> </table>		OST-FFMP Schedule	Pepacton	L2 Storage Zone, Summer Season (cfs)			L2	L2-a	Cannonsville	Neversink	PCN	Table 4a	100	225	75	400	Total	Table 4b	110	245	80	435		Table 4c	125	275	90	490		Table 4d	140	325	100	565		Table 4e	140	400	100	640		Table 4f	140	500	110	750		Table 4g	140	500	110	750	
OST-FFMP Schedule	Pepacton		L2 Storage Zone, Summer Season (cfs)																																																		
	L2	L2-a	Cannonsville	Neversink	PCN																																																
Table 4a	100	225	75	400	Total																																																
Table 4b	110	245	80	435																																																	
Table 4c	125	275	90	490																																																	
Table 4d	140	325	100	565																																																	
Table 4e	140	400	100	640																																																	
Table 4f	140	500	110	750																																																	
Table 4g	140	500	110	750																																																	

Selected Schedule: Table(s) 4f/4g^j^jRelease rates are identical for the current storage zone**Fig. 31** Sample FFMP summary page for June 1, 2016

L2 storage zone. The schedule that most closely matches the average release target from Section B is selected. In this case, the maximum release schedule is selected as 750 CFS is the closest to the 833 CFS target.

DEP typically reevaluates the Delaware conservation release schedule on at least a weekly basis and provides public updates on any schedule changes.

4 Conclusion

Streamflow forecasts are critical for operational decision-making by NYC water supply managers. However, any forecast includes one or more types of uncertainty, and traditional deterministic streamflow forecasts do not provide any information on uncertainty. Lack of knowledge regarding the nature, magnitude, and timing of uncertainty limits the usefulness of such forecasts. Ensemble forecasts quantify and express some of this uncertainty, providing decision-makers with information on range and probability of potential future conditions. This allows managers to make quantitative, risk-informed decisions, potentially taking more assertive management actions than could be taken in the absence of such information. These actions can result in enhanced system performance. Ensemble forecasts thus have the potential to improve system operations, but they also complicate the decision-making process because decision-makers must interpret and apply the ensemble information in a context-sensitive manner specific to the decision at hand.

References

- E.A. Anderson, N.H. Crawford, *The Synthesis of Continuous Snowmelt Runoff Hydrographs on a Digital Computer* (Department of Civil Engineering, Stanford University, Stanford, 1964)
- R. Burnash, R. Ferral, R. McGuire, *A Generalized Streamflow Simulation System, Conceptual Modeling for Digital Computers* (California-Nevada River Forecast Center, Sacramento, 1973)
- T. Cole, S. Wells, *CE-QUAL-W2: 2: A Two-Dimensional, Laterally Averaged, Hydrodynamic and Water Quality Model* (U.S. Army Engineering and Research Development, Vicksburg, 2006)
- G. Day, Extended streamflow prediction using NWSRFS. *J. Water Resour. Plan. Manag.* **111**, 157–170 (1985)
- Decree Parties to the 1954 U.S. Supreme Court, *Flexible Flow Management Program* (Decree Parties to the 1954 U.S. Supreme Court Decree, Trenton, 2016)
- Delaware River Basin Commission, *Water Quality Programs of the Delaware River Basin Commission* (Delaware River Basin Commission, West Trenton, 2015)
- J. Demargne, L. Wu, S. Regonda, J. Brown, H. Lee, M. He, . . . Y. Zhu, The science of NOAA's operational hydrological ensemble forecast service. *Bull. Am. Meteorol. Soc.* 79–98 (2014)
- R. Gelda, S. Effler, Modeling turbidity in a water supply reservoir: advancements and issues. *J. Environ. Eng. Div.* **133**, 139–148 (2007a)
- R. Gelda, S. Effler, Simulation of operations and water quality performance of reservoir multilevel intake configurations. *J. Water Resour. Plan. Manag.* **133**, 78–86 (2007b)
- R. Gelda, S. Effler, Testing and application of a two-dimensional hydrothermal model for a water supply reservoir: implications of sedimentation. *J. Environ. Eng. Sci.* **6**, 73–84 (2007c)

- G. Gong, L. Wang, L. Condon, A. Shearman, U. Lall, A simple framework for incorporating seasonal streamflow forecasts into existing water resource management practices. *J. Am. Water Resour. Assoc.* **46**(3), 574–585 (2010)
- R. Hirsch, Stochastic hydrology model for drought management. *J. Water Resour. Plan. Manage.* **107**, 303–313 (1981)
- New York City Department of Environmental Protection, *History of New York City's Water Supply System* (2016), Retrieved from New York City Department of Environmental Protection: http://www.nyc.gov/html/dep/html/drinking_water/history.shtml
- New York Codes, Rules, and Regulations, *6 NYCRR Part 701. Classifications – Surface Waters and Groundwaters* (New York State Department of Environmental Conservation, Albany, 1991)
- New York State Department of Environmental Conservation/New York City Department of Environmental Protection (DEC/DEP), *New York State Department of Environmental Conservation/New York City Department of Environmental Protection (DEC/DEP) Interim Ashokan Release Protocol*. DEC/DEP (2011)
- NYCDEP, Refinement of New York City Water Demand Projections. Prepared by Hazen and Sawyer (2014)
- NYCDEP, *History of Drought and Water Consumption* (2016), Retrieved from New York City Department of Environmental Protection: http://www.nyc.gov/html/dep/html/drinking_water/droughthist.shtml
- NYSDEC, *The Lower Hudson River Basin Waterbody Inventory and Priority Waterbodies List* (Division of Water, Bureau of Watershed Assessment and Management, Albany, 2008)
- Office of the Delaware River Master, *Historical Background* (2015), Retrieved from Office of the Delaware River Master: <http://water.usgs.gov/osw/odrm/intro.html#background>
- R. Seager, N. Pederson, Y. Kushnir, J. Nakamura, The 1960s drought and the subsequent shift to a wetter climate in the Catskill Mountains region of the New York City watershed. *J. Clim.* **25**, 6721–6742 (2012)
- W. Sittner, C. Schauss, J. Monro, Continuous hydrograph synthesis with an API-type hydrologic model. *Water Resour. Res.* **5**, 1007–1022 (1969)
- W.J. Weiss, G. Pyke, W. Becker, D. Sheer, R. Gelda, P. Rush, T. Johnstone, Integrated water quality modeling to support long-term planning. *J. Am. Water Works Assoc.* **105**, 217–228 (2013)
- L. Zhao, Q. Duan, J. Schaake, A. Ye, J. Xia, A hydrologic post-processor for ensemble streamflow predictions. *Adv. Geosci.* **29**, 51–59 (2011)



Probabilistic Shipping Forecast

Dennis Meißner and Bastian Klein

Contents

1	Introduction: Inland Waterway Transport in Europe	1372
2	Hydrological Impacts on Shipping	1373
3	Hydrological Forecasts for Inland Waterway Transport	1376
3.1	Utilization and User Requirements	1376
3.2	Components of Hydrological Forecasting Systems for Shipping	1378
3.3	Demonstrating the Added Value of Probabilistic Forecasts for Navigation	1379
4	Outlook	1381
5	Conclusion	1383
	References	1384

Abstract

Inland waterway transport is an important even though often neglected economic sector relying on hydrological forecasts in order to increase its operating efficiency. Besides river ice and floods, low stream flow is the main hydrological impact for shipping along the European inland waterways as it sustainably affects the load capacity of vessels and thus transportation costs several times a year. For this reason, inland waterway transport benefits from water-level forecasts in order to take preventive action and adjust the draft, especially during stream flow droughts.

Although most navigation-related water-level forecasts are still deterministic, the waterway transport sector is a well-suited customer of probabilistic forecast products for several reasons: The number of decisions to be taken is quite high, especially in

D. Meißner (✉)

German Federal Institute of Hydrology (BfG), Koblenz, Germany
e-mail: meissner@bafg.de

B. Klein

Department Water Balance, Forecasting and Predictions, Federal Institute of Hydrology (BfG),
Koblenz, Germany
e-mail: klein@bafg.de

comparison to the operation of protection measures against rare flood events. Furthermore, in waterway transport, the user's costs and losses associated with possible forecast-based decisions are well known and monetary valuation of losses, like nonoperation times or additional effort due to lighterage, is more feasible as it is for example with regard to human lives or environmental pollution. Last but not least shipping is an inhomogeneous stakeholder as different vessel types and routes cause different cost structures and sensitivities due to navigation conditions. Selecting one "best-guess" forecast, being optimal for all users, is impossible.

In this chapter, hydrological forecasts as one component to support inland waterway transport are presented and the added value of probabilistic forecasts is demonstrated applying a simulation based cost model for the River Rhine, being one of the world's most frequented inland waterways.

Keywords

Cost structure model · Flood · Inland waterway · Inland waterway transport · Low stream flow · Navigation · River ice · River Information Service (RIS) · River Rhine · Seasonal forecast · Shipping · Traffic · Water-level forecast

1 Introduction: Inland Waterway Transport in Europe

Inland navigation has a long history of providing safe and environment-friendly transport. As one of the modes of transport – besides road, railway and air – numerous types of cargo are carried by inland waterway vessels. The European inland waterways offer a more than 40,000 km network of canals, rivers, and lakes connecting cities and industrial regions across the continent. Some 18 out of the 28 European member states have inland waterways, most of them being part of interconnected waterway networks (European Union 2013). The European waterway network is particularly dense in the north-western part of the continent where the large waterways (especially Rhine, Danube, Elbe) in combination with their tributaries and canals enables inland shipping to reach many destinations and, for example, to travel from the North to the Black Sea. All important industrial areas in Western Europe as well as the major sea ports (like Rotterdam, Antwerp, Hamburg) are accessible to inland shipping vessels (Fig. 1).

In light of continuing transport growth within the European Union – freight transport performance is expected to grow up to 32% by 2020 and up to 50% by 2030 (Petersen et al. 2009) – there is a need to use the free capacity inland navigation offers more consequently in order to release the overloaded road and railway networks. As currently freight transport by inland waterway accounts for approximately 5% of all freight transport in the European Union, it is a political objective to promote and strengthen the position of inland waterways within the overall European transportation network in order to increase its modal share (European Union 2013). One main disadvantage of inland waterway transport is the comparatively long time of travel. A trip from Rotterdam to Basel by ship takes approximately 5 days, whereas the travel times of the competing transport modes are just a



Fig. 1 Important inland waterways in Central Europe (Modified source: German Federal Ministry of Transport and Digital Infrastructure, Fachstelle für Geoinformation Süd, Regensburg, Germany)

small fraction of these. Therefore, inland waterway transport is particularly suitable for the transport of large quantities of cargo, especially bulk cargo and containers. A standard inland shipping vessel with a length of 135 m can carry up to 3,800 t, which is equal to nearly 150 lorries.

As with any transportation system, reliability and availability is the main driver to use inland waterway transport as traffic carrier. Unscheduled or excessive scheduled delays are one of the key factors for determining reliability of the transport system (International navigation association 2002). Although the inland waterways offer a nearly congestion-free network (sometimes standby time occurs due passing sluices), the ease, safety, and efficiency of inland waterway transport are sensitive to hydrological impacts. Therefore inland waterway transport is one of the economic sectors, like hydro power or agriculture, relying on hydrometeorological and hydrological forecasts in its short- to long-term business planning and decision making (Moser et al., 2012).

2 Hydrological Impacts on Shipping

The main hydrological hazards concerning inland waterway transport in Europe are low stream flows, floods, and river ice (Nilson et al. 2012). While river ice and floods limit the operational availability of waterways due to closing affected stretches, low stream flows do not cause abandoning of navigation, but it leads to restrictions of

load carrying capacity and as a consequence to higher operation costs. The impacts are of different relevance depending on climate conditions as well as characteristics of the waterway. River ice primarily occurs in waterways with low or nearly no flow velocities (canals, impounded rivers) in areas with low air temperature over longer periods (like Scandinavia and Eastern Europe). Besides blocking the waterway and interrupting its trafficability, ice run can damage vessels and harm technical structures, like weirs, locks, or harbor facilities (Ashton 1986; Carstensen 2008). Besides this purely economic impact, river ice might cause so-called ice jam floods by restraining the discharge which causes the water-level of the jam to rise significantly (Beltaos 1995). Therefore, river ice forecasts, indicating affected stretches as well as estimating ice thickness, are a valuable information in order to coordinate the operation of icebreakers (e.g., pooling the vessels at hot-spots), trying to clear the waterway as long as possible, as well as to take into account limitations of waterway availability (e.g., shifting transport to another mode). The fact that rive ice has a significant impact on inland navigation is indicated by Fig. 2 (top part), showing the number of days with suspension of navigation due to river ice at the waterways Odra along the German-Polish border and Main-Danube canal connecting River Rhine an River Danube and therefore the North with the Black Sea.

In most parts of Europe, river ice is relevant for shipping over a limited period of the year, whereas the water-level – high as well as low – is the hydrological parameter influencing navigation most time of the year. Therefore, this chapter focuses on water-level forecasts for inland waterways. High water-levels affect rivers, regulated as well as free flowing, whereas floods are normally not relevant for canals due to missing natural inflows. Restrictions related to floods depend on the absolute water height as above a given level, river traffic is halted (e.g., Donaukommission 2005). In addition to the protection of the infrastructure the security of navigation is the main motivation, because high flow velocities occurring during floods reduce the maneuverability of the vessels traveling downstream. Additionally the guaranteed clearance below bridges might become too low and limits the possible layer of containers. Therefore, water-levels are not solely relevant for the flood protection community but also for navigational user (Belz et al. 2013). Although the duration and frequency of occurrence of floods is significantly lower than low flows (see Fig. 2, bottom part), floods could cause relevant costs with regard to inland waterway transport. For example, the big floods in 1993 and 1995 caused costs due to failed proceeds of more than 25 million Euros in the international Rhine basin (International Commission for the Hydrology of the Rhine basin, 1999). Figure 2 (bottom part) shows the high variability of restrictions due to high and low stream flows between different years at the station Vienna/Danube using the highest navigable discharges (HSQ) and the (low) flow value that is exceeded in 95% of the days within a reference period (FlowDurationCurve_Q95%) as relevant indicators.

Restrictions caused by low stream flows occur in free flowing waterways, as discharges and water-levels are correlated and the interannual flow regime leads to corresponding water-level conditions. In canals and impounded rivers, the water-level is determined artificially and therefore just indirectly affected by hydro-meteorological drivers. Here, water-levels might be affected during low flows

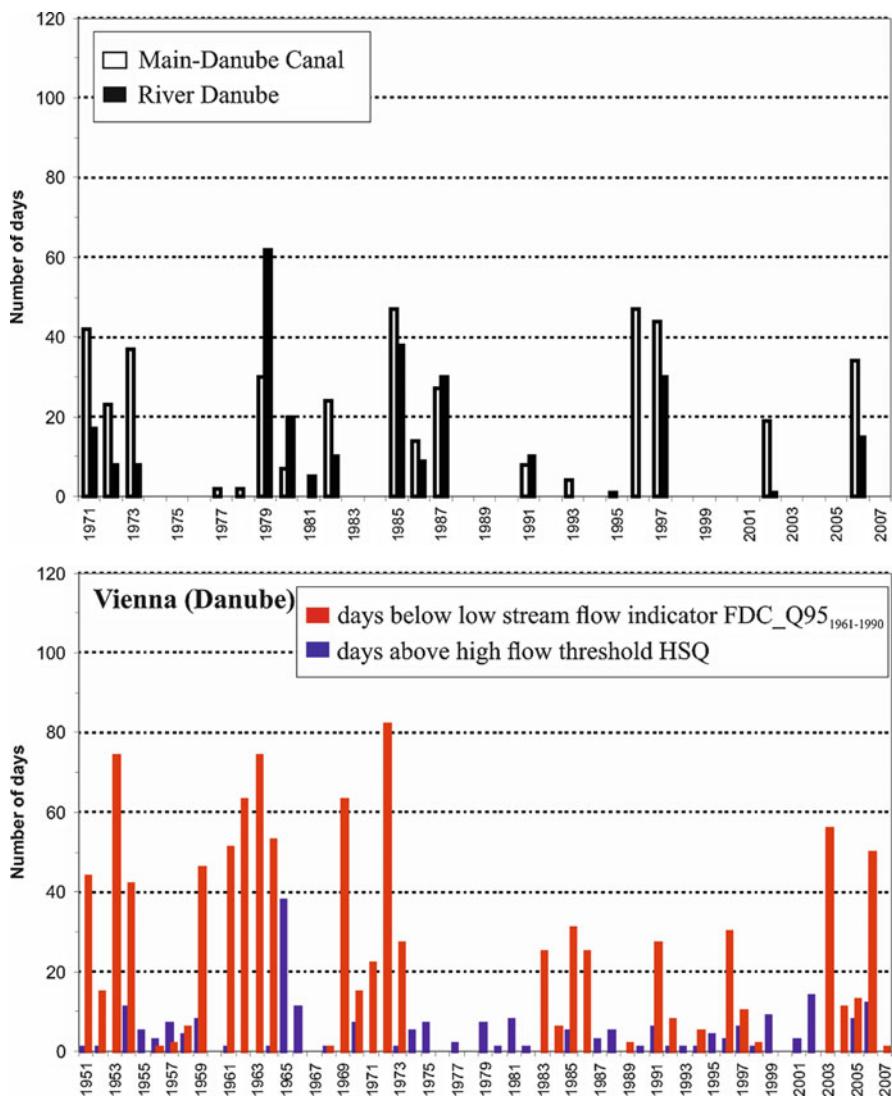


Fig. 2 top part: Suspension of navigation on the Main-Danube Canal and the Odra along the German-Polish border in the period 1971–2007 (left); bottom part: Annual number of days above the high-flow threshold HSQ (blue columns) and below the low-flow indicator FDC_Q95 (red columns) for the station Vienna/Danube in the period 1951–2007

when the operation rules of the canal/weirs do not longer allow for abstraction or retention of water. Unlike floods, there is no threshold beyond which navigation is prohibited due to low stream flows. It is the responsibility of each vessel's skipper to decide whether it is possible to travel within a given section of the waterway despite the reduced water depth. So, it is an individual evaluation of risk (in terms of safety

and cost-effectiveness of the transport) given the intensity of the low flow situation, the ship as well as the cargo type and the destination of the transport. Low water-levels reduce the load factor of inland waterway vessels and thereby increases costs per transported unit (cost per ton). At the same time, the danger of ship-grounding or ship-to-ship collisions increases due to reduced depth and width of the fairway. As low flow situations occur more often than floods and as they are relatively long lasting, they are regarded as the major threat to the reliability of inland waterway transport (see Fig. 2, bottom). The estimated damage in shipping caused by the extreme drought 2003 was, for example, about 91 million for the Rhine basin (Jonkeren et al. 2007).

3 Hydrological Forecasts for Inland Waterway Transport

3.1 Utilization and User Requirements

Water-level as well as river ice forecasts are a fundamental part of many River Information Services (RIS). Those services provide harmonized information in order to support traffic and transport management in inland waterway transport as part of the intermodal transport chain (International navigation association 2002). At the same time, RIS aims at increasing safety of transport and reducing accidents. The main use of hydrological forecasts aiming at inland waterway transport is to optimize the amount of cargo a ship can take before it starts its trip. That is why the water-level, which directly determines the available water depth, is the most important forecast parameter. Figure 3 gives an example on how operational water-level forecast published via RIS are used by the waterway transport sector.

Figure 3 shows the number of accesses per day (black dots) on the forecast for the station Kaub, which is one of the main bottlenecks of the international waterway Rhine (Fig. 3, right), published via the German RIS-component ELWIS (www.elwis.de). It is obvious that decreasing water-levels increase the need for water-level forecast information. Also in case of water-levels tending to the highest navigable level (indicated as “HSW” in Fig. 3), more users are interested than in case of medium water-levels offering sufficient water-depths to fully load the vessels. The link between user demand and economic sensitivity is visible in Fig. 3, too. The transport costs increase with decreasing water-levels as well (blue dots), initially moderate, subsequently exponentially. At high water-levels, large-sized vessels have advantages (economy of scale), which inverses to disadvantages compared to smaller vessels at low fairway depths.

Although the sensitivity of the transport costs differs as a function of the absolute water-level and the vessel type, hydrological shipping forecasts are in demand along the complete range of water-levels as waterway vessels are operated throughout the whole year (despite of suspension time, e.g., due to river ice or floods). Therefore, availability and soundness are basic demands for navigation-related forecasts. This requires a high and stable forecast quality over the complete

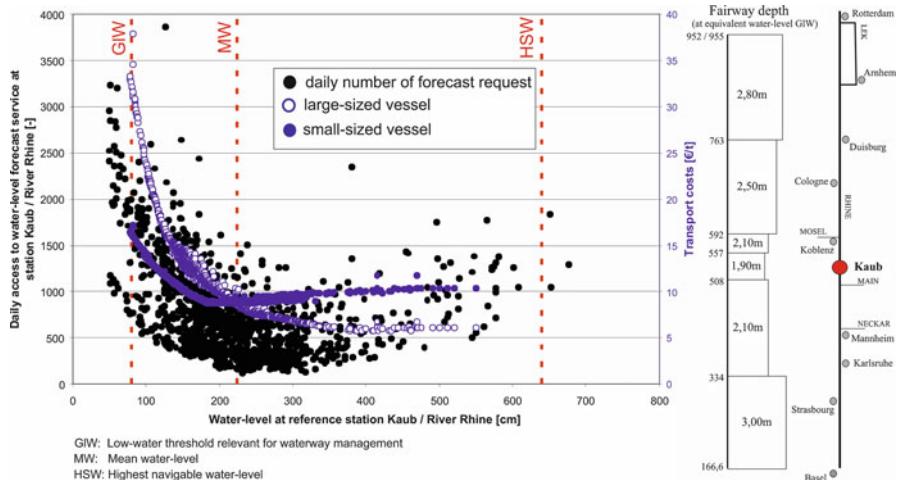


Fig. 3 Left part: Number of daily users of the operational low flow forecast for the River Rhine at station Kaub (black dots) and simulated transport costs (blue dots) both plotted against the absolute water-level, right part: schematic overview of the River Rhine waterway and the fairway depths guaranteed

water-level range. As users in the waterway transport sector have to take a large number of decisions over the year based on hydrological forecasts, they benefit from the “law of large numbers,” which means that single wrong decisions are compensated by a majority of good choices. This aspect is also relevant with respect to probabilistic forecasts. According to probability theory, decisions based on (the expected value of) numerous probabilistic forecasts tend to result in an optima on the long-term, although several single decisions might produce losses.

Besides the absolute height, the timing of forecasted water-levels is also important for transport usage, because the skipper calculate quite exactly when they will reach load limiting locations along their trip, which is the point in time the forecasted water-level is particularly relevant for. That is why the vast majority of waterway user is interested in the shape of the water-level hydrograph instead of just looking at the forecast overtopping or dropping below single threshold levels. As especially during low stream flows the water-levels are sensitive to the height and shape of the river bed, it is important for shipping forecasts to take current bed levels into account, especially within morphodynamically active river stretches, like Elbe or the Lower Rhine.

Forecast lead times should be equal to the travel time of the vessels or at least cover the period the ships need to pass the main bottlenecks of a waterway leaving the lading port. Normally this takes several days. If forecast lead times are too short the forecasts are not completely usable to optimize the load before starting the trip and cargo capacity is wasted or the ships are overloaded and they have to dump cargo on their way leading to additional costs (due to unloading, stocking, further transport via truck or rail, etc.).

The required forecast frequency is relatively low when compared to flood forecasts for example. The water-level dynamic is lower than in case of floods, and the navigational users are not able to adjust the load of their vessels every time a new forecast is published anyway. The skipper has to take his decision on the load carrying capacity at a specific point in time before leaving the port. Most navigation-related forecasts are issued once a day.

3.2 Components of Hydrological Forecasting Systems for Shipping

Forecasting systems for shipping are quite similar to those for flood forecasting or power production. The systems are a cascade of different models which is driven by hydrometeorological forecasts. Precipitation and air temperature are the meteorological variables used most often. Precipitation directly influences discharges and as a consequence water-levels along the rivers. Air temperature is relevant as it influences the snow processes as well as evaporation, but of course all processes related to river ice almost completely depend on the air temperature. These hydrometeorological inputs are transferred into discharges by a hydrological model covering the whole basin of the waterway of interest.

As navigational users are primarily interested in water-levels, a hydrodynamic model, normally one-dimensional, representing the waterway itself with all its relevant structures (e.g., groins, weirs), is part of the forecasting system. Quite often forecast locations relevant for waterway transport are affected by backwater effects or the regulation of structures, which could be represented within the hydrodynamic model adequately. In most cases, the hydrodynamic model is the computational most demanding hydrological component, especially when using ensemble forecasts. The hydrodynamic models require a comparatively high spatial (one grid point every 100–500 m) and temporal resolution (less than 1 h) in order to guarantee numerical stability and reliable results. The hydrodynamic model has to contain current data of the river morphology, as it influences the calculation of the water-level. So, the frequency of model updates is relatively important and comparatively high when compared to hydrological models.

The models within a forecasting system for shipping are not calibrated solely focusing on a specific discharge or water-level range (e.g., floods), but the aim is to find a balanced model-setup giving satisfying forecast results over the whole hydrological range. The quality measures used to evaluate calibration have to allow for these characteristics. For example, care should be taken that single major discrepancies of flood peaks are not overrated in the quality index used. On the other hand, the amount and in most cases also the quality of training data in order to calibrate the models is significantly larger as if the focus is on extreme events only.

3.3 Demonstrating the Added Value of Probabilistic Forecasts for Navigation

Shipping forecasts have value if the users are able to base upon decisions that reduce the transport costs and therefore maximize their profit over the years. There are several methods available to estimate the economic value of meteorological or hydrological forecasts (Katz and Murphy 1997; Roulin 2007). Simulation-based cost modeling is an approach to assess the economic value in an objective way with close relation to practice. It is beyond question that hydrological forecasts are a valuable information for inland waterway transport. The following examples of the application of a cost structure model demonstrate that forecast uncertainty information is able to yield an additional economic benefit for the inland waterway transport sector using the example of the Rive Rhine. The River Rhine is one of the world's most frequented waterways and the backbone of the European inland waterway network. Approximately 600 vessels are passing the Dutch-German border per day.

A cost structure model considers various influences and cost components of inland waterway transport. The relevant expense factors are composed on the one hand of fixed costs, like capital costs for investment and insurance and labor costs depending on ship size and operation mode. On the other hand, the variable operation costs have to be considered, which depend on the navigation conditions, mainly water depth and hydrodynamic properties of the vessels. The model considers the impact of water-levels and stream velocities on load carrying capacity, power demand, speed and the resulting fuel consumption. Of course, the water-level determines the maximum draft and therefore maximum payload, but also fuel consumption and possible speed are influenced significantly. In order to analyze the effect of water-level forecasts on the performance of inland waterway transport, a deterministic forecast and three water-level quantiles (0.25-quantile, 0.5-quantile, 0.75-quantile) of a probabilistic forecast are coupled to a cost structure model. For each forecast variant, cost analysis simulations have been performed for seven representative vessel types traveling on three origin-to-destinations (starting point: the port of Rotterdam) in upstream direction for a period of approximately 4.5 years (1530 forecasts). Within the model each day of the time span a ship of each type is traveling to all of the three destinations as far as the waterway isn't closed due to a flood (Holtmann and Bialonski 2009; Bruinsma et al. 2012).

Based on these results Fig. 4 compares the best probabilistic variant with the deterministic forecast for five vessel types and the three routes of different lengths and waterway characteristic (see Fig. 3, right panel). Negative values indicate lower costs than using the deterministic forecast. The analysis shows that probabilistic forecasts have the potential to improve decisions for waterway transport with relation to costs. Even a cost reduction of just a few percentage points per ton leads to relevant effects due to the large amounts of tons transported per year. It is also visible that the longer the trip and therefore the longer the required lead times, the higher the sensitivity of the hydrological forecast variant becomes. The

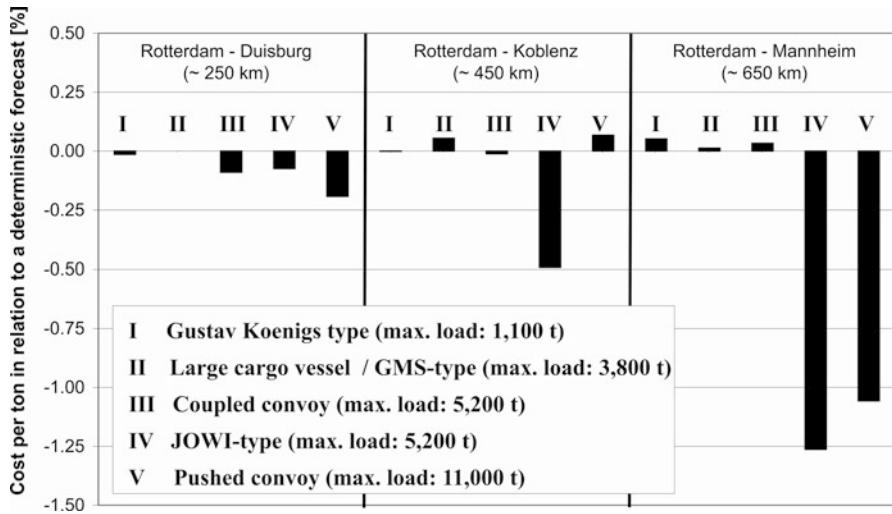


Fig. 4 Cost differences between deterministic and probabilistic forecast for five vessels and three routes based on cost structure modeling for the River Rhine

probabilistic forecasts become more and more advantageous – also in economic terms – with increasing uncertainty. Furthermore, larger vessels show higher potential to profit from probabilistic forecasts.

Figures 5 and 6 illustrate two single forecasts out of the 1530 forecasts analyzed with the cost structure model for the trip from the port of Rotterdam to the industrial region of Rhine – Neckar near Mannheim. The bottleneck near Kaub is passed after 3 days, which is therefore the relevant lead time to determine the load capacity. Figure 5 demonstrates a positive example of the probabilistic forecast, which is defined by the area between the 5%- and 95%-quantiles. The uncertainty range is relatively narrow, and the measured values (black squares) are quite close to the median (blue line). The deterministic forecast (red line) underestimates the water-level from day one onwards, with the consequence that less cargo than possible is carried (up to nearly 150 t for large-sized vessels). The effects on transport costs are depicted in Fig. 5 as well.

As for every probabilistic forecast and also for navigation-related forecasts, it is disadvantageous if the range of forecasted water-levels does not cover the measured values. Figure 6 shows such a forecast bust, as the forecast clearly overestimates the water-level rise at the station Kaub for the coming days. Subsequently skippers tend to overload their vessels and they have to lighter surplus goods or wait until the water-level raises high enough to pass the bottleneck along their trip (although safety margins already included). The commercial consequences depend on the size of the vessel, but on average the additional costs due to lighterage represent more than half of the total transport costs per ton, for smaller ships significantly more.

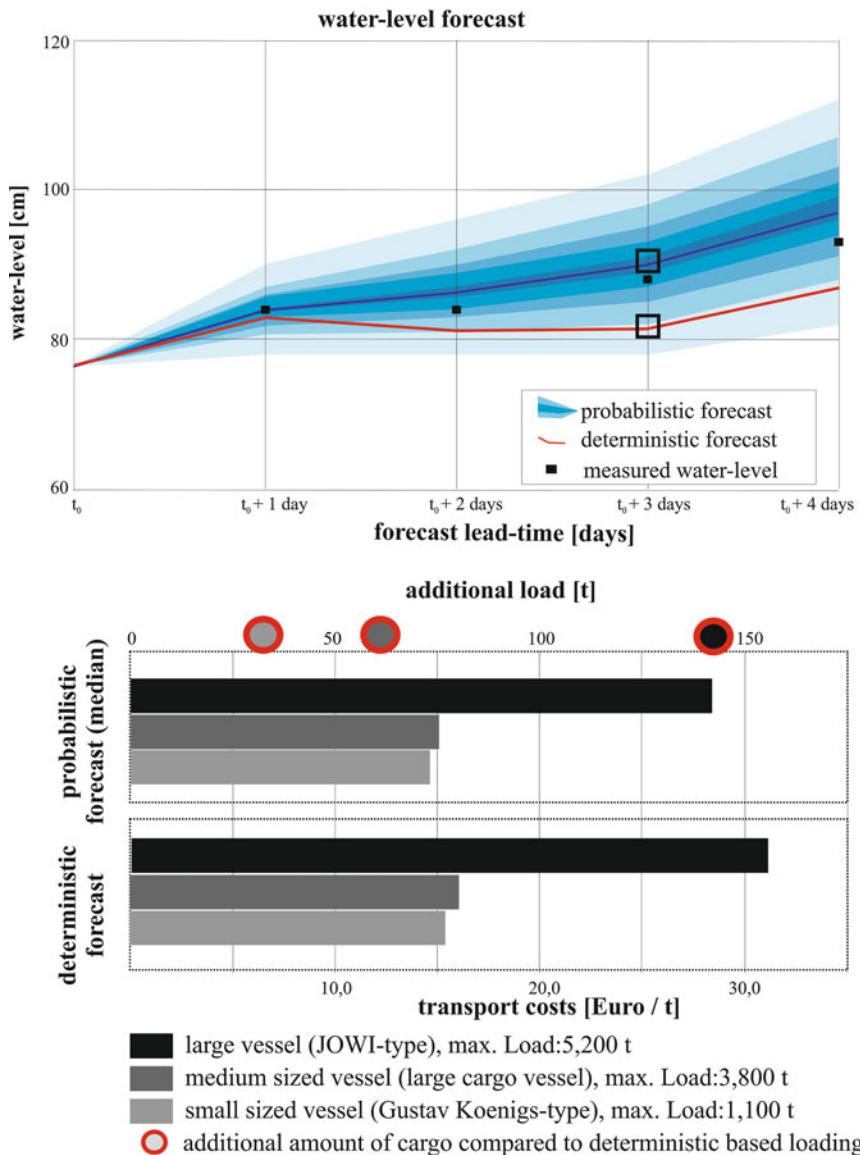


Fig. 5 Example of a successful probabilistic forecast with a tight uncertainty range

4 Outlook

The growth of transport requires added shifting of cargo on the inland waterways with its vast capacities in order to relieve roads and railways. To trap the full potential of inland waterway transport as an environmentally friendly and energy

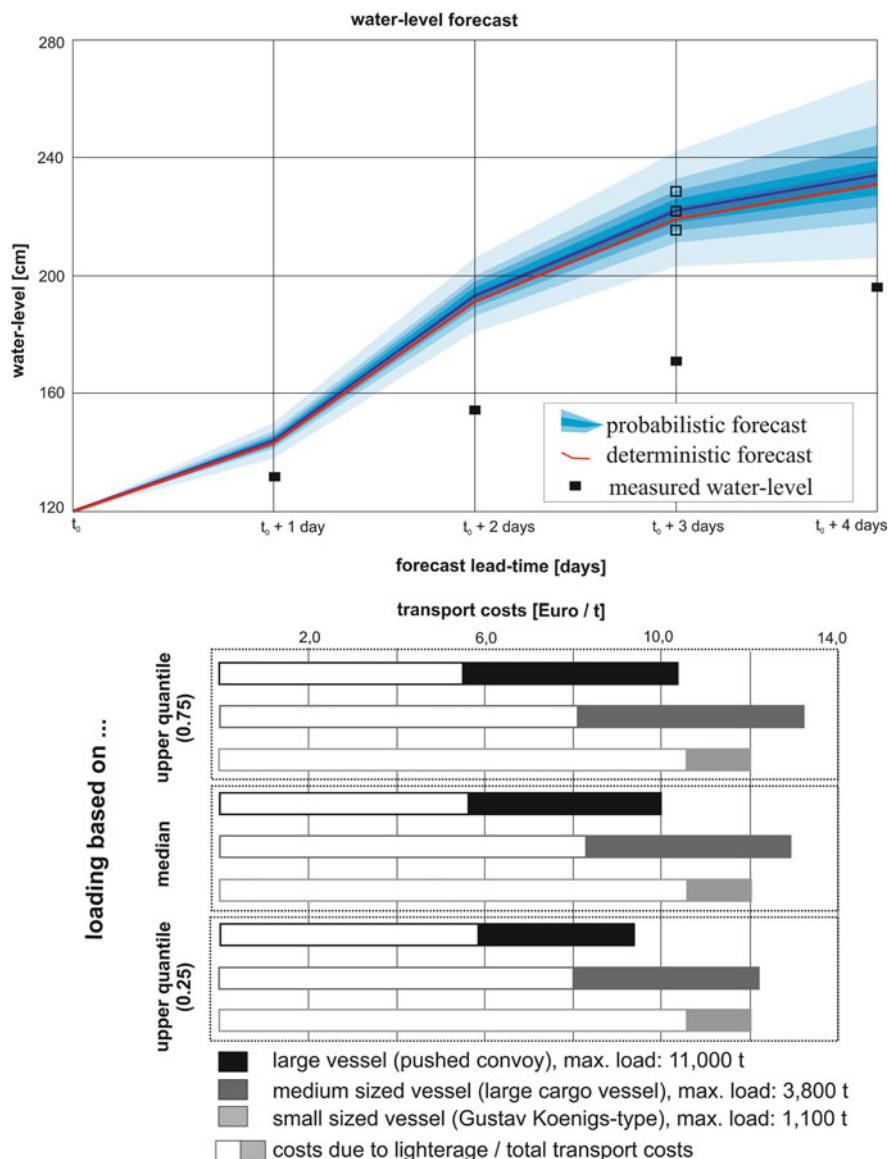


Fig. 6 Example of a forecast bust – the measured water-levels are completely outside the uncertainty range (top), transport costs including charges due to excess load based on different water-level quantiles for three vessel types (bottom)

efficient transport mode, the importance of hydrological shipping forecasts will increase in the future. The effort to incorporate water-level forecasts into River Information Services will go on and have to be extended to probabilistic forecast products as they offer additional benefit to navigational user.

A second challenge is the development and design of medium-range up to seasonal forecast products for waterway transport. Additional lead-time is needed in order to strengthen the inclusion of waterway transport within multi-modal transport chains. Seasonal forecasts with lead times of several months have great potential to become a valuable tool for the optimization of waterborne transport and the medium- to long-term waterway management. Extended lead times offer the possibility to optimize the fleet structure of shippers as well as the stock management of enterprises. By taking into account periods with above- or below-average water-levels for the coming months, the timing of transport could be rescheduled to an earlier or later date or multiple smaller ships could be ordered in times of lower water-levels to execute the transport efficiently. Of course, seasonal forecast have to be probabilistic to be useful and communicated/explained to end-users in an adequate way.

Last but not least the (probabilistic) forecast of additional parameter relevant for navigation along the inland waterways is a matter of development. In addition to optimize the amount of cargo a ship can take before it starts, fuel consumption can be minimized by following the optimal track based on the flow velocity and using the optimal ship speed in relation to the water depth and the desired time of arrival. So, hydrological forecasts for shipping could be used to set up advanced trip advisors, which gives the optimal track and ship speed in addition to the maximum load capacity or container height for which it is possible to pass the critical points on the particular route.

5 Conclusion

Switching from deterministic to probabilistic information is beneficial also for navigation-related water-level forecasts. This is particularly true if longer lead-times up to monthly and seasonal scales are required. But as analyses based on the coupling of probabilistic water-level forecasts with a cost-structure model for the River Rhine demonstrate, probabilistic forecast information could be vital for short- to medium-range forecasts, too. Of course, forecast busts are not excluded, but in the end probabilistic forecasts enable the waterway user to take decisions on an objective and consistent basis according to their own individual sensitivity to the occurrence or nonoccurrence of a hydrological event. Publishing one “best-guess forecast” is out of reach as inland waterway transport covers a wide range of stakeholders with extremely different sensitivities related to navigation condition. The essential transition from deterministic to probabilistic forecasts is a common process of several years for forecast providers and forecast users. And only if we succeed to transform the undisputed theoretical advantage of probabilistic forecasts into practical use, establishing probabilistic shipping forecasts will go beyond being a purely academic exercise.

References

- G.D. Ashton, *River and Lake Ice Engineering* (Water Resources Publications LLC, Littleton, 1986)
- S. Beltaos, *River ice jams* (Water Resources Publications LLC, Littleton, 1995)
- J. Belz, N. Busch, M. Hammer, M. Hatz, P. Krahe, D. Meißner, A. Becker, U. Böhm, A. Gratzki, F.J. Löpmeier, G. Malitz, T. Schmidt, Das Juni-Hochwasser des Jahres 2013 an den Bundeswasserstraßen – Ursachen und Verlauf, Einordnung und fachliche Herausforderungen. Korrespondenz Wasserwirtschaftschaft (2013). <https://doi.org/10.3243/kwe2013.001>
- F. Bruinsma, P. Koster, B. Holtmann, E. van Heumen, M. Beuthe, N. Urbain, B. Jourquin, B. Ubbels, M. Quispel, Consequences of climate change for inland waterway transport. Deliverable 3.3 Consequences of climate change for inland waterway transport (2012). http://www.ecconet.eu/deliverables/ECCONET_D3.3_final.pdf (last access: 11/07/2016)
- D. Carstensen, *Eis im Wasserbau – Theorie, Erscheinungen, Bemessungsgrößen*. Dresdner Wasserbauliche Mitteilungen, vol. 37 (TU Dresden, Dresden, 2008). ISBN 978-3-86780-099-0
- Donaukommission, *Lokale Schiffahrtsregeln auf der Donau (Sonderbestimmungen)* (Budapest, 2005) http://www.danubecommission.org/uploads/doc/publication/Lokale_Schifff_Reg/Lokale_Schiffahrtsregeln.pdf (last access: 11/07/2016)
- European Union, *EU transport in figures – statistical pocketbook 2013* (2013). <https://doi.org/10.2832/19314>
- B. Holtmann, W. Bialonski, *Einfluss von Extremwasserständen auf die Kostenstruktur und Wettbewerbsfähigkeit der Binnenschifffahrt*, ed. by BMVBS Tagungsband KLIWAS – Auswirkungen des Klimawandels auf Wasserstraßen und Schifffahrt in Deutschland. 1. Statuskonferenz am 18. und 19. März 2009 (Bonn, 2009)
- International navigation association, *Vessel traffic and transport management in the inland waterways and modern information systems*. Report of Working Group 24 – INCOM (2002). ISBN 2-87223-124-2
- O. Jonkeren, P. Rietveld, J. van Ommeren, Climate change and inland waterway transport: welfare effects of low water levels on the river Rhine. *J. Transp. Econ. Policy* **41**(3), 387–411 (2007)
- R.W. Katz, A.H. Murphy, *Economic value of weather and climate forecasts* (Cambridge University Press, Cambridge, 1997). ISBN 9780521435710
- H. Moser, J. Cullmann, S. Kofalk, S. Mai, E. Nilson, S. Rösner, P. Becker, A. Gratzki, K.J. Schreiber, in *An integrated climate service for the transboundary river basin and coastal management of Germany*, ed. by WMO. Climate ExChange, (2012), ISBN 978-0-9568561-4-2
- E. Nilson, I. Lingemann, B. Klein, P. Krahe, Impact of hydrological change on navigation conditions. Deliverable 1.4 ECCONET – Effects of climate change on the inland waterway transport network (2012)
- M.S. Petersen, R. Enei, C.O. Hansen, E. Larrea, O. Obisco, C. Sessa, P.M. Timms, A. Ulid, Report on Transport Scenarios with a 20 and 40 year Horizon, Final report TRANSVisions project, (Copenhagen/Denmark, 2009)
- Internationale Kommission für die Hydrologie des Rheingebietes, *Eine Hochwasserperiode im Rheingebiet – Extremereignisse zwischen Dez. 1993 und Febr. 1995* (1999). ISBN 90-70980-28-2
- E. Roulin, Skill and relative economic value of medium-range hydrological ensemble predictions. *Hydrol. Earth Syst. Sci.* **11**, 725–737 (2007)



Probabilistic Inundation Forecasting

A. Mueller, C. Baugh, P. Bates, and Florian Pappenberger

Contents

1	Introduction	1386
2	Methodology	1387
2.1	Modeling Framework	1387
2.2	Numerical Weather Prediction	1387
2.3	Hydrological Modeling	1388
2.4	Hydraulic Modeling	1388
2.5	Forecast Evaluation	1389
3	Event Description	1390
4	Model Framework Settings and Simplifications	1390
5	Results	1392
5.1	Probability Maps	1392
5.2	Forecast Skill Scores	1394
6	Conclusion	1396
	References	1396

A. Mueller (✉)

Geography and Environmental Science Department, University of Reading and European Centre for Medium-Range Forecast (ECMWF), Reading, UK

e-mail: a.mueller@reading.ac.uk

C. Baugh (✉)

European Centre for Medium-Range Forecast (ECMWF), Reading, UK

e-mail: calum.baugh@ecmwf.int

P. Bates (✉)

Department of Geography, School of Geographical Sciences, University of Bristol, Bristol, UK
e-mail: paul.bates@bristol.ac.uk; gppdb@bristol.ac.uk

F. Pappenberger (✉)

European Centre for Medium-Range Weather Forecasts, ECMWF, Reading, UK
e-mail: florian.pappenberger@ecmwf.int

Abstract

Many existing operational hydrological ensemble forecasting systems only produce forecasts of river discharge. It is possible to convert discharge forecasts into inundation extents, in particular because there are well-established tools for the estimation of inundation hazard. The basic components of the modeling framework from which to produce inundation forecasts are: (1) meteorological forcing; (2) a hydrological model; (3) a hydraulic model; and; (4) a methodology to derive probabilistic inundation maps. We perform all those steps using the example of the 2013 River Elbe event. We validate the maps of flooding probability against the observations. We stress the importance of the spatial discretization of the digital elevation maps (DEM) and the influence of the resolution of the flood defense topographic features. This study shows that up to 80% of the flooded area along the Elbe in 2013 could have been forecasted to inundate 7 days in advance, using the probabilistic modeling framework proposed.

Keywords

Probabilistic forecast · Ensemble forecast · Inundation forecast · Forecast skill

1 Introduction

Prior to a forecasted emergency flooding situation, it is important that responders have access to predictions of flood inundation extent (Wetterhall et al. 2013; Dale et al. 2014). Many existing operational hydrological ensemble forecasting systems only produce forecasts of river discharge (Croke et al. 2009; Pappenberger et al. 2014; Voisin et al. 2011; Pappenberger et al. 2011; Croke and Pappenberger 2009); therefore, it is necessary to translate these into forecasts of flood inundation extent. Pappenberger et al. (2005) and Schumann et al. (2013) illustrate that it is possible to convert discharge forecasts into inundation extents, in particular because there are well-established tools for the estimation of inundation hazard (Pappenberger et al. 2012). Therefore, the basic components of the modeling framework from which to produce inundation forecasts are: (1) meteorological forcing, (2) a hydrological model, (3) a hydraulic model, and (4) a methodology to derive probabilistic inundation maps (see Fig. 1). However, floods are by definition extreme events and establishing prediction skill in such cases is notoriously difficult due to the extreme difference against normal conditions. Hence, the evaluation of skill can be undertaken by an assessment of case studies to establish the suitability and usability of such forecasts (Alfieri et al. 2013; Pappenberger et al. 2011). This study evaluates the skill of a probabilistic flood inundation forecast for the June 2013 floods in Central Europe using meteorological forcing from the ECMWF ensemble system, the operational set-up of the European Flood Awareness System (EFAS) to forecast discharges, and the hydraulic model LISFLOOD-FP for inundation predictions.

2 Methodology

2.1 Modeling Framework

The modeling framework in this study, as described above, consists of four components (Fig. 1): (1) meteorological forcing from a numerical weather prediction (NWP) system to provide forecasts of variables such as precipitation, (2) a hydrological model to route precipitation through the catchment and river network providing discharge forecasts at a given river location, (3) a hydraulic model to route the forecasted discharges through the river and floodplain. In this study, the meteorological forcing comes from an ensemble forecasting system; hence, multiple predictions of flood inundation extent will be produced. The fourth modeling component deals with the evaluation and presentation of these probabilistic forecasts.

2.2 Numerical Weather Prediction

Probabilistic meteorological forcing data are taken from the ECMWF Integrated Forecasting System (IFS) ensemble. This is a global forecasting system with one control member and 50 perturbed members according to uncertainties in the initial model conditions and deficiencies in the model. The spatial resolution of this forecasting system is 32 km for the first 10 days lead time, thereafter it is 64 km.

Fig. 1 Schematic procedure for probabilistic flood prediction

Numerical Weather Prediction

Hydrological rainfall-runoff model
(EFAS)

Hydraulic model
Lisflood-FP

Flooding probability map

2.3 Hydrological Modeling

The meteorological data are used to force a hydrological model in order to produce forecasts of river discharge. In this study, the LISFLOOD hydrological model which is part of the operational EFAS suite was used. LISFLOOD is a distributed, hydrological rainfall-runoff model which simulates canopy and surface processes as well as flow and wave routing in the river channel (Van der Knijff et al. 2010). EFAS provides probabilistic flood forecasts and early warnings, based on discharge predictions, to a network of European national hydrological forecasting institutions and civil protection agencies (Thielen et al. 2009). The hydrological model is run at a 5 km resolution; therefore, the meteorological data are downscaled using a nearest neighbor approach.

EFAS has been evaluated in multiple scientific studies and has been proven to provide skillful discharge forecasts for rivers in Europe (Pappenberger et al. 2011; Bogner et al. 2014; Demeritt et al. 2013; Ramos et al. 2013; Alfieri et al. 2012; Bogner et al. 2011).

2.4 Hydraulic Modeling

The LISFLOOD-FP hydraulic model is used to route the EFAS discharge predictions across the floodplain, thus providing forecasts of inundation extent. LISFLOOD-FP simulates the spreading of fluvial or coastal flooding by solving the St. Venant shallow water equations (Bates and de Roo 2000; Bates et al. 2010; Neal et al. 2012) of mass (1) and momentum (2) conservation.

The equations in the horizontal direction x are as follows:

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = 0 \quad (1)$$

$$\frac{\partial Q_x}{\partial t} + \frac{\partial}{\partial x} \left(\frac{Q_x^2}{A} \right) + gA \frac{\partial h}{\partial x} = gA(S_0 - S_f) \quad (2)$$

where: $S_0 = -\frac{\partial z}{\partial x}$ is the bed slope, and $S_f = \frac{Q^2 n^2}{A^2 R^{4/3}}$ is the friction slope. A is the cross-sectional area of the flow, t is time, Q_x is the volumetric flow rate in x direction of the flow, g is the gravity acceleration, h is the depth of the flow, n Manning coefficient of friction, z bed elevation, and R is the hydraulic radius.

In this study, a version of the code called Lisflood-ACC (or acceleration) was used (Bates et al. 2010; de Almeida and Bates 2013). It solves the Eqs. 1 and 2 omitting term $\frac{\partial}{\partial x} \left(\frac{Q^2}{A} \right)$, which is called the convective acceleration. When solved numerically the above equations provide a good simulation of wave propagation when flow is subcritical. The discretization in LISFLOOD-FP is achieved on a regular square (raster) mesh by means of a Finite Difference Method with explicit

time-stepping. The solver uses adaptive time-stepping with the Courant-Friedrichs-Lowy (CFL) number set to 0.7 for stability reasons (de Almeida and Bates 2013).

2.5 Forecast Evaluation

Forecasts of flood extent provided by this model cascade can be evaluated against an observed extent with multiple measures.

The possible outcome of a binary (yes/no – wet/dry) event, such as a flood, can be represented using a contingency table (Mason 2003). There are two ways for the forecast to be correct (correct hit (A) and correct rejection (D)) and two ways to fail (false alarm (B) and missed occurrence (C)) – see Table 1. In this case, each cell of the domain can fall into categories A, B, C, or D. We perform such a classification for each pixel of the domain. The forecast performance for the 2013 flood event is measured by the skill scores detailed in Table 2.

Another way of evaluating the forecasts is the Brier Score which can present the skill of the probabilistic forecast information as a single value. It is defined as the mean squared error of N samples:

$$BS = \frac{1}{N} \sum_{i=1}^N (p_i - O_i)^2 \quad (3)$$

where $p_i \in <0, 1>$ is the probability of the forecast of the observation $O_i = 0$ or 1. Our sample size N is the number of pixels in the model. Each pixel or cell has 51 possible 0/1 states; 0 when pixel is dry and 1 when pixel is wet. This way we get the probability p_i for each pixel separately (51 being the number of ensemble members):

$$p_i = \frac{1}{51} \sum_{j=1}^{51} (\text{pixel}_{\text{wet}} = \text{true}) \quad (4)$$

Table 1 Contingency table for a flood event

		Simulated Wet	Dry
Observed	Wet	A (hit)	C (miss)
	Dry	B (false alarm)	D (correct dry)

Table 2 Forecast verification measures used

Name	Definition
Hit rate (H)	$\frac{A}{A+C}$
False alarm rate (F)	$\frac{B}{B+D}$
Peirce Skill score (PSS)	$\frac{AD-BC}{(B+D)(A+C)} = H - F$
Critical Success Index (CSI)	$\frac{A}{A+B+C}$

3 Event Description

Extreme flooding in Central Europe began after several days of heavy rain in late May and early June 2013. Prior to the torrential rains, the SMOS (Soil Moisture and Ocean Salinity satellite mission) showed that soils in Germany reached record levels of saturation due to the effect of the unusually wet spring (Pappenberger et al. 2013). The areas affected included south and east German states along the Elbe, Danube, and their tributaries, leading to high water and flooding along their banks. The area around the city of Wittenberge has been chosen for this demonstration. The flood wave reached the city on 8 June 2013. The inundation extent data (Radarsat-2, 50 m resolution) were provided by the German Center for Satellite Based Crisis Information (DLR-ZKI) (see Fig. 2a). The dark blue areas represent the normal body of water, while the light blue shows the extent of flooding. The next figure (Fig. 2b) presents the bitmap of the body of water obtained from the polygon supplied by DLR-ZKI. The areas circled in green were removed from the mask before the comparison with the simulation results. They represent a lake and inundation not connected directly with the inundation caused by the river Elbe. The area marked with dashed blue line is a bridge and a road leading to it; these were also removed from the comparison polygon, as they do not appear in 100 m resolution Digital Elevation Model (DEM) which is used to run LISFLOOD-FP. The water flows under the bridge, including it in the simulation would mean that it is a dry area obstructing the flow in the river channel.

4 Model Framework Settings and Simplifications

Meteorological predictions from the 51 member ECMWF IFS ensemble for the forecast on 30 May 2013 were used to force the LISFLOOD hydrological model to produce a 10 day 6-hourly ensemble discharge series (Fig. 3) at the location marked with the red cross in Fig. 2a. The date of this forecast was chosen in order to assess the ability of the model framework to forecast the flood extent with a 1 week lead time.

The LISFLOOD-FP model domain was set to the same extent as in Fig. 2a (lower left corner ETRS-LAEA coordinates 4417000,3308000), it extended 40 km to the east and 24 km to the north. The input data requirements are a DEM of floodplain topography, definitions of channel geometry, a declaration of Manning's n friction parameter values across the domain, and flow input and output boundary conditions. The Shuttle Radar Topography Mission – SRTM DEM was used to define the floodplain topography (Fig. 4a). According to the global assessment of the SRTM mission (Rodriguez et al. 2006), the absolute height error in Eurasia is of the order of 6 m. The mission radar did not penetrate the vegetation canopy, so the DEM contains the vegetation and buildings artifacts. Version 4.1 of the CGIAR-CSI (Consultative Group for International Agricultural Research – Consortium for Spatial Information) SRTM DEM (downloaded from <http://srtm.cgiar.org/>) was used. It was resampled from the WGS84 to the ETRS-LAEA coordinate system at 100 m resolution. Channel

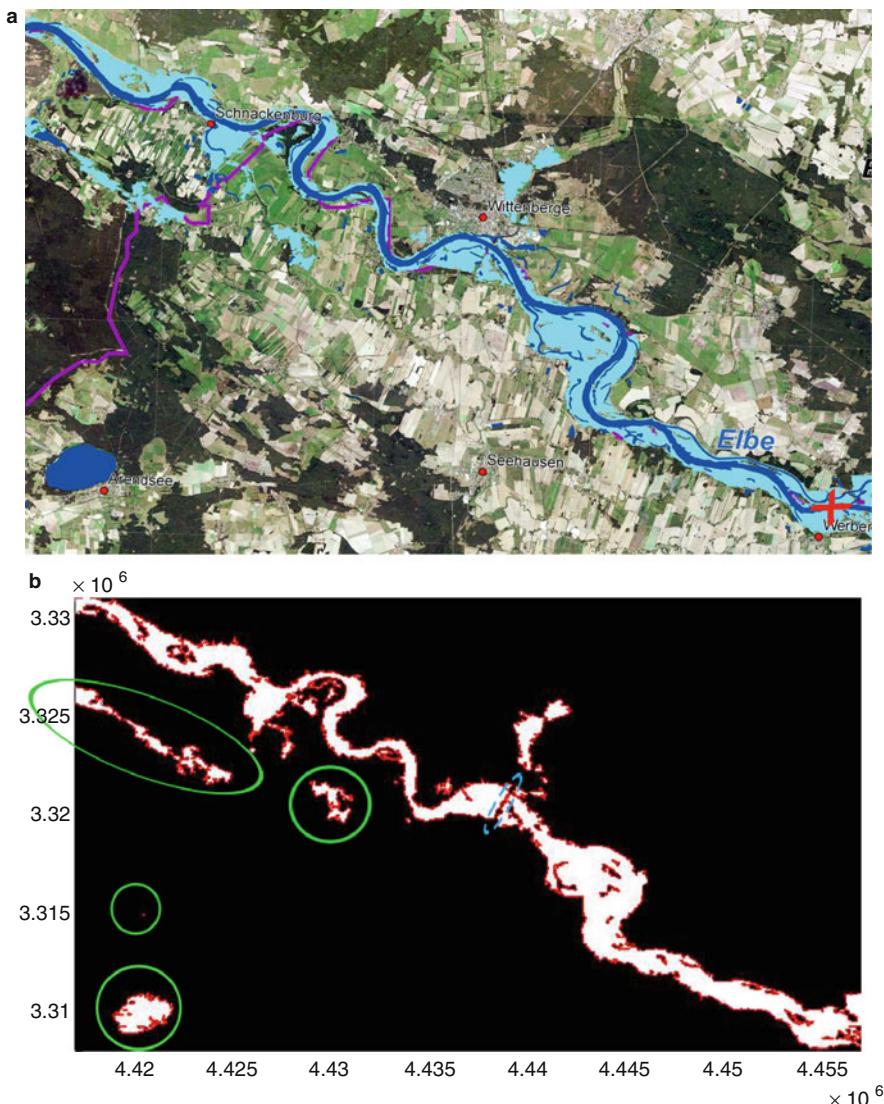


Fig. 2 Flood extent on 8 June 2013. Picture courtesy of DLR-ZKI (a), A bitmap of flooded region. White – observed flooding, black – no flooding. Circled areas excluded when comparing with simulations (b)

geometry requires the definition of bed elevation and channel widths along the river course. Since this information was not available to this study, the SRTM DEM was also used to define this input. Manning's n friction parameter values were defined across the domain (Fig. 4b) for different landcover classes taken from the Corine 2006 dataset (<http://www.eea.europa.eu/data-and-maps/data/corine-land-cover-2006-raster-2>). The

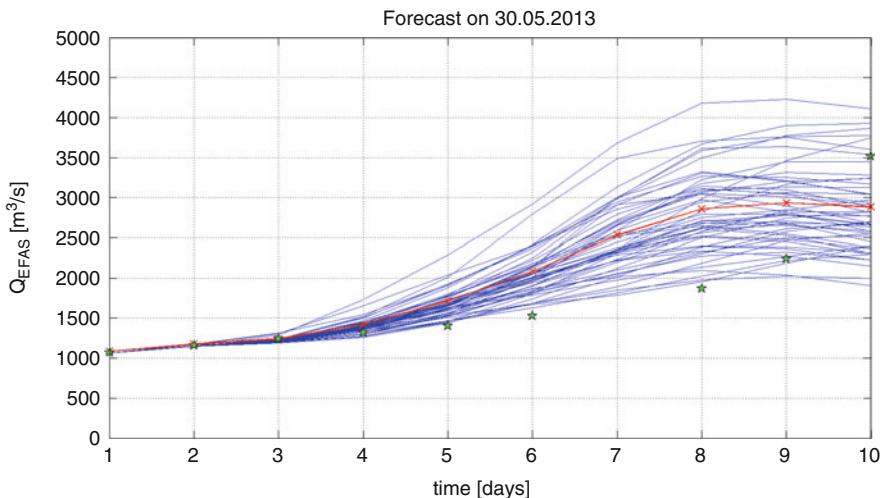


Fig. 3 Members of the EFAS ensemble forecast used as a forcing term. Red line is the mean of the ensemble, green stars are observations

lowest values were declared in the channel, thus allowing for preferential flow in these areas. Flow inputs, as described above, were taken from the results of the LISFLOOD hydrological model within EFAS driven by the ECMWF-IFS ensemble. To account for the lack of channel geometry representation in the SRTM DEM, owing to the inability of the radar signal to penetrate the water surface, the flow inputs were corrected and the mean annual flow was removed before the simulation. This meant that LISFLOOD-FP in this study would be used to route only the flood event flow discharge anomaly, as using the entire discharge would result in the overestimation of flood extent. Flow outputs were declared in the form of free boundary conditions at the northern and western edges of the domain. Finally the soil was treated as impermeable, and it was assumed that no precipitation or evaporation occurred within the floodplain.

These input data are available on a global or European scale and were used in an “as-is” way to show if, without modification, they are suitable for producing a probabilistic flooding forecast in this area. Moreover, we also sought to identify what modifications and/or additional information would be needed in order to reproduce the flooding observed in June 2013.

5 Results

5.1 Probability Maps

LISFLOOD-FP was run 51 times for all the members of ensemble forecast. These were independent simulations with different forcing discharges as depicted in Fig. 3. Each run provides the water levels and flood extent over time. Combining those

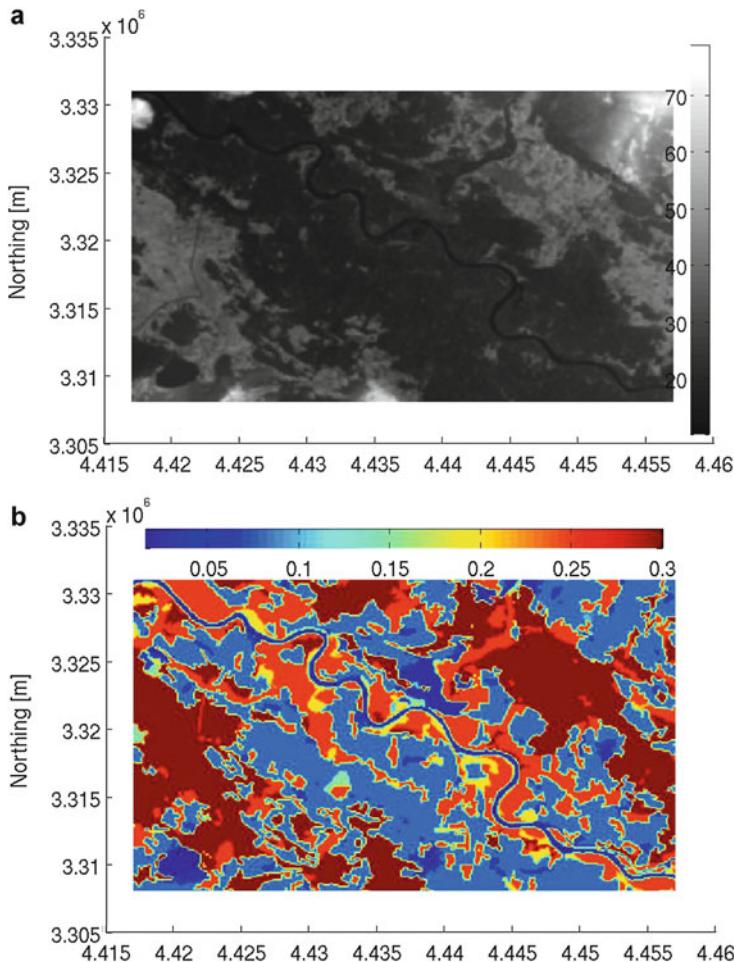


Fig. 4 Digital Elevation MAP (SRTM 100 m resolution) (a), Manning coefficient (b)

51 realizations, we can estimate a map of probable flooding for this particular forecast, as calculated from Eq. 4, where 0 means the area is never flooded, and 1 means the area is flooded in all 51 cases.

In Fig. 5, one can see the result of the simulations run on the original DEM with 100 m resolution compared to the observed flood extent marked with blue polygon line (from DLR-ZKI). Pixels are colored with the probability p_i (Eq. 4). The simulation overestimates the area of flooding which is partly due to the low resolution of SRTM DEM, as flood defenses, dykes, and roads are not continuously represented in this data set. Instead, there are gaps in these linear features through which the entire valley can be flooded. For example, see the satellite terrain image of the area contained in yellow rectangle in Fig. 5. The inability of the SRTM DEM to

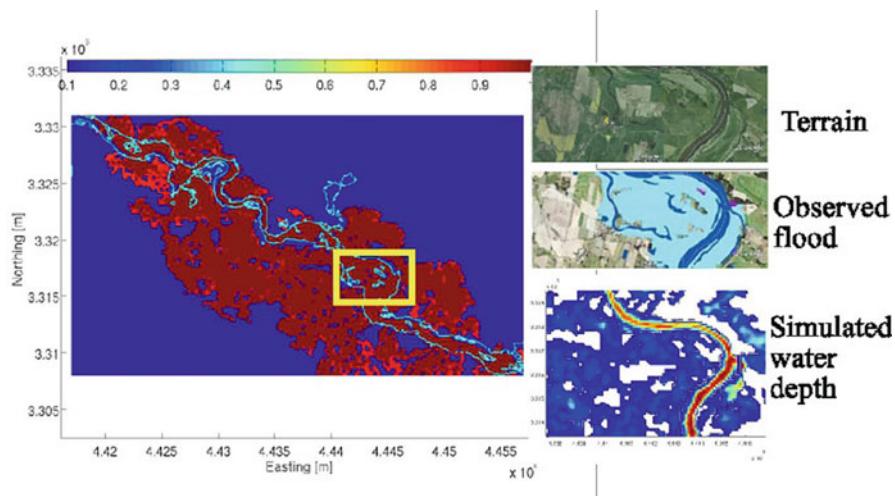


Fig. 5 Map of flooding probability on 08.06.2013. EFAS forecast from 30.05.2013

resolve such topographic features and the consequent hindrance on flood extent prediction has also been noted in previous studies (Sanders 2007).

It was decided to introduce an approximate representation of the flood defenses. The original DEM was modified by increasing the elevation of pixels lying along the course of flood defenses. The locations of flood defenses and dykes were approximated from open access data sources such as Google Earth. The city of Wittenberge is protected by flood defenses, while in peripheral areas elevated road networks acted as dykes, the locations of these are shown by the yellow line in Fig. 6. Flood defense and dyke pixels at these locations were raised by 2 m relative to their neighbors. The decision to use this particular height of approximate flood defenses was based on the average elevation of the roads in Google Earth (yellow lines in Fig. 6). The resulting probability map of inundation can be seen in Fig. 6. Here again, the blue line represents the flood extent observed on 8 June 2013. The flood defenses in this simulation contained water within the banks throughout the majority of the domain. Of the overbank flow which did occur, approximately 80% was onto the southern bank.

Neither of the simulations captured the inundation around the city of Wittenberge itself (northern bank, center of Fig. 2a). Flooding here may have happened due to inundation from the small Stepenitz river which flows along the eastern flank of the city. Flow inputs for this river were not defined in LISFLOOD-FP model.

5.2 Forecast Skill Scores

Both flood predictions discussed above were obtained applying the EFAS ensemble discharge forecast as a forcing term. In the first case, an unmodified SRTM DEM of

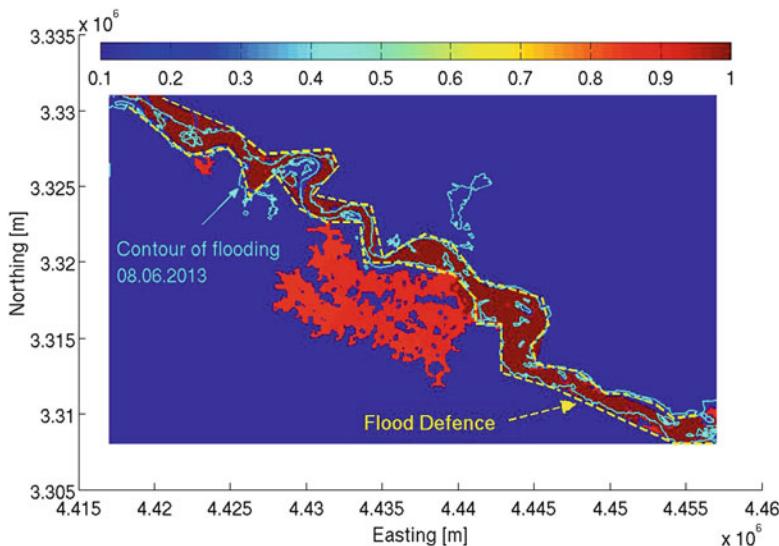


Fig. 6 Map of flooding probability on 8 June 2013. EFAS forecast from 3 May 2013. Modified DEM including flood defenses (*dashed yellow line*)

100 m resolution was used to describe the terrain. In the second case, we decided to include an approximation of the flood defenses, by modifying the original DEM. The spatial resolution and roughness in both cases remained the same. We ran 51 different simulations for both DEMs which provided us with a probability distribution of flooded cells. Next we quantified the skill scores of the two ensemble simulations. Table 3 compares the scores from the simulations with the different DEMs. Q1, Q2, and Q3 are the first (25%), second (50%), and third (75%) quartiles, respectively.

The hit rate (H) increased after introducing the flood defenses, this is because the water is kept within the secondary channel and covers more of the in-channel islands within the area of the observed flood. The false alarm rate (F) is small in both cases, despite the fact that the flooding extent is significant in the case without flood defenses. This is likely due to the inclusion of upland, never-flooded areas into the calculation of correctly predicted dry areas (D). Still, the false alarm rate (F) decreased significantly after introducing the flood defenses into the simulation. This resulted in a reduction in forecasted inundation extent and hence a reduction in the number of false-positive cells. Both Peirce and Critical Success Index score values were also increased after the introduction of flood defenses, which reflect the increase in the proportion of correct hits over false positives and misses.

The Brier Score for the forecast with the original DEM was equal to $BS_{\text{original DEM}} = 0.28$ and after introducing flood defenses, $BS_{\text{with defences}} = 0.08$. It means that the mean squared error of the simulation against observations is reduced when

Table 3 Skill scores distribution, comparison with the observation on 8 June 2013 (DLR-ZKI, Fig. 2b). Q1, Q2, and Q3 are the first, second, and third quartile, respectively

	Original DEM min,Q1, Q2 ,Q3,max	DEM with flood defenses min,Q1, Q2 ,Q3,max
H [%]	62.2, 67.8, 70.6 , 74.3, 79.9	70.0, 73.8, 75.0 , 76.1, 78.8
F [%]	17.4, 22.4, 24.3 , 27.1, 31.1	1.5, 3.0, 6.1 , 12.1, 21.5
PSS [%]	43.9, 45.7, 46.4 , 47.1, 47.8	57.2, 63.8, 68.8 , 70.3, 71.4
CSI [%]	18.0, 18.8, 19.2 , 19.6, 21.5	23.5 32.6, 45.1 , 55.4, 61.1

modifying the DEM. It is also shown in the Brier Skill Score (BSS) relating the forecast with original DEM and the modified one:

$$BSS = 1 - \frac{BS_{\text{withdykes}}}{BS_{\text{originalDEM}}}$$

The resulting value of $BSS = 0.71$ shows that the forecast using the DEM with flood defenses has more skill for the 2013 event than when using the original SRTM DEM.

6 Conclusion

The probabilistic forecasting of flood inundation is important for emergency management. It can give an advanced warning and quantify the probability of flood occurrence. In this chapter, we demonstrate using the example of the June 2013 floods in Central Europe a modeling framework that can derive probabilistic inundation maps. In this framework, meteorological forcing from the ECMWF-IFS ensemble was used to drive the LISFLOOD hydrological model within EFAS to produce discharge forecasts, in turn these forecasts were routed across the floodplain using the LISFLOOD-FP hydraulic model. These ensemble forecasts of inundation extent were then combined to produce a probabilistic forecast which highlights the areas most likely to flood. The forecasted inundation extent is extremely sensitive to the accuracy of the input DEM. Using the original SRTM DEM resulted in the overestimation of inundation, modifying this to include topographic components such as flood defenses reduced this problem and improved the forecast skill. Therefore, the quality of the input DEM will be an important consideration when establishing an operational flood inundation forecast framework. Overall this study shows that up to 80% of the flooded area along the Elbe in 2013 could have been forecasted to inundate 7 days in advance, using the probabilistic modeling framework proposed.

References

- L. Alfieri, P. Salamon, F. Pappenberger, F. Wetterhall, J. Thielen, Operational early warning systems for water-related hazards in Europe. Environ. Sci. Policy **21**, 35–49 (2012). 10.1016/j.envsci.2012.01.00

- L. Alfieri, P. Burek, E. Dutra, B. Krzeminski, D. Muraro, J. Thielen, F. Pappenberger, GloFAS global ensemble streamflow forecasting and flood early warning. *Hydrol. Earth Syst. Sci.* **17**, 1161–1175 (2013). <https://doi.org/10.5194/hess-17-1161-2013>
- P. Bates, A. de Roo, A simple raster-based model for flood inundation simulation. *J. Hydrol.* **236**, 54–77 (2000)
- P. Bates, M.S. Horritt, T.J. Fewtrell, A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling. *J. Hydrol.* **387**, 33–45 (2010)
- K. Bogner, H. Cloke, F. Pappenberger, A. de Roo, J. Thielen, Improving the evaluation of hydrological multi-model forecast performance in the Upper Danube catchment. *Int. J. River Basin Manag.* (2011). <https://doi.org/10.1080/15715124.2011.625359>
- K. Bogner, D. Meißner, F. Pappenberger, Korrektur von Modell- und Vorhersagefehlern und Abschätzung der prädiktiven Unsicherheit in einem probabilistischen Hochwasservorhersagesystem. *Hydrol. Wasserbewirtsch.* **58**(2), 73–75 (2014). https://doi.org/10.5675/HyWa_2014,2_2
- H.L. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**(3–4), 613–626 (2009)
- H.L. Cloke, J. Thielen, F. Pappenberger, S. Nobert, P. Salamon, R. Buizza, G. Blint, C. Edlund, A. Koistinen, C. de Saint-Aubin, C. Viel, E. Sprokkereef, Progress in the implementation of Hydrological Ensemble Prediction Systems (HEPS) in Europe for operational flood forecasting. *ECMWF Newsletter No 121* (2009), pp. 20–24
- M. Dale, J. Wicks, J. Mylne, F. Pappenberger, S. Laeger, Probabilistic flood forecasting and decision-making: an innovative risk-based approach. *Nat. Hazards* **70**(1), 159–172 (2014). <https://doi.org/10.1007/s11069-012-0483-z>
- A.M. de Almeida, P. Bates, Applicability of the local inertial approximation of the shallow water equations to flood modelling. *Water Resour. Res.* **49**, 4833–4844 (2013)
- D. Demeritt, S. Nobert, H.L. Cloke, F. Pappenberger, The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. *Hydrol. Process.* **27**, 147157 (2013). <https://doi.org/10.1002/hyp.9419>
- I.B. Mason, Binary Events, in *Forecast Verification a Practitioners Guide in Atmospheric Science*, ed. by I.T. Jolliffe, D.B. Stephenson (Wiley, Hoboken, 2003), 240 pp
- J. Neal, G. Schumann, P. Bates, A subgrid channel model for simulating river hydraulics and floodplain inundation over large and data sparse areas. *Water Resour. Res.* **48**, W11506 (2012)
- F. Pappenberger, K.J. Beven, N.M. Hunter, P.D. Bates, B.T. Gouweleeuw, J. Thielen, Cascading model uncertainty from medium-range weather forecasts (10 days) through a rainfall-runoff model to flood inundation predictions within European Flood Forecasting System (EFFS). *Hydrol. Earth Syst. Sci.* **9**(4), 381–393 (2005)
- F. Pappenberger, J. Thielen, M. del Medico, The impact of weather forecast improvements on large scale hydrology: analysing a decade of forecasts of the European Flood Alert System. *Hydrol. Process.* **25**(7), 1091–1113 (2011). <https://doi.org/10.1002/hyp.7772>
- F. Pappenberger, E. Dutra, F. Wetterhall, H.L. Cloke, Deriving global flood hazard maps of fluvial floods through a physical model cascade. *Hydrol. Earth Syst. Sci.* **16**, 4143–4156 (2012). <https://doi.org/10.5194/hess-16-4143-2012>
- F. Pappenberger et al., Floods in Central Europe in June 2013. *ECMWF Newsletter*, No. 136, Summer (2013)
- F. Pappenberger, L. Stephens, S.J. van Andel, J.S. Verkade, M.H. Ramos, L. Alfieri, J.D. Brown, M. Zappa, G. Ricciardi, A. Wood, T. Pagano, R. Marty, W. Collischonn, M. Le Lay, D. Brochero, M. Cranston, D. Meissner, Operational HEPS systems around the globe (2014), <http://hepex.irstea.fr/operational-heps-systems-around-the-globe/>. Accessed 10 Apr 2014
- M.H. Ramos, S.J. van Andel, F. Pappenberger, Do probabilistic forecasts lead to better decisions? *Hydrol. Earth Syst. Sci.* **17**, 2219–2232 (2013). <https://doi.org/10.5194/hess-17-2219-2013>
- E. Rodriguez, C.S. Morris, J.E. Belz, A global assessment of the SRTM performance. *Photogramm. Eng. Remote. Sens.* **72**(3), 249–260 (2006)
- B. Sanders, Evaluation of on-line DEMs for flood inundation modeling. *Adv. Water Resour.* **30**, 1831–1843 (2007)

- G.J.-P. Schumann, J.C. Neal, N. Voisin, K.M. Andreadis, F. Pappenberger, N. Phanthuwongpakdee, A.C. Hall, P.D. Bates, A first large scale flood inundation forecasting model. *Water Resour. Res.* **49**, 6248–6257 (2013). <https://doi.org/10.1002/wrcr.20521>
- J. Thielen, J. Bartholmes, M.-H. Ramos, A. de Roo, The European flood alert system – part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125140 (2009)
- J.M. Van der Knijff, J. Younis, A.P.J. de Roo, LISFLOOD: a GIS-based distributed model for river basin scale water balance and flood simulation. *Int. J. Geogr. Inf. Sci.* **24**, 189–212 (2010)
- N. Voisin, F. Pappenberger, D.P. Lettenmaier, R. Buizza, J.C. Schaake, Application of a medium-range global hydrologic probabilistic forecast scheme to the Ohio River Basin. *Weather Forecast.* **26**, 425446 (2011)
- F. Wetterhall, F. Pappenberger, L. Alfieri, H.L. Cloke, J. Thielen-del Pozo, S. Balabanova, J. Dahelka, A. Vogelbacher, P. Salamon, I. Carrasco, A.J. Cabrera-Tordera, M. Corzo-Toscano, M. Garcia-Padilla, R.J. Garcia-Sanchez, C. Ardilouze, S. Jurela, B. Terek, A. Csik, J. Casey, G. Stanknavius, V. Ceres, E. Sprokkereef, J. Stam, E. Anghel, D. Vladikovic, C. Alionte Eklund, N. Hjerdt, H. Djerv, F. Holmberg, J. Nilsson, K. Nyström, M. Sunik, M. Hazlinger, M. Holubecka, HESS Opinions “Forecaster priorities for improving probabilistic flood forecasts”. *Hydrol. Earth Syst. Sci.* **17**, 4389–4399 (2013). <https://doi.org/10.5194/hess-17-4389-2013>



Challenges of Decision Making in the Context of Uncertain Forecasts in France

Caroline Wittwer, C. de Saint-Aubin, and C. Ardilouze

Contents

1	Introduction	1400
2	Vigilance Procedure and Communication	1400
3	Flood Vigilance	1401
4	Operational Activity at Local and National Level	1402
5	Added Value of Ensemble Platforms	1407
6	Summary	1409
7	Conclusion	1410
	References	1411

Abstract

Flood is a major risk in France and an operational hydrometeorological organization (a team of 450 employees for national coordination and local services) is in place since 2003 for surveying the main river courses, about 22,000 km which can be damaged by flash flood, fluvial flood, and coastal flood hazard. Besides the building of a national network of hydrometric station, with real-time access of data on the Internet, the services are responsible for flood vigilance, over the next 24 h, and for more detailed forecasts at shorter lead-time. The information is disseminated since 2006 on the www.vigicrues.gouv.fr website. Hydrological and hydraulic deterministic models are used to increase the forecast lead-time for some 500 hydrometric stations located close to

C. Wittwer (✉)
BRGM, Orléans, France
e-mail: cwittwer@wmo.int

C. de Saint-Aubin
Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondations (SCHAPI),
Toulouse, France

C. Ardilouze
Météo France, Toulouse, France

vulnerable flood areas. Within the set of information integrated into the operational procedure, ensemble meteorological forecasts are used by the hydro-forecasters to evaluate rainfall distribution in the following days. Besides those routine activities, the European Flood Awareness System (EFAS) platform, the ensemble based forecasting system, has been tested since 2009 for some 60 flood events to study how the system can be best used within the French flood vigilance and warning procedure. During the first years of the evaluation, the results stayed quite difficult to be accounted for in the operational activities, as the number of false alarms and missed events was too high. Since 2013, the tendency is completely different, with the production of Flash Flood reports for watersheds smaller than 3,000 km². The localization of the impacted area and the timing is better foreseen; this tendency should be even improved when real-time discharges of main gauging stations will be transferred to EFAS.

Keywords

Vigilance · France · Operational flood forecasting services

1 Introduction

Two thirds of the 36,000 French cities are subject to one or more natural disasters (floods, earth quakes, clay deformation), with more than 6 million people living in flood prone areas. Reduction of the natural risk is therefore a national priority in order to adapt to these phenomena and to minimize their impact. Risk prevention policy is put in place with the aims of reducing the consequences of potential damages.

The risk prevention policy is based on a large set of activities including long-term adaptation measures (control of urban development and construction following legal standards and risk prevention master plans, reduction of vulnerability, adaptation to risk and climate change to prepare for future crises); resilient infrastructure development (protection, security of protection infrastructures); provision of training and real-time information to the citizen; and other short-term crisis management measures. This covers not only the surveillance, forecasting, and warning activities during crises but also all the preparative measures linked to the awareness of hazards and vulnerability, the gathering of historical information and for analyzing past crises, as well as preparedness to crises situations.

Here, we discuss how the French hydrometeorological services translate flood forecast information, including probabilistic forecasting, into so-called vigilance maps and written information to the public services and the population.

2 Vigilance Procedure and Communication

A new disaster awareness system called vigilance procedure was established in France following the devastating storm of 1999 along the Atlantic coast. Although the storm level (intensity, location) had been well estimated by meteorological

forecasters, the public underestimated the effect of the storm. Vigilance procedure provides clear instructions to the public on what to do in the case of a possible natural hazard over the following 24 h. It combines information on current situation and forecasts of the vulnerability of the area, in order to determine the right level of action to be taken by the services in charge of the security of the citizen and by the citizen themselves. Vigilance implies that the extents of natural hazards are continuously assessed and that threshold levels are set based on historical events, model simulations, local knowledge, and any type of information that allows linking hazard to impact on citizens and assets. Vigilance ensures that not only the technical and civil security services are aware about the level of any future events but also that the public is informed and become responsible for its own behavior. For this, communication means need to be in place for real-time assessment and coordination, and responsibility and information flow between the various technical services, civil security, and local services must be agreed. The vigilance procedure is used for hydrometeorological disasters including heavy rainfall and floods, as well as for meteorological and health related hazards, such as strong wind, snow-haze, avalanche, storm, heat wave, cold wave, wave-storm surge.

In the French risk prevention policy, the terminology makes a difference between *vigilance*, a warning procedure produced by technical services (either meteorological or hydrological) to the civil services and to the public for preparing to react properly in case of natural hazard, and *alert*, which triggers civil protection actions and is launched by the Prefect (the administrative coordinator at each of the 100 departments) when the danger is confirmed, when the forecasts indicate that safety measures and assistance become necessary.

3 Flood Vigilance

Flood vigilance information is produced by a coordinated effort of some 20 local flood forecasting services across France, responsible for a series of watersheds, and one national hydrometeorological and flood forecasting center (SCHAPI). Together with the services in charge of the hydrometric stations, this organization is operated by about 450 persons, performing either daily in normal situation or 24/24 h and 7/7 days during crises. All of them are technicians or engineers of the Government at national and regional level.

The goal of the vigilance procedure is to deliver twice daily, or more frequently, information on the level of hydrometeorological hazard before and during a disaster to the responsible party in charge of alerting the responsible services of the government, such as mayors at local level, the prefects at department level, the civil security services, and the general public. Vigilance information is provided for the upcoming 24 h. Four levels of vigilance are used for meteorological hazards (by Météo France) and for hydrological hazard (by the flood forecasting services). The levels are represented by four color codes: green, yellow, orange, and red. Green stands for no flood risk, yellow for localized floods with minor damages or rapid increase of water level, orange for more dangerous events with people at risk and damages on infrastructures and houses, and red stands for major floods extensively threatening

people and property and involving the support of national civil security services. For each of those vigilance levels, behavioral guidelines have been provided to decrease the risk on people and assets. The visualization of vigilance does not include information about the probability or uncertainties, e.g., there is no color shading or changing intensity of colors reflecting the probability range of the event. Instead, the vigilance maps are accompanied by bulletins where the uncertainties are expressed verbally and ranges of waterlevel forecasts are described either in the text itself or in attached files.

The flood hazard level classification is performed for linear segments of river network covering 22,000 km of streams under surveillance. The 250 linear segments, called river sections, ranging up to 40 km long, were identified depending on the need for detailed forecasts, the localization of vulnerable urbanized areas, and the availability of measuring hydrometric stations. All the gauging stations are linked to one of these river sections, and they relate water levels in the river course to damages in the area. Each of the past events can be classified on the vigilance scale, as shown on Fig. 1.

The products delivered by the flood forecasting services range from hydrographs at hydrometric stations to maps and bulletins for the river network. Since 2006, the website www.vigicrues.gouv.fr has provided direct access to real-time measures of some 1,500 stations, as well as to national and local vigilance information. This information is either sent to the decision-making public authorities or made available for free access on the Internet (Fig. 2). Radio and TV are also major tools of the dissemination process. Two of the regional services also provide applications to registered users when threshold levels are reached at selected measuring stations. This functionality will be implemented at national level in the near future, with the upgrading of the flood vigilance web services. Real-time communication gives therefore fast and easy access to field data, as well as to flood forecasting (vigilance levels and local forecasts). With continuous improvements (for example, access to historical data and event description), it is a very effective tool for increasing the participation of people to hazard and crisis management (Fig. 2).

The vigilance procedures for hydrological risks are fully aligned with the ones for meteorological risks – from the exchange of data to the dissemination of a joint meteorological and hydrological vigilance map, produced since 2007 at the request of the Ministry of Interior in order to deal with a unique concerted and expertized access to decision-making information.

4 Operational Activity at Local and National Level

A set of meteorological and hydrological information is used by SCHAPI for assessing the global hydrometeorological situation over the current day and the next days across France. This information at the national scale is valuable for ensuring that the local vigilance levels and forecasts produced by the 20 regional services are consistent and in line with the behavior of the main watersheds, as most of the main river courses are surveyed by a series of regional services. For the next

River section GARONNE TOULOUSE (1)		REFERENCE STATIONS OF THE RIVER SECTION	
Vigilance	Explanation and expected consequences	A reference station is the station selected for defining the vigilance color of the river section	
		STATION : CAZERES / GARONNE	
		Historical flood	Water level
R E D	Level 4 : Major risk of flood directly and extensively threatening people and property Rare and catastrophic flood, numerous human lives threatened, generalized overflow, generalized and simultaneous evacuations, disruption at large scale of urban, agricultural and industrial activities	June 1875	8.60 m
O R A N G E	Level 3 : Risk of flood with considerable overflow liable to significantly affect the daily life and security of people and property Generalized and damaging overflow, threatened human lives, evacuations, greatly affected circulation, partly disruption of social, agricultural and economic activities	February 1952	4.74 m
Y E L L O W	Level 2 : Risk of high or rapid rising water not involving significant damage but requiring particular vigilance in the case of seasonal and/or outdoor activities. Disruption of activities related to the river courses, localized overflows, localized secondary road closures, isolated houses involved, inundated cellars, disruption of agricultural activity	June 2000	3.70 m
G R E E N	Level 1 : no particular vigilance required Normal situation	January 2004	3.52 m
Warning: the choice of the color will also account for special cases, such as extremely rapid increase of water level, unusual event of the season or seasonal dangerous activity			
SPC GARONNE-TARN-LOT			

Fig. 1 Vigilance section of the Garonne River in the Toulouse area: compilation of past flood events for two gauging stations in relationship with vigilance levels

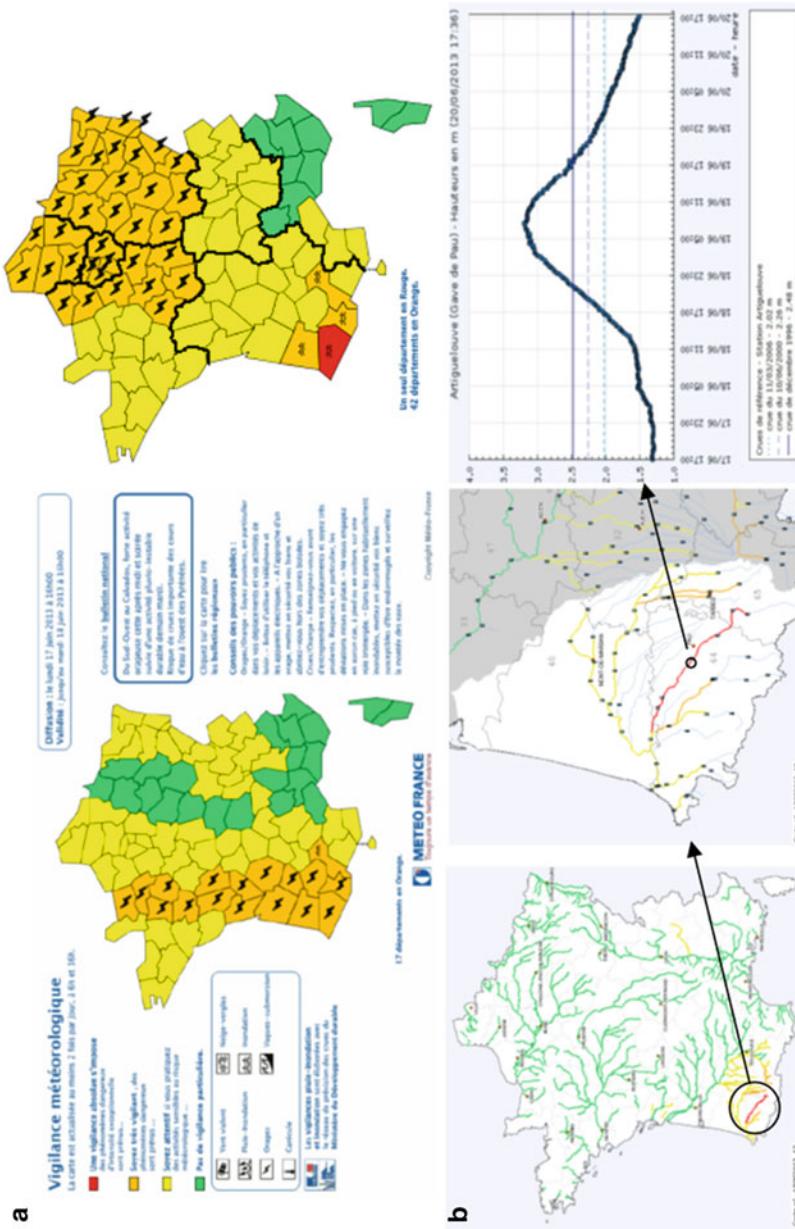


Fig. 2 Example of vigilance information available on www.vigicrues.gouv.fr during the June 2013 event in the Pyrenees. **(a)** Publication of the joint meteorological and hydrological map on June 17th and 18th. **(b)** Hydrological maps at national and regional level on June and related hydrograph for one of the gauging stations

days, the information is used for preparing and organizing work of the forecasting team (shifts, 24/24 h presence in the office). In any case, a first telephone call occurs between SCHAPI and the national civil security, the Inter-ministerial Center for Operational Crisis Management (COGIC) as soon an orange or red vigilance map is foreseen, therefore before any information is pushed via Internet. During crises, the whole process is continuously repeated, in order to follow the gravity and evolution of the situation.

The operational dataset used by the forecasters to assess the hydrometeorological hazard includes the following information:

- Operational radar-gauge rainfall grids (Tabary 2007) and moisture maps calculated with the SIM model (Noilhan and Mahfouf 1996), both provided by Météo France
- Observed hydrological data records (waterlevels and discharges)
- Deterministic meteorological forecasts at large scale from ARPEGE model from Météo France
- Rainfall quantitative forecasts for the next 3 days, with expertized quantitative forecasts for some 100 meteorologically homogenous subzones over the French watersheds at various ranges, one latest method being produced on 3 h timesteps, as well as comparisons of the available numerical models AROME (Seity et al. 2011) and ECMWF-ENS
- Meteorological hazard at midterm range, with probability rainfall maps of European Centre for Medium-Range Weather Forecasts (ECMWF)
- Medium-range, ensemble hydrometeorological models covering the entire national area, such as EFAS (Thielen et al. 2009) and the SIM-PE provided by Météo France (51 members of the ECMWF ENS forcing the SAFRAN-ISBA-MODCOU hydrometeorological model, Habets et al. 2008)

Each of the local forecasting services is responsible for assessing the extent of future flood hazard along the set of river sections under its responsibility. This means that various types of decision-aiding tools have been developed and are regularly used to provide a vigilance color on each of its river sections, usually at 10 am and 4 pm, for the following 24 h. The vigilance level, color, is often defined by using simple graphic representations of water level, or discharge, increase for ranges of precipitation forecasts over the next day. Additional criteria include initial conditions (e.g., river water level, soil moisture, snow cover, and frozen soil). For some 500 stations, quantitative forecast can be calculated using numerical models to provide values at different lead-times ranging from few hours to 10 days for probabilistic systems. In order to account for the heterogeneity of hydrological conditions at local level, from mountainous to fluvial and coastal watersheds, the forecasters use a quite large number of modeling tools, from empirical, conceptual, global, and distributed rainfall-runoff models to hydrological routing models. Hydraulic propagation is also represented by 1 D models. All models are running with deterministic rainfall distributions, either accumulation or forecasts at various time steps (12 h, down to 3 h) and spatial grids (8×8 km and $2,5 \times 2,5$ km, 1×1 km)

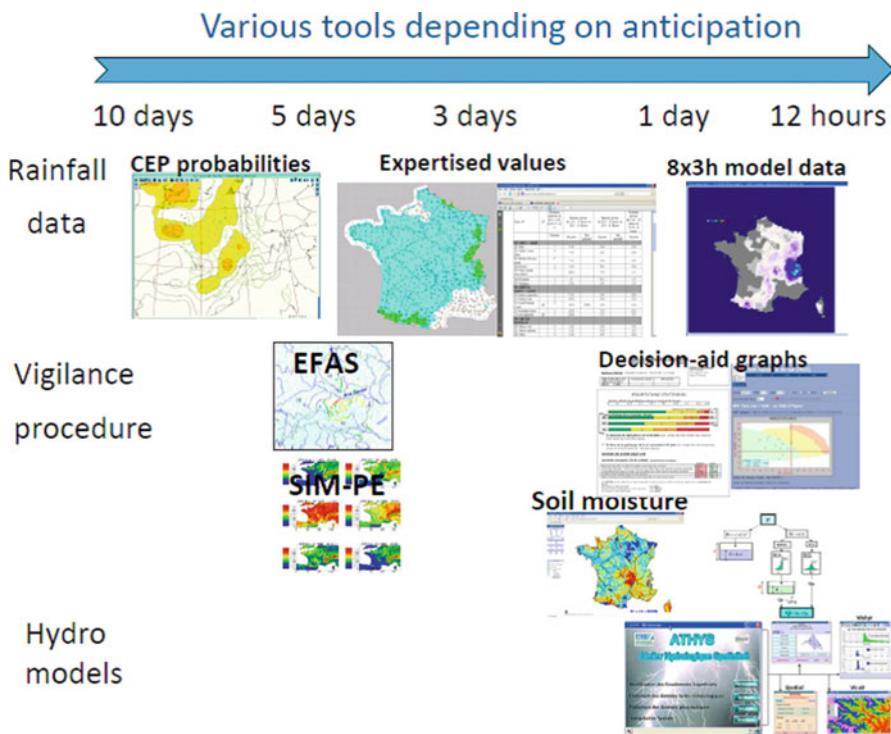


Fig. 3 Use of various meteorological and hydrological information by the forecasters over a 10-day sliding period

issued from numerical weather prediction models and/or from the radar network. When several models are providing a set of forecasts at the same station, the forecaster uses either expert decision or simple routines to determine the uncertainty around the chosen forecast value. These forecasts, and their uncertainty, are disseminated in the bulletin within the comment line of the river sections or as an attached file. The visual representation of the forecast and related uncertainty on the hydrograph is planned in the forthcoming new version of the Vigicrues website.

An illustration of the types of information involved into the assessment of the forecaster over a 10-day period before any event is given on Fig. 3. It shows when the various products provide valuable information and how they complement.

Decision, either for selecting a vigilance color on the 250 river sections or for coordinating the operational activities inside the flood forecasting services, at national and regional level, and for warning the civil security on future events, is based on the whole set of data. No action is taken based on only one source of information. Ensemble products are particularly valuable for early warning, as their forecast range is in line with the reaction time of most of the large watersheds, for which a first signal within 3–5 days before event is an additional criteria to inform the civil security services about a possible flood occurrence.

5 Added Value of Ensemble Platforms

Besides the operational procedure, SCHAPI is also involved into the testing of new products and the development of the national modeling strategy. Ensemble methods are used to assess longer term flood hazard including EFAS and are part of this program since 2009, when the first agreement was signed with the Joint Research Center for testing the platform. Results from SIM-PE (Thirel et al. 2008; Cloke et al. 2009), a similar system developed by Météo France at national scale, are also integrated to this testing procedure. During the last 5 years, EFAS and SIM-PE signals have been checked daily to identify possible areas with future flood events with as much anticipation as possible to compare the signals with the routinely used forecasting tools and the observed vigilance level during the event.

Over the period 2009–2014, each event announced either with EFAS alerts and watches or with orange and red flood vigilance has been evaluated on various criteria, such as the size of the watershed, spatial distribution and type of precipitation (rain/snow), alert level compared to real situation, and lead-time. More than 60 events with orange vigilance and almost 10 with red vigilance have been analyzed and classified into “hits,” “missed,” and “false alarm” groups. Results have shown that the medium-range forecasts provided for watersheds larger than 4000 km² are less useful for forecasters in their decision making for the vigilance maps. With the introduction of the so-called flash flood layers into EFAS in 2012 (see Alfieri in this handbook), the warning on watersheds below 3000 km² is now also provided by the system. In 2013, a ratio of “hits” of about 50% for the relatively large watersheds was observed and a very high ratio of 80% for the Flash Flood reports on smaller watersheds. These improving results are providing more valuable information to the operational forecasters: the system can provide warnings up to 4 days in advance, which is greatly valuable for organizing the forecasting activities and also for first exchanges with the civil security services. Warnings sent only 1 or 2 days before an event by the EFAS system are less informative, as the situation is already localized by other models and the timing is also better defined.

An example from EFAS system for one of the November 2014 events on the Mediterranean shore is shown in Fig. 4, while the corresponding vigilance maps sent over the same period are presented below on Fig. 5. Besides the very good performance of EFAS for the Aude, Têt, and Tech rivers, it is worthwhile to notice that EFAS Flood Watch had been received on November 22 for Hérault (2525 km²) for November 27th, therefore longer anticipation. On the other hand, no signal had been sent for the Berre river, although it reached red vigilance level on Vigicrues on November 30th.

A detailed evaluation of the EFAS flash flood results for 2014 is still underway but first results show that valuable information concerning the area covered by an incoming event and its timing are provided to the operational teams by the ensemble system.

The active participation of the operational team in the testing of two ensemble EFAS and SIM-PE platforms during their development phase and thereafter has been a very valuable training for the operational forecasters. It allows them not only to learn about the methodology but also to assess in real-time, as well as a posteriori, the value of the warnings within the overall set of hydrometeorological information

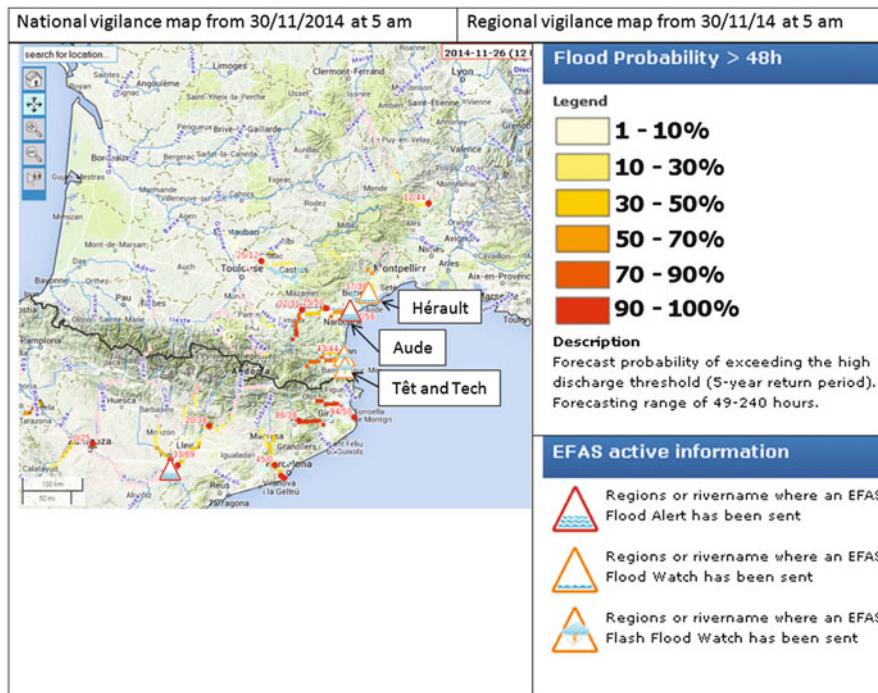


Fig. 4 Flood probability forecast provided by EFAS system for Aude river ($5,500 \text{ km}^2$) for November 26, 2014 issued 3 days ahead. The red triangles show the flood alert and the orange triangles show thunderstorm for Tech (731 km^2) and Têt ($1,412 \text{ km}^2$)

from various sources which are provided to the forecaster in normal situation and during crises. The forecasters are using additional information to evaluate if the level of warnings issued by the ensemble are in agreement with the local hydrological models and with the real water levels reached during events. This is for example the accuracy of the location and surface of the impacted area, the timing, duration, lead-time and consistency of the signal. In return, the possibility of providing feedback to the developers has allowed improving the systems, reducing false alarm rates, and visualizing information in the best way for the expert users.

In line with the testing of EFAS and the SIM-PE platform, several national projects have been developed, and implemented, by the flood forecasting services and its research partners, in order to concentrate on more local floods and short-term forecasting. The purpose there is to provide new methodologies and tools to deal with the spatial and temporal variability of precipitation forecasts, which are influencing the behavior of smaller, upper parts of watersheds and causing devastating floods over the last years. These new systems aim at providing forecasts for areas which are not yet covered by the river network under "hydrological vigilance surveillance," and responding therefore to the expectation of the citizens affected by these rapid floods, as well as to the public authorities. Following two tragic events in

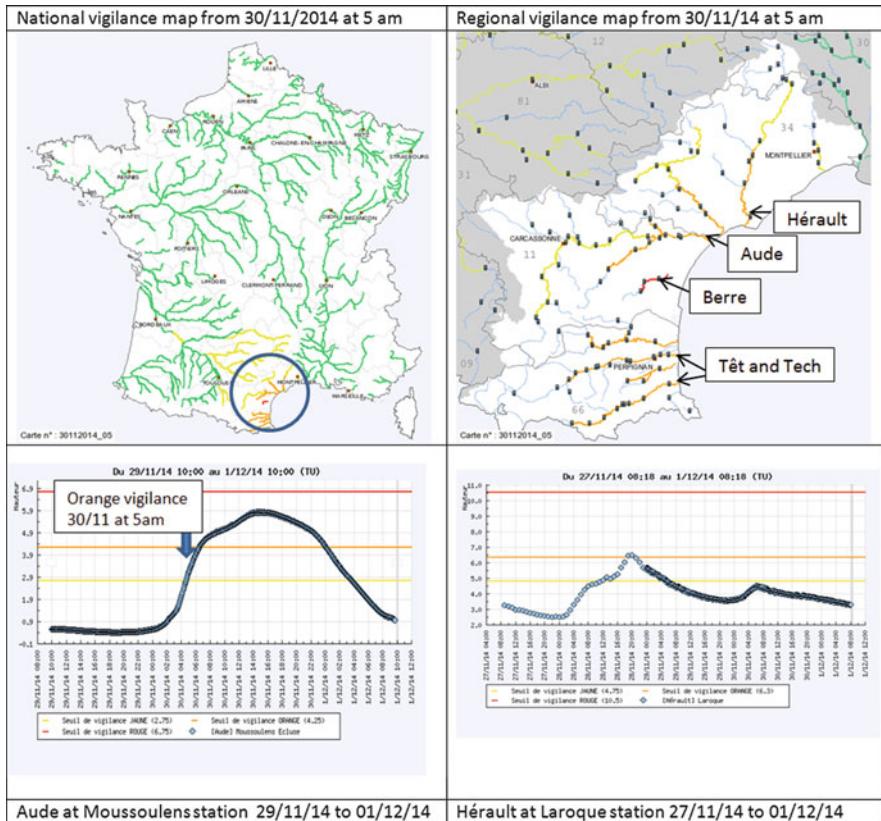


Fig. 5 Vigilance maps and hydrographs of two stations on Aude and Hérault rivers for the November 30, 2014 event

2010 along the Atlantic coast and the Var department Mediterranean area, the ministry in charge of sustainable development acted a national management plan for rapid submersions (MEDDE 2011), to decrease the effects of coastal and small watersheds floods. These two projects are one of the measures of this large program covering four main areas: (a) control of urban planning and adaptation of constructions; (b) improvement of hazard understanding, warning systems, forecasting, vigilance, and alert; (c) security of waterworks and protecting infrastructures; and (d) improvement of resilience.

6 Summary

In France, a major reform of the national and regional flood services in 2002 has been enforced to modernize the former procedures, based on observation of river levels, into a national real-time operational system, delivering forecasting and

warning information on the main river courses (Tanguy et al. 2005). The network covers about 22,000 km of streams (out of the 1,20,000 km of streams larger than 1 m width) and improves the anticipation before flood events for 90% of population at risk, on the basis of a variety of data and numerical hydrometeorological models, including ensemble prediction systems. A complementary procedure is under study for flash floods on small watersheds in order to inform an even larger percentage of the population.

Up to now, the probabilistic information, whether meteorological or hydrological, is mainly used by the operational expert teams, who evaluate the platforms results and provide a “deterministic” assessment to the decision makers (Prefects, civil security). A first success has already been reached by Météo France, the French National Meteorological Service, to convey uncertain warnings to forecasters, technical public authorities, and civil security, by producing a daily map of possible extreme (orange and red) events for the next 2 and 3 days at regional level. The probabilistic information or the degree of uncertainty is not yet translated to the public but a new version of the Vigicrues website will include uncertainty ranges on the forecasts. However, the aim of the vigilance procedure is to make the public aware that a potentially dangerous situation is expected and that adapted behaviors must be followed.

The situation has also evolved inside the flood services: 5 years ago, only the national center, SCHAPI, was convinced of the value of ensemble warnings and ready to invest time to be educated to this new approach and to assess regularly the platform results (EFAS and SIM-PE). Now, several local services, mainly those concerned by short-term rapid floods (6–12 h reaction time of the watersheds), are also interested into this type of information, as EFAS has produced particularly well-located and timely forecasts over the last 2013 and 2014 flood events. Furthermore, the tests performed with the current models under development, CHROME and ([AIGA](#) and [CHROME](#)), are also providing excellent forecasts, especially in the case of orographic effects and rapid floods.

Through working groups and evaluation of the different components of the EFAS and SIM-PE systems, improvements and needs for adaptation to the national procedures and local hydrological behavior have been identified by the French forecasters. These results can not only be implemented into the European platform but they are also valuable to develop additional dedicated numerical models and approaches. The future of ensemble forecasting is therefore large for operational forecasters, not only in testing and developing new products but also in transferring the information to the end-users, at technical and decisional level, as well as to the public at longer term.

7 Conclusion

With more than 6 million people living in flood prone areas, two thirds of the French cities are subject to one of more natural disasters (floods, earth quakes, clay deformation).

References

- P. Arnaud, J. Lavabre, Coupled rainfall model and discharge model for flood frequency estimation. *Water Resources Research*, **38**–6, (2002). <https://doi.org/10.1029/2001WR000474>
- H. Cloke, J. Thielen, F. Pappenberger, S. Nobert, G. Balint, C. Edlund, A. Koistinen, C. De Saint-Aubin, E. Sprokkereef, C. Viel, P. Salamon, R. Buizza, Progress in the implementation of hydrological ensemble prediction systems (HEPS) in Europe for operational flood forecasting. *ECMWF Newslett.* **121**, 20–24 (2009)
- F. Habets, A. Boone, J.L. Champeaux, P. Etchevers, L. Franchisteguy, E. Leblois, E. Ledoux, P. Le Moigne, E. Martin, S. Morel, J. Noilhan, P. Quintana Segui, F. Rousset-Regimbeau, P. Viennot, The SAFRAN-ISBA-MODCOU hydrometeorological model applied over France. *J. Geophys. Res.* **113**, D06113 (2008). <https://doi.org/10.1029/2007JD008548>
- MEDDE, Plan Submersion Rapide. Document of the Ministry for Ecology, Sustainable Development and Energy (2011)
- J. Noilhan, J.F. Mahfouf, The ISBA land surface parameterisation scheme. *Global Planet. Change* **13**(1–4), 145–159 (1996)
- Y. Seity, P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, V. Masson, The AROME–France convective-scale operational model. *Mon. Weather Rev.* **139**, 976–991 (2011)
- P. Tabary, The New French operational radar rainfall product. Part I: methodology. *Weather Forecast.* **22**(3), 393–408 (2007)
- J.-M. Tanguy, J.-M. Carrière, Y. Le Trionnaire, R. Schoen, Réorganisation de l'annonce des crues en France. *Houille Blanche – Revue Internationale De L'Eau* **2**, 44–48 (2005). <https://doi.org/10.1051/lhb:200502005>
- J. Thielen, J. Bartholmes, M.-H. Ramos, A. de Roo, The European flood alert system–Part 1: concept and development. *Hydrol. Earth Syst. Sci.* **13**, 125–140 (2009). <https://doi.org/10.5194/hess-13-125-2009>
- G. Thirel, F. Rousset-Regimbeau, E. Martin, F. Habets, On the impact of short-range meteorological forecasts for ensemble streamflow prediction. *J. Hydrometeorol.* **9**, 1301–1317 (2008)
- B. Vincendon, V. Ducrocq, O. Nuissier, B. Vié, Perturbation of convection-permitting NWP forecasts for flash-flood ensemble forecasting. *Nat. Hazards Earth Syst. Sci.* **11**, 1529–1544 (2011).



Hydrological Ensemble Prediction Applied in China

Guangsheng Wang, Zhijie Yin, Jianqing Yang, and Yuhong Yan

Contents

1	Brief Introduction of Hydrological Forecasting in China	1414
2	Short-Range Ensemble Predictions for Flood Control	1415
3	Long-Range Ensemble Prediction	1419
3.1	The Rainfall Runoff Ensemble Prediction	1419
3.2	Conditional Probability	1422
4	Conclusion	1426
	References	1426

Abstract

In China, to prolong lead-time prediction in order to meet with the requirement of flood defense, short range ensemble stream flow prediction is implemented. Based on the forecasting and the historical storm model, a group of precipitation data is generated as input of rainfall-runoff model. Furthermore, to meet with the requirement of water resources management, in the humid regions rainfall-runoff ensemble prediction has been applied for monthly runoff prediction, in large basins conditional probability method has also been applied to assess the next monthly runoff probability under a fixed runoff initial condition.

Keywords

Hydrological ensemble prediction · Short range flood forecasting · Monthly runoff prediction · Rainfall runoff ESP · Conditional probability prediction

G. Wang (✉) · Z. Yin · J. Yang · Y. Yan

Bureau of Hydrology, Ministry of Water Resources, Beijing, China

e-mail: gshwang@mwr.gov.cn; yinzhijie@mwr.gov.cn; jianqing.yang@mwr.gov.cn;
yyh@mwr.gov.cn

1 Brief Introduction of Hydrological Forecasting in China

In China hydrological forecasting is undertaken mainly in seven major basins, from north to south in turn are Songhua, Liaohe, Hailuan, Yellow, Huaihe, Yangtze, and Pearl River basins. From north to south, the amount of rainfall is increasing gradually, and annual precipitation of major basins is shown in Table 1. Huaihe, Yangtze, and Pearl River basins are located in humid regions in south China; Songhua and part of Liaohe River basins are located in semi humid regions in north China mainly because of their high latitude low evaporation, and other river basins are located in arid and semiarid regions. Rain season in south China is normally from May to August, while in north China is from June to September, where the amount of rainfall exceeds 50% of annual precipitation, especially in north China.

At present, there are 2867 hydrological stations (monitoring precipitation, water level, discharge elements), 1677 water level stations (monitoring precipitation, water level elements), 11,925 precipitation stations which collect and transmit real time data for the purpose of hydrological forecasting and flood warning. Among these, hydrological forecasting is undertaken for 1039 hydrological stations.

Models used in hydrological forecasting in China include Xinanjiang (Rrenjun 1992), API, Horton infiltration, Sacramento, Smar, Muskingum, and Lag & K for channel routing, etc. The Xinanjiang model is a conceptual model, runoff production occurs on repletion of storage to capacity values which are assumed to be distributed throughout the basin, runoff was separated into surface flow, interflow, and ground flow in the model. From 1990s, the distributed Xinanjiang Model was developed. Applying these models in humid and semi humid regions can get reliable forecasting.

Hydrological forecasting is mainly used for flood control, so most of forecasting is short range storm flood forecasting (within 3 days), where rain gauge observed precipitation is input to the forecasting system to produce discharge and stage forecasting. In the last decade, as the quantitative precipitation forecast has been released by meteorology service, therefore prolonged lead time hydrological forecasting has been made using observed and forecast precipitation. Following the use of forecast precipitation, the ensemble prediction was then implemented for flood

Table 1 Mean annual precipitation and area of major river in China (China Water Conservancy and Hydropower Planning and Design Institute 2014)

Basin	Precipitation (mm)	Basin Area (km ²)
Songhua	504.8	557,180
Liaohe	545.2	228,960
Hailuan	534.8	308,531
Yellow	445.8	752,443
Huaihe	838.5	269,283
Yangtze	1086.6	1,808,500
Pearl	1549.7	453,690

control. In recent years, long-range ensemble prediction, usually monthly runoff prediction, is also undertaken for water resources management.

2 Short-Range Ensemble Predictions for Flood Control

Storm flood forecasting is crucial for flood control decision-making. When rain gauge observed precipitation is taken as the only precipitation input of hydrological model, foresight time is generally less than basin concentration time, and if there are rainfall data available in foresight time then the forecast would be more valuable. So from year 2000, after the quantitative precipitation forecast was released by meteorology service, the forecast precipitation is used as input to hydrological forecast model, as well as the observed precipitation. The resolution of forecast precipitation is usually 1° of longitude multiply 1° of latitude, which is about 110 km multiply 110 km. To obtain the precipitation at basin scale, the grid of forecast precipitation need to be changed into equal area projection and smaller grids using GIS tool.

The weather system is a nonlinear and complex system. It is impossible to forecast exact rainfall amount; the forecast precipitation that meteorology service issued is single deterministic volume. Usually there is difference between the forecast precipitation and real precipitation, thus the uncertainty of precipitation forecasting would be brought into hydrological forecasting, and this uncertainty is more concerned than the uncertainty of hydrological model. For this reason the ESP was introduced into short-range flood forecasting to account for the uncertainty of rainfall. Based on the forecast precipitation, a group of precipitation data is generated by hydrologist, including those less than forecast volume, bigger than forecast volume, and also equal forecasting volume; this group of precipitation is input to the hydrological model to produce a group of forecast result.

The time intervals of forecast precipitation are 12 h or 24 h, which cannot meet the need of hydrological models, so the 1 day and 3 days forecast precipitation need to be divided into smaller time interval, i.e., 6 h or smaller than 6 h. For hydrologists, the most reasonable and realistic method to assign forecast precipitation into every small time interval is to consult historical storms. Therefore, several important historical storm processes are used as reference models of time distribution of storm rainfall, similar type of weather system such as typhoon rain, frontal rain etc., and similar type of season (month) historical storm is selected for the time distribution models of forecast precipitation, as shown in Table 2. The amount of forecast precipitation is divided into every time interval of hydrological forecasting models as same percent as the time interval of selected historical storms.

Here an ESP example of severe flood in year 2010 at Fengman reservoir is presented. Fengman reservoir is located at upstream Second Songhua River in Songhua River basin, northeast China's Jilin province. The basin area is 42,500 km², and normal annual precipitation is about 750 mm, among which 70% falls in rainy season (from June to September). This area is a semi-humid region and mountainous area, where 194 km upstream of Fengman reservoir is Baishan

Table 2 Precipitation process of several historical storm Fengman basin at time interval 6 h in mm

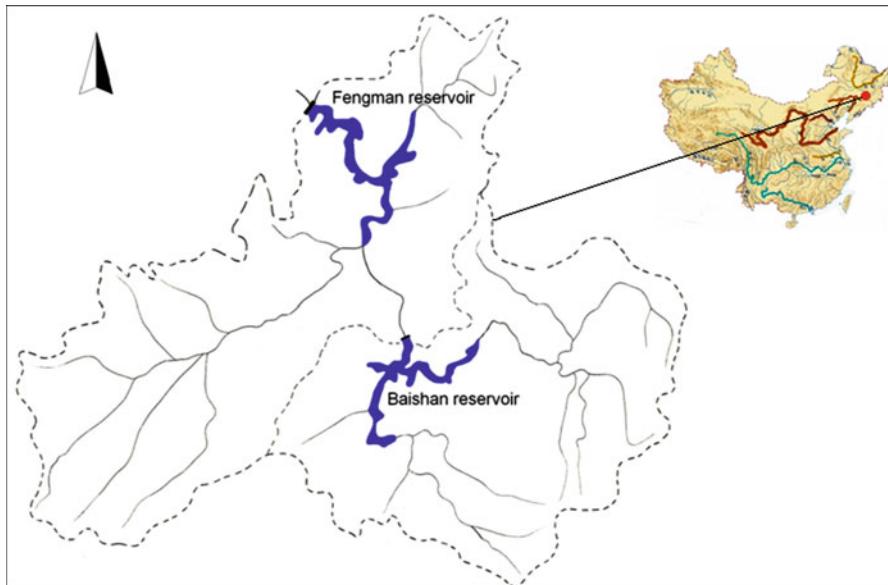


Fig. 1 Basin area and river network of upstream Second Songhua River

reservoir, and basin area between Baishan and Fengman is $23,500 \text{ km}^2$. The basin areas, river networks, and locations of two reservoirs are shown in Fig. 1.

The average rainfall runoff concentration time is about 10 h; forecasted inflow of Fengman reservoir consists of two parts, that are, Xinanjiang rainfall runoff modeling in the area between Baishan and Fengman, and L&K channel routing of released flow from Baishan reservoir.

The river flow rose after several continuous raining days, at 8 am, 28 July 2010, and a precipitation forecasting was released by meteorology service that in future 24 h the precipitation amount would be 40 mm. Based on the forecasted precipitation, 30 mm, 40 mm, 50 mm, 60 mm was set down as input precipitation of ESP. Meanwhile the 21 July 1971 storm was selected as a process model, the total precipitation was assigned to every 6 h time interval. Forecasted discharge hydrographs is shown in Fig. 2, and details are shown in Table 3. In this case, the 60 mm precipitation forecast result is finally recommended for flood control decision-making, which takes into account the optimization of dam operation in order to protect safety of downstream area and dam. For the recommended forecast result, flood peak and volume are forecasted satisfactory; the real precipitation is 56 mm similar with the recommended, but because some difference between real precipitation process and the 21 July 1971 model, the forecasting flood peak is lag.

From year 2000, this kind of ESP has been applied in short-range flood forecasting, and strong supports are provided for flood control decision-making, especially for the severe flood of Huaihe River in year 2003 and 2007, and the severe flood of Yangtze and Songhuai Rivers in year 2010. However, there are still issues remaining

Fig. 2 Forecast and observed discharge hydrographs of Fengman July 2010

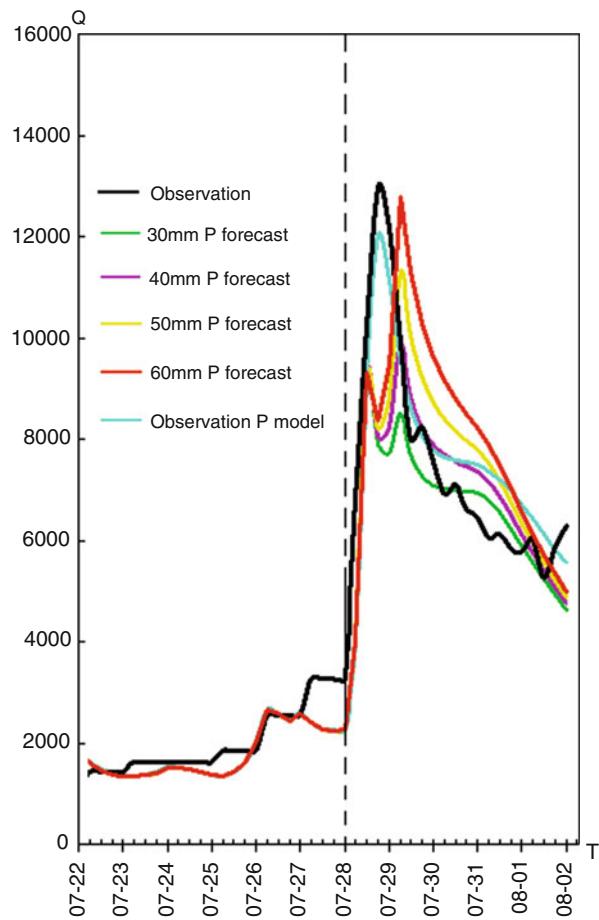


Table 3 Forecast and observed flood peak and volume of Fengman July 2010

	P 30 mm forecast	P 40 mm forecast	P 50 mm forecast	P 60 mm forecast	Observed P 56 mm model	Observed discharge
Flood peak Discharge (m^3/s)	9320	9862	11,302	12,791	12,013	12,973
Flood volume ($\text{m}^3 \cdot 10^9$)	2.93	3.12	3.32	3.52	3.35	3.31

in the application of ESP in flood control. That is, a group of hydrological forecasting results are produced, but the probability of every forecasting result is not presented or every forecast is assumed to have the same probability. In the future, if probabilistic quantitative precipitation forecasting is provided by the meteorology

service, or otherwise, if the probability distribution of forecast precipitation can be estimated through analysis of a series of error of forecast precipitation, then a group of flood forecasting results, each with a probability, can be produced. Nonetheless, there is a gap between ESP and flood control regulation. At present, all these regulations are based on fixed flood index, i.e., flood peak discharge or level, flood volume. Hydrologists usually have to select the preferred one for flood control decision-making depending on their forecasting experiences, and the comparative safe forecast result is generally preferred.

3 Long-Range Ensemble Prediction

Long-range runoff prediction is important for water resource management especially in low-flow period, as the water resource issue become more important and the reinforcement of water resources management long-range runoff prediction is attached more importance.

In early years, to predict long-range future river runoff, the normal mean monthly volume was normally used as predicting volume, and the probability method was also used. For the long-range predictions in low-flow period, flow recession model was used, and based on the principle of flow recession correlation diagram, initial time discharge and mean monthly discharge were also developed for monthly runoff prediction. In the probability method, the initial condition cannot be used to reduce the uncertainty, while the uncertainty cannot be considered in other methods. For the long-range hydrological prediction, the uncertainty is more notable than for short range hydrological forecasting, so the ESP is undertaken in monthly time step.

Long-range ESP is used to support dam operation and river flow regulation in order to solve water resources issues including ensuring water supply for agriculture irrigation, industry consumption, and household consumption, across basin water diversion and aquatic environment protection.

3.1 The Rainfall Runoff Ensemble Prediction

From last decade, rainfall runoff ESP has been studied and carried out in monthly runoff prediction in humid regions. Usually, in humid region hydrological models can produce reliable forecast; the initial soil moisture condition has more determinative effect on runoff generation. In arid region soil moisture generally is stable, remains at very low level, and runoff generation is less affected by initial soil moisture. Future runoff are mainly determined by future rainfall and initial conditions, where the future precipitation is rather a random term, while the initial soil moisture can be simulated reliably and affect the future runoff in humid regions. Thus, the rainfall runoff ESP is more valuable in humid regions.

The Xinanjiang model is used in rainfall runoff ESP. Usually it is operated as daily model whose time interval is 1 day. Most of hydrological models are used for flood forecasting, while long-range ESP meets the need of water resources

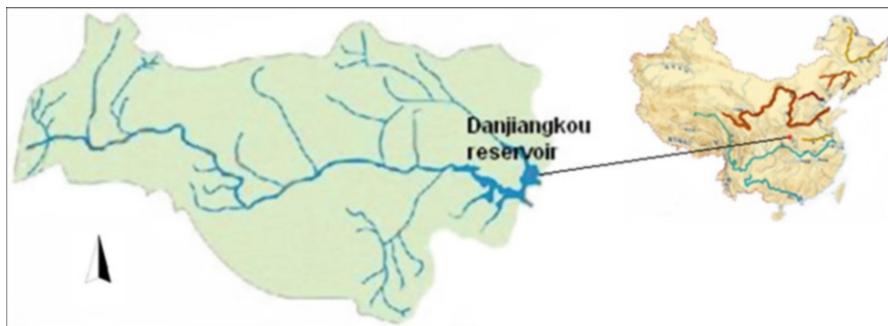


Fig. 3 Basin area and river network of upstream Han River

management. The low-flow forecast accuracy is important, especially for water resources management. In this case, it is likely that model parameters need to be rectified to fit for low-flow forecasting. Precipitation series used in ESP require data series exceeding 30 years in general. In China, there are more than 50 years observation records for most of rain gauges. Practically, 1 week forecast precipitation can be taken for valuable reference, so precipitation of the first week in each historical monthly precipitation process used in ESP should be replaced by the forecast precipitation, to reduce the uncertainty of model forecast. Mathematical expectation value and probability distribution of mean monthly discharge are forecasted, and Pearson III, normal and log-normal probability distribution, is used for forecast discharge probability analysis.

In China, application of the long-range rainfall runoff ESP is still in primary phase. A number of research studies have been carried out (Yan et al. 2008; Gobena and Gan 2010; Bao and Linna 2012; Long et al. 2013; Tao et al. 2014), but operations have been undertaken only in part of southern China rivers. The most significant case is the Danjiangkou reservoir (Yan et al. 2008; Bao and Linna 2012).

Danjiangkou reservoir is located at the upstream of Han River, in Yangtze River basin, south China Hubei province. The basin area is 95,200 km², the normal annual precipitation is 800–1100 mm, where 80% of the precipitation falls in the period of May to October, and the normal mean annual discharge is 1200 m³/s. This area is a humid region and mountainous area. Basin area, river network, and location of the reservoir are shown in Fig. 3. Danjiangkou reservoir is an important water source for local area, and the special importance is that the reservoir is the source of water for the south-to-north water diversion project, which diverts water from south China to north China.

The Xinanjiang model was used in rainfall runoff modeling of Danjiangkou reservoir; 30-years historical precipitation data series and Pearson III probability distribution function was used for forecasting runoff probability. This is a humid large basin so that the initial condition affects more on the next month runoff. Figure 4 is the forecast discharge hydrographs in October 2007, and Table 4 is the forecast and observed mean monthly discharge from July to October 2007.

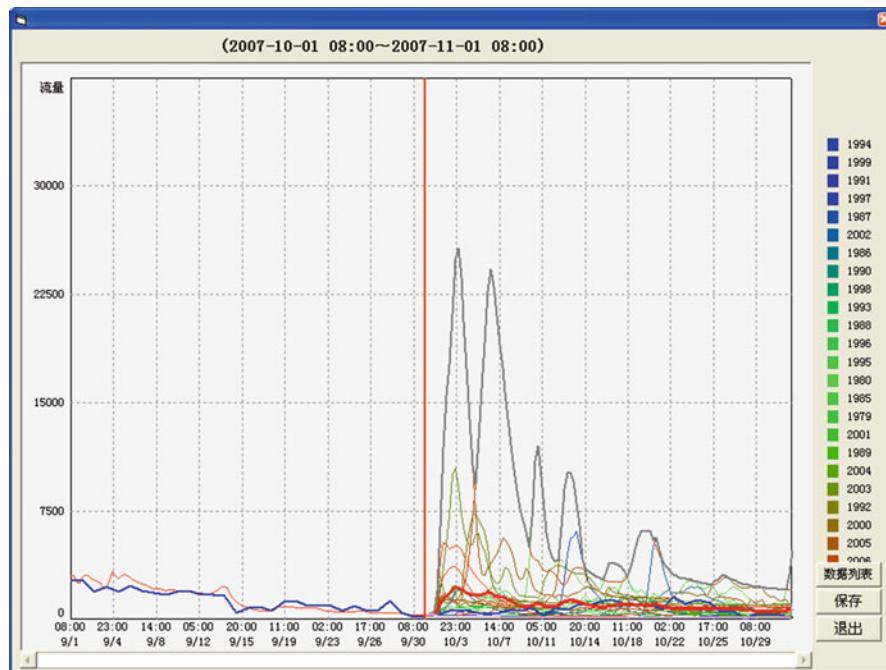


Fig. 4 Forecast discharge hydrographs in October 2007 Danjiangkou

Table 4 Observed and forecast mean monthly discharge from July to October 2007 Danjiangkou

Month	Observed	Prediction						
		Expectation	10%	25%	50%	75%	90%	
7	4408.2	2056.5	3169.1	2613.1	1994.8	1376.9	821.0	
8	2567.2	1992.6	3394.9	2703.1	1932.9	1164.1	471.2	
9	1381.6	1969.5	3761.2	2893.9	1930.2	966.8	99.2	
10	565.6	672.4	2265.9	1344.8	659.0	265.8	96.0	

For the operation of ESP, the correlation of initial soil moisture and river flow with next mean monthly river flow should be analyzed, the better correlation the more valuable ESP. ESP takes the same initial soil moisture and base flow, input historical precipitation series data to hydrological model, so the coefficient of variation of probability distribution of predicted mean monthly discharge should be obviously less than the coefficient of variation of historical mean monthly discharge series that correspond with the historical precipitation series ESP model inputted. Otherwise, the rainfall runoff model is not fine or the next month runoff is not decided by initial condition obviously.

In probability computing, a few high-return period historical severe storms, whose return period significantly exceed the length of ESP historical precipitation

series, ought to be specially treated. A better way is to analyze their return period with longer hydrologic records, instead of within the ESP historical precipitation series, otherwise the forecast result is likely to be on the high side.

3.2 Conditional Probability

The conditional probability method (Wang 2008) is used in the mean monthly discharge prediction from late 1990s, originally in spring low-flow prediction for agriculture irrigation in Songhua River basin northeast China. Conditional probability means future runoff probability under a fixed initial condition. It is desired by the use of initial condition to reduce the mean quadratic error of probability distribution of the predicted monthly runoff, and accuracy of prediction can thus be improved.

Discharge of the last day of current month is taken as initial condition to predict next mean monthly discharge. The worthiness of conditional probability method is determined by the correlation between initial condition and next month runoff. Generally, large basins have longer memories, with good correlations between initial condition and future runoff. Correlation coefficient of discharge of the last day of current month and mean monthly discharge of the next month at selected river sections in major basins are shown in Table 5, where control areas of those sections are over ten thousands or hundred thousand square kilometers, and rivers are less regulated by projects such as dams and gats. In low-flow season, correlation of most sections are strong in part of northern region of Songhua River and Yellow River, it is even better in summer rain season, especially in Songhua basin.

Two-dimensional log normal probability distribution function is used for conditional probability study. This function usually has been used for hydrological frequency analysis. It is a two-dimensional analytic function from which conditional probability distribution can be derived.

From x and y joint probability distribution $f(x, y)$, marginal distribution of x and y can be derived

$$f_x(x) = \int_{-\infty}^{\infty} f(x, y) dy \quad (1)$$

$$f_y(y) = \int_{-\infty}^{\infty} f(x, y) dx \quad (2)$$

Also conditional probability distribution of y at $x = x^*$, $f(y | x^*)$ can be derived

$$f(y | x^*) = \frac{f(x^*, y)}{f_x(x^*)} \quad (3)$$

Two-dimensional log normal probability distribution function is

Table 5 Correlation coefficient between discharge of the last day of current month and mean monthly discharge of the next month

Month	Songhua River basin			Yellow River basin			Huaihe River basin			Yangtze River basin			Pearl River basin	
	Jiangqiao	Haerbin	Jiamusi	Tongguan	Hishiguan	Huaxian	Lutaizzi	Cuntan	Xiangtan	Waizhou	Wuzhou			
1	0.97	0.75	0.93	0.54	0.96	0.94	0.64	0.87	0.84	0.83	0.83			
2	0.95	0.92	0.96	0.72	0.93	0.90	0.63	0.84	0.58	0.59	0.62			
3	0.78	0.84	0.95	0.67	0.88	0.85	0.57	0.70	0.63	0.73	0.90			
4	0.70	0.68	0.67	0.11	0.54	0.56	0.54	0.56	0.21	0.46	0.50			
5	0.69	0.79	0.90	0.82	0.43	0.42	0.80	0.24	0.45	0.48	0.48			
6	0.84	0.85	0.71	0.89	0.50	0.25	0.55	0.31	0.32	0.40	0.56			
7	0.78	0.83	0.74	0.63	0.16	0.35	0.57	0.43	0.59	0.68	0.58			
8	0.82	0.85	0.80	0.68	0.47	0.19	0.67	0.38	0.59	0.30	0.32			
9	0.76	0.90	0.93	0.86	0.63	0.30	0.79	0.49	0.78	0.86	0.78			
10	0.85	0.90	0.93	0.90	0.61	0.89	0.74	0.55	0.60	0.62	0.47			
11	0.97	0.96	0.95	0.95	0.88	0.81	0.62	0.78	0.56	0.62	0.83			
12	0.96	0.90	0.70	0.79	0.98	0.92	0.91	0.79	0.74	0.82	0.82			

$$f(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(\ln x - a_x)^2}{2\sigma_x^2} - 2\rho\frac{(\ln x - a_x)(\ln y - a_y)}{\sigma_x\sigma_y} + \frac{(\ln y - a_y)^2}{2\sigma_y^2}\right]} \quad (4)$$

where ρ is correlation coefficient between $\ln x$ and $\ln y$; a_x is mean $\ln x$; σ_x is mean quadratic error of $\ln x$; a_y is mean $\ln y$; σ_y is mean quadratic error of $\ln y$.

From Eqs. 1, 2, and 4, marginal distribution of x and y can be derived

$$f_x(x) = \frac{1}{\sigma_x\sqrt{2\pi}} e^{-\frac{(\ln x - a_x)^2}{2\sigma_x^2}} \quad (5)$$

$$f_y(y) = \frac{1}{\sigma_y\sqrt{2\pi}} e^{-\frac{(\ln y - a_y)^2}{2\sigma_y^2}} \quad (6)$$

Marginal distribution functions of x and y , Eqs. 5 and 6, are one dimensional Log normal distribution function of x and y .

From Eqs. 3, 4, and 5, the following conditional Log normal probability distribution function can be derived:

$$f(y|x=x^*) = \frac{1}{\sigma_y\sqrt{2\pi(1-\rho^2)}} e^{-\frac{1}{2\sigma_y^2(1-\rho^2)}[\ln y - a_y - \rho\frac{\sigma_y}{\sigma_x}(\ln x^* - a_x)]^2} \quad (7)$$

Equation 7 is still a Log normal distribution function, the mathematical expectation value of $\ln y$ is $a_y + \rho\frac{\sigma_y}{\sigma_x}(\ln x^* - a_x)$; mean quadratic error of $\ln y$ is $\sigma_y\sqrt{1-\rho^2}$, where $\rho > 0$; there is correlation between x and y , mean quadratic error $<\sigma_y$, $\rho = 0$ mean quadratic error $= \sigma_y$, the high correlation between x and y , the less mean quadratic error of probability distribution of y . Make the Log normal distribution of x and y , then a_x , σ_x , a_y , σ_y can be obtained, and compute the correlation coefficient ρ , parameters of conditional distribution of y can be fixed.

Figure 5 is the cumulative log normal distribution of discharge of 30 June, Haerbin Songhua River; Fig. 6 is the comparison of conditional and nonconditional cumulative log normal distributions of mean monthly discharge of July, Haerbin Songhua River, where, the mean quadratic error of conditional distribution is less than nonconditional distribution. Observed and forecast mean monthly discharge of July 1999, Haerbin Songhua River, are shown in Table 6.

Conditional probability model is applied mainly in large rivers monthly runoff prediction. Usually, large river basin area is 100,000 km², upstream area is brumal mountainous, and the density of rainfall gauge network cannot satisfy the need of rainfall runoff model. This model is commonly used for such kind of prediction as it does not need precipitation data. In low-flow season year 2005 and 2006, there were severe low water in Pearl River, and sea water surge intruded the downstream river. To safeguard fresh water supply and protect aquatic environment, dam operation was

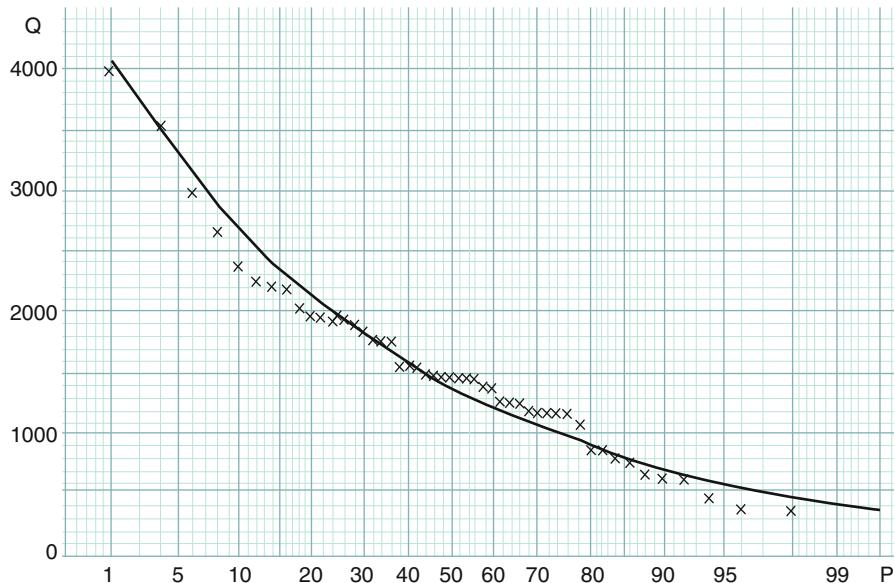


Fig. 5 Cumulative log normal distribution of discharge of 30 June, Haerbin Songhua River

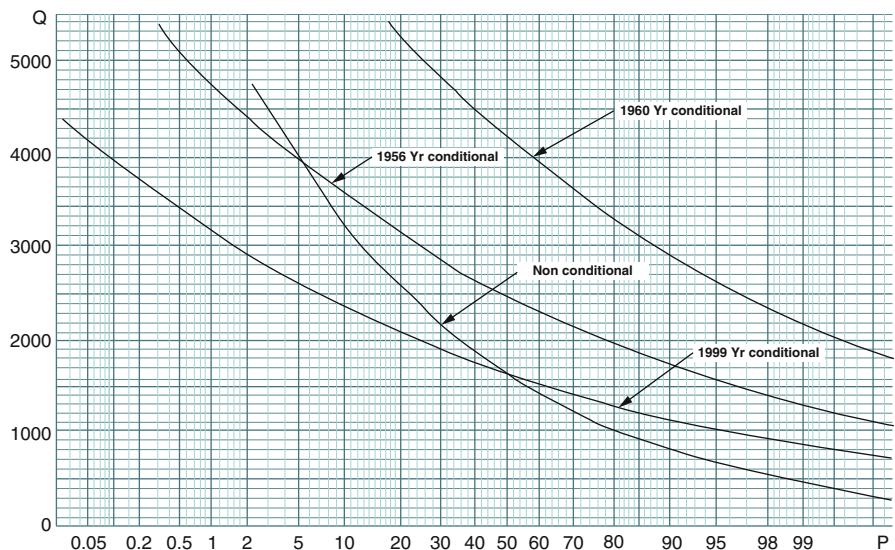


Fig. 6 Conditional cumulative log normal distribution of mean monthly discharge of July, Haerbin Songhua River

Table 6 Observed and forecast mean monthly discharge of July 1999, Haerbin Songhua River

Observed	Expectation	10%	25%	50%	75%	90%
1160	1680	2460	2000	1680	1370	1170

optimized to release water to overwhelm the salty water tide; conditional probability model was used to predict inflow of upstream Pearl River at Wuzhou station, which provided a firm support to the decision-making.

4 Conclusion

Applications of hydrological ensemble predictions have provided strong support to decision-making of flood control and water resources issues in China. In the future, following the implementation of the most strict water resources management policy in China, the hydrological ensemble prediction will receive more attention. Another potential application of ESP is snow melt runoff prediction in northeast China and northwest China, where snow melting flow consists of notable part of the river flow, and is an important part of water resources especially in spring season. Snow melt runoff is mainly decided by initial conditions, i.e., snow amount and base flow, and these initial conditions can be monitored so the prediction will be more reliable and application of the ESP can be more resultful. The problem faced in this case is still the lack of snow gauge data. Application of satellite remote sensing data may be helpful in getting the snow data in future.

References

- H. Bao, Z. Linna, Flood forecast of Huaihe River based on TIGGE ensemble predictions. *J. Hydraul. Eng.* **23**(2), 216–224 (2012) (in Chinese)
- China Water Conservancy and Hydropower Planning and Design Institute, Water resources and its development and utilization assessment in China. China Water & Power Press, Beijing China (in Chinese), (2014)
- A.K. Gobena, T.Y. Gan, Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. *J. Hydrol.* **385**(1–4), 336–352 (2010)
- G. Wang, etc., Applied hydrological forecasting method. China Water Power Press Beijing Chian. (2008) (in Chinese)
- Y. Long, T. Fuqiang, H. Heping, Modified ESP with information on the atmospheric circulation and teleconnection incorporated and its application. *J. Tsinghua Univ. (Sci. Technol.)* **53**(5), 606–612 (2013) (in Chinese)
- Z. Rrenjun, The Xinanjiang model applied in China. *J. Hydrol.* **135**, 371–381 (1992)
- Y. Tao, Q. Duan, A. Ye, etc., An evaluation of post-processed TIGGE multimodel ensemble precipitation forecast in the Huai river basin. *J. Hydrol.* **519**, 2890–2905 (2014)
- L. Yan, H. Jun, W. Jinxin, Application of ensemble streamflow prediction (ESP) to medium-and-long-term water resources prediction. *J. China Hydrol.* **28**, 25–28 (2008) (in Chinese)

Part XI

Mathematical and Statistical Fundamentals for Hydrometeorological Ensemble Forecasting



Probability and Statistical Theory for Hydrometeorology

Zengchao Hao, Vijay P. Singh, and Wei Gong

Contents

1	Introduction	1430
2	Random Variable and Probability Distribution	1431
2.1	Random Variables	1431
2.2	Probability Distributions	1432
2.3	Expectations and Moments	1433
3	Total Probability Theorem and Bayes' Theorem	1434
3.1	Multiplication Rule	1434
3.2	Total Probability Rule	1434
3.3	Bayes' Theorem	1434
4	Statistical Dependence and Joint Probability	1435
4.1	Definition of Statistical Dependence and Independence	1435
4.2	Measure of Dependence	1435
4.3	Joint and Conditional Probability	1435
5	Probability Distributions in the Univariate Case	1437
5.1	Parametric Distributions	1437
5.2	Nonparametric Distributions	1439
5.3	Mixed Variables and Distributions	1443
5.4	Parameter Estimation	1444

Z. Hao (✉)

College of Water Sciences, Beijing Normal University, Beijing, China

e-mail: haozc@bnu.edu.cn

V. P. Singh

Department of Biological and Agricultural Engineering and Zachry Department of Civil Engineering, Texas A&M University, College Station, TX, USA

e-mail: vsingh@tamu.edu

W. Gong

State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China

Institute of Land Surface System and Sustainable Development, Faculty of Geographical Science, Beijing Normal University, Beijing, China

e-mail: gongwei2012@bnu.edu.cn

5.5	Random Number Generation	1445
6	Probability Distributions in the Multivariate Case	1446
6.1	Methods for Constructing Joint Distributions	1446
6.2	Multivariate Parametric Distributions	1446
6.3	Copulas	1448
7	Hypothesis Testing	1450
8	Bayesian Inference	1451
9	Hydrometeorological Forecasting with Statistical Methods	1452
9.1	Linear Methods	1453
9.2	Nonlinear Models	1454
9.3	Ensemble Forecast	1456
10	Summary	1457
	References	1458

Abstract

The hydrometeorological forecasting plays an important role in water resources planning and management. The fundamentals of probability and statistics in hydrometeorology are reviewed in this chapter to aid the forecasting practices. We first introduce the elements of probability theory in the classical statistics, including random variables, probability distribution, joint probability, and total probability theorem. The probability estimation is among the key topics in hydrometeorology for statistical inferences, such as uncertainty analysis and statistical forecasting. The probability inference in the univariate case is first introduced with different methods, including parametric distribution, nonparametric distribution, and mixed distribution. Many hydroclimatic variables are mutually correlated, and the dependence modeling of multivariate random variables through the construction of the joint distribution is then introduced. Most of the context is introduced from the view of classical statistics, while a preliminary introduction of the Bayesian approach is also provided. At last, some commonly used methods for hydrometeorological forecasting are introduced, along with a short summary of this chapter.

Keywords

Probability · Statistics · Distribution · Forecasting

1 Introduction

Hydrometeorological forecasting of multiple variables, such as precipitation, temperature, soil moisture, and river discharge, at different scales plays an important role in water resource management and early warning of nature hazards, such as flood and drought, to reduce potential impacts. Short-term hydrometeorological forecasting (e.g., hourly and daily) is of critical importance for operational flood forecasting systems, while long-term forecasting (e.g., monthly or seasonal) is critical for reservoir operations and drought early warning, among other applications. For example, drought often caused huge losses to agriculture, energy, and

even famine, and substantial efforts have been devoted to the seasonal forecasting of drought to establish early warning systems to enhance our capacity to reduce the impacts of drought (Pozzi et al. 2013; Wood et al. 2015). Improving the skill of hydrometeorological forecasting is important to enhance the capacity to cope with natural hazards.

Broadly speaking, hydrometeorological forecasting can be classified into two types of approaches. The first type of approach is the dynamical forecasting based on general circulation models (GCMs) of the weather or climatic systems (Kirtman et al. 2014; Yuan et al. 2015a) or rainfall-runoff modeling in hydrological systems (Beven 2011; McEnery et al. 2005; Singh and Woolhiser 2002). For example, hydrological models based on rainfall-runoff modeling have been widely used for operational streamflow or flood forecasting to aid water resources planning and management (Bourdin et al. 2012; Cloke and Pappenberger 2009; Thielen et al. 2008). The second type of approach is the statistical forecasting based on statistical relationships of hydrometeorological observations without explicitly considering the relationships in the underlying systems or processes. A variety of statistical models have been used for hydroclimatic forecasting, such as regression models, autoregressive moving average models (ARMA), and artificial neural networks (ANN). Moreover, the statistical methods also play an important role in the post-processing of forecasts from the short-term weather forecast or long-term seasonal forecast of hydroclimatic variables. For example, Bayesian model averaging (Duan et al. 2007; Hoeting et al. 1999; Luo et al. 2007; Raftery et al. 2005; Yuan and Wood 2012) has been widely used to quantify the uncertainty of the drought or streamflow forecasts obtained from multiple GCMs or multiple hydrologic models.

The objective of this chapter is to provide an introduction of the fundamentals of probability and statistics for hydrometeorological forecasting. Due to the vast majority of materials from the probability and statistics, the introduction of this chapter is confined to the fundamentals of the concepts and tools that are closely related to hydrometeorological forecasting. Some commonly used methods, such as the regression model and ensemble forecast, in recent decades for hydrometeorological forecasting are introduced at the end of this chapter.

2 Random Variable and Probability Distribution

2.1 Random Variables

In probability and statistics, a random variable is a real value function with specific values in the sample space. In hydrology or climatology, certain characteristics of rainfall, temperature or streamflow value, rainfall duration, or drought can be regarded as a random variable. Specific examples of random variables are annual extreme rainfall for a given duration, number of rainfall events in a year, annual peak discharge, drought severity, drought duration, number of floods occurring in

a year, number of days with maximum temperature exceeding a given value in a year, and so on. By convention, a *random variable* (*r. v.*) is usually denoted by an upper case letter (e.g., X , Y , or Z), while its observed values (or realizations) are denoted by a lower case letter (e.g., x , y , and z). A *discrete* random variable takes on a finite number of values constituting a finite set, while a *continuous* random variable takes on an infinite number of values constituting a continuous or infinite set.

2.2 Probability Distributions

A random variable is commonly characterized by probability density and distribution functions. Accordingly, these functions are either discrete or continuous. For example, a continuous variable X has a *probability density function* (PDF), typically written as $f(x)$ and a *cumulative distribution function* (CDF) $F(x)$. The CDF is defined as the probability of the value of X less than or equal to a specific value x , which can be expressed as:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(s)ds \quad (1)$$

The PDF of a continuous random variable X is the derivative of the CDF that can be expressed as:

$$f(x) = \frac{dF(x)}{dx} \quad (2)$$

which satisfies the property:

$$\int_{-\infty}^{\infty} f(x)dx = 1 \quad (3)$$

In a similar vein, the probability density function of a real-valued discrete random variable X is called *probability mass function* (PMF). X takes on a specific set of values x_1, x_2, \dots with each value x_i having a given probability p_i of occurrence. Specifically, the PMF of X for the value x_i can be expressed as:

$$p_X(x_i) = P(X = x_i) \quad (4)$$

The cumulative distribution function $F_X(x)$ is defined as the sum of probabilities of x_i less than or equal to a specific value x and can be expressed as:

$$F_X(x) = \sum_{x_i \leq x} p_X(x_i) \quad (5)$$

2.3 Expectations and Moments

The expectation of a function of a random variable represents the average value of the function of the random variable. Based on the PDF $f(x)$ of a continuous random variable X , the expectation of a function $g(x)$ can be defined as:

$$E[g(x)] = \int_{-\infty}^{\infty} g(x)f(x)dx \quad (6)$$

The distribution function of a random variable describes the whole behavior of a random variable. In many cases, it is useful to define certain measures to describe the properties of a random variable, such as the central tendency or range. The moments are commonly used measures to describe the properties of the distribution of a random variable, such as the shape, center, spread, skewness, and peakedness. The two commonly used moments of a distribution are the mean μ and variance σ^2 (σ is generally called standard deviation), which can be expressed as:

$$\mu = E(X) = \int_{-\infty}^{\infty} xf(x)dx \quad (7)$$

$$\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx \quad (8)$$

In reality, the probability distribution function of a random variable is unknown. Instead, parts of its realization are known, based on a set of observations $[x_1, x_2, \dots, x_n]$. The mean and variance can be estimated based on these sample observations. The sample estimate of the mean (or sample mean, denoted as \bar{x}) and of the variance (or sample variance S_x^2) can be expressed as:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (9)$$

$$S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{X})^2 \quad (10)$$

The mean is the first moment about the origin and other moments are generally defined based on the mean. The mean in Eq. 9 is generally referred to as the “arithmetic mean.” The harmonic mean and geometric mean are also used in certain cases. The variance is the second moment about the arithmetic mean. The third moment and fourth moment, which measure the asymmetry and peakedness of a distribution, respectively, are also commonly used.

3 Total Probability Theorem and Bayes' Theorem

3.1 Multiplication Rule

The conditional probability provides a general expression of the probability of the intersection of two events, which is commonly referred to as the “multiplication rule” of probabilities. For two events A and B , the probability of the event $A \cap B$ can be expressed as:

$$P(AB) = P(B|A)P(A) = P(A|B)P(B) \quad (11)$$

3.2 Total Probability Rule

For two mutually exclusive events B and B' with $P(B) + P(B') = 1$ (i.e., the union is the sample space), the totally probability of an event A can be expressed as::

$$P(A) = P(A|B)P(B) + P(A|B')P(B') \quad (12)$$

where B' is the complement of B . The total probability rule states that given a collection of events that are mutually exclusive, the probability of an event can be written as the sum of probabilities of the intersection of the event with elements of the collection.

3.3 Bayes' Theorem

The Bayes theorem can be computed based on the multiplication rule and total probability rule defined above. The conditional probability of event B given event A can be expressed as:

$$P(B|A) = \frac{P(AB)}{P(A)} \quad (13)$$

The probability $P(AB)$ can be replaced with the multiplication rule, and the probability $P(A)$ can be replaced with the total probability rule. Accordingly, the probability of $P(B|A)$ can be expressed as:

$$P(B|A) = \frac{P(AB)}{P(A)} = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B')P(B')} \quad (14)$$

Equation 14 is generally called Bayes' theorem. When $P(A|B)$ is known, the Bayes theorem can be used to compute the conditional probability $P(B|A)$. The Bayes theorem can be used to compute the conditional probability when the probability has to be computed indirectly, which is commonly used to update the probability that is consistent with the new information.

4 Statistical Dependence and Joint Probability

4.1 Definition of Statistical Dependence and Independence

The statistical independence of two events implies that the occurrence of event B conveys no information on the occurrence of event A . Two events A and B are statistically independent if and only if the joint probability is the product of the probability of each event, i.e.:

$$P(AB) = P(A)P(B) \quad (15)$$

Two events are dependent if the occurrence of event B affects the occurrence of event A . For two dependent events, the joint probability $P(AB)$ and conditional probability $P(B|A)$ or $P(A|B)$ are commonly used to describe the two events A and B .

4.2 Measure of Dependence

The covariance and correlation are commonly used to measure the dependence between random variables. The covariance of two random variables X and Y is expressed as:

$$\text{cov}(X,Y) = E([X - E(X)][Y - E(Y)]) \quad (16)$$

and the correlation ρ is defined as:

$$\rho = \frac{E([X - E(X)][Y - E(Y)])}{\sigma_x \sigma_y} \quad (17)$$

The correlation in the equation above is generally called Pearson product-moment correlation coefficient that measures the linear dependence between random variables.

Other dependence measures, such as the Spearman or Kendall rank correlation, are also commonly used to measure the nonlinear dependence (or rank correlation) between random variables.

4.3 Joint and Conditional Probability

When a statistical characterization of a random vector (X_1, X_2, \dots, X_n) is of particular interest, the joint or multivariate distribution is generally used for modeling the dependence structure of the random vector. Since random variables may be dependent on each other, it is commonly needed to describe the

dependence among random variables (or random vector). The joint and conditional distribution function is commonly used for describing the joint behavior of multivariate random variables or random vectors, which are introduced as follows (Loucks et al. 2005).

For two continuous random variables X and Y , the joint cumulative distribution function $F_{XY}(x, y)$ can be expressed as:

$$F_{XY}(x, y) = P(X \leq x, Y \leq y) = \int_{-\infty}^y \int_{-\infty}^x f_{XY}(x, y) dx dy \quad (18)$$

where $f_{XY}(x, y)$ is the joint PDF of random variables X and Y .

The marginal distribution function and conditional distribution function are two related concepts to describe random variables. For two continuous random variables X and Y with the joint PDF $f_{XY}(x, y)$, the marginal PDF of Y can be expressed as:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx \quad (19)$$

The conditional distribution $F_{X|Y}(x, y)$ of X given a specific value of y can be defined as:

$$F_{X|Y}(x, y) = P(X \leq x | Y = y) = \int_{-\infty}^x f_{x|y}(s | y) ds \quad (20)$$

where $f_{x|y}(s | y)$ is the conditional probability density function that can be expressed as:

$$f_{x|y}(x | y) = \frac{f_{XY}(x, y)}{f_Y(y)} \quad (21)$$

For discrete random variables X and Y , the joint distribution function $F_{XY}(x, y)$ can be expressed as:

$$F_{XY}(x, y) = \sum_{s \leq x} \sum_{t \leq y} p_{XY}(s, t) \quad (22)$$

where p_{XY} is the joint PMF of random variables X and Y .

The marginal PMF $p_Y(y)$ can be expressed as:

$$p_Y(y) = \sum_x p_{XY}(x, y) \quad (23)$$

Accordingly, the conditional PMF of X given Y can be expressed as:

$$p_{X|Y}(x|y) = \frac{p_{XY}(x,y)}{p_Y(y)} \quad (24)$$

5 Probability Distributions in the Univariate Case

Probability distributions are usually used to describe hydroclimatic data. Depending on the nature of data, several types of distributions, including parametric distributions, nonparametric distributions, and mixed distributions, are used. In this section, different types of probability distributions in the univariate case are introduced, along with the parameter estimation and random number generation.

5.1 Parametric Distributions

Gaussian Distribution

The first type of probability distribution is the continuous parametric distribution, which can be used to describe variables, such as annual streamflow. The Gaussian (or normal) distribution plays a central role in classical statistics and is particularly useful because of the central limit theorem (CLT), which states that, given certain conditions, the arithmetic mean (or sum) of independent identically distributed random variables will have a normal distribution when the sample size is sufficiently large (Wilks 2011).

For a random variable X , the normal probability density function (PDF) can be expressed as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] \quad (25)$$

where μ is the mean and σ^2 is the variance.

A typical figure of the PDF of the normal distribution is shown in Fig. 1 with different parameters. The normal distribution is a symmetric distribution with a single central peak at the mean of the data (well known as “bell-shaped”). Parameter μ controls the location of the center of the distribution (compare Fig. 1a and Fig. 1b), while parameter σ controls the spread of the distribution (compare Fig. 1c and Fig. 1d).

The distribution of certain hydroclimatic variables is generally nonsymmetric and skewed, which generally occurs when there is a physical limit of the variable (e.g., precipitation is nonnegative). Since the normal distribution is symmetric, it may not be suitable for describing the distribution of some variables, such as daily precipitation or streamflow. A variety of other parametric distributions have been used to

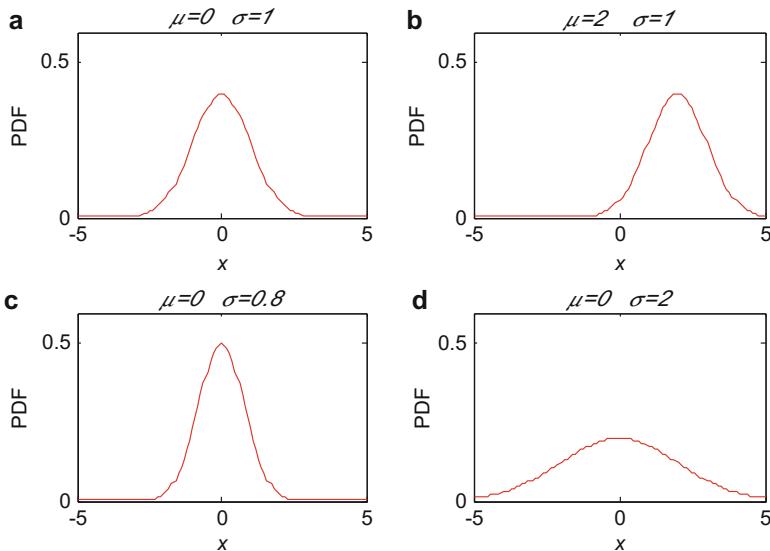


Fig. 1 Normal distribution with different parameters μ and σ . (a) $\mu = 0, \sigma = 1$; (b) $\mu = 2, \sigma = 1$; (c) $\mu = 0, \sigma = 0.8$; (d) $\mu = 0, \sigma = 2$

describe these hydroclimatic data, such as lognormal, Gumbel, Weibull, and gamma distribution. These distributions exhibit different types of PDF, thereby allowing for modeling different properties of the data, such as skewness.

Gamma Distribution

The gamma distribution is commonly used due to the advantage that it is defined only for positive values, since hydrological variables, such as rainfall and runoff, are generally positive. The gamma distribution can be expressed as:

$$f(x) = \frac{(x/\beta)^{\alpha-1} \exp(-x/\beta)}{\beta \Gamma(\alpha)} \quad x > 0 \quad \alpha > 0, \quad \beta > 0 \quad (26)$$

where α and β are the shape and scale parameters, respectively, and $\Gamma(\alpha)$ is a complete gamma function.

A typical figure of the PDF of the gamma distribution is shown in Fig. 2. For the location parameter $\alpha < 1$, the PDF of the gamma distribution is highly skewed (skewness = $2\alpha^{1/2}$). A larger α results in less skewed distribution (Fig. 2a) and the distribution function approaches the normal distribution with very large values of α (Fig. 2b). The scale parameter β stretches or squeezes the PDF curve (Fig. 2c, d).

The gamma distribution includes two distributions as special cases, the exponential distribution and the chi-square distribution. When the shape parameter $\alpha = 1$, the gamma distribution reduces to the exponential distribution, i.e.:

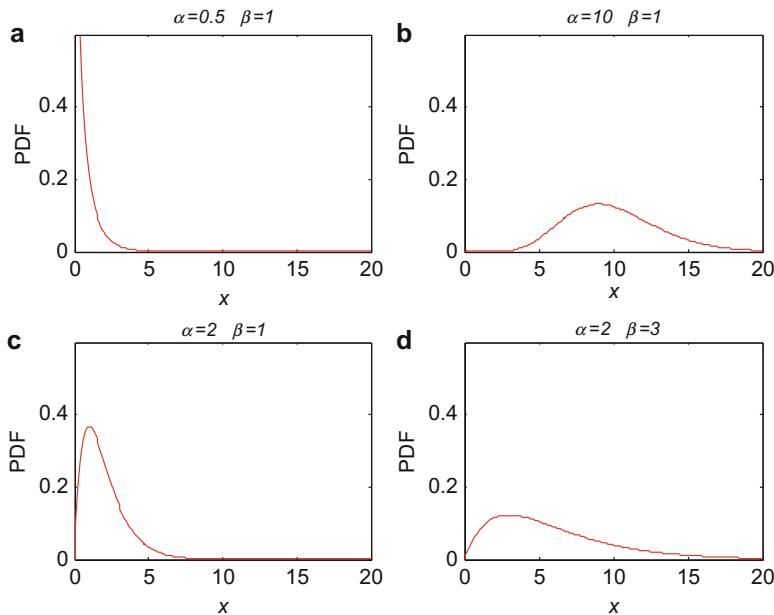


Fig. 2 Gamma distribution with different values of the location parameters α and scale parameter β . (a) $\alpha = 0.5$, $\beta = 1$; (b) $\alpha = 10$, $\beta = 1$; (c) $\alpha = 2$, $\beta = 1$; (d) $\alpha = 2$, $\beta = 3$

$$f(x) = \frac{\exp(-x/\beta)}{\beta} \quad x > 0 \quad (27)$$

When the scale parameter $\beta = 2$, the gamma distribution reduces to the chi-square distribution, i.e.:

$$f(x) = \frac{(x/2)^{\nu/2-1} \exp(-x/2)}{2^{\nu/2} \Gamma(\nu/2)} \quad x > 0 \quad (28)$$

where the integer parameter ν ($=2\alpha$) is generally called the *degree of freedom*. The chi-square distribution is commonly used in the context of statistical testing.

5.2 Nonparametric Distributions

The parametric distributions suffer from drawbacks in that assumptions about the distribution forms of data have to be made. In reality hydroclimatic variables may exhibit rich features, such as bimodality, skewness, and tail property, which may not be well captured by prescribed function forms. Hence, nonparametric distributions have been commonly used for modeling rich features of hydroclimatic data due to the advantage that they allow the data to speak for themselves. In other words, the

nonparametric methods attempt to estimate the density function directly from the sample without assuming a particular form for the underlying distribution.

Histogram

The simplest form for nonparametric density estimation is the histogram, which gives a rough visualization of the density of the underlying distribution of the data at hand. For a specific sample, the range of data is divided into a number of bins (or intervals) and the number of the data points falling into each bin is then counted. Usually bins with equal width are widely used, though varying bin width may be used in certain cases. In the case with bins of equal size, the height of each bar of the histogram is proportional to the frequency (or the number of data points in each bin). It is generally convenient to normalize the histogram by changing the vertical axis values (i.e., frequency) without changing the shape, which can be conducted in different ways. The frequency of the histogram can be converted to the probability by dividing each frequency by the total sample size, resulting in a discrete probability distribution of the data. The discrete probability can be converted to a probability density distribution by dividing each probability by the equal bin width Δx (in this case the total area of the histogram is 1).

Here a sample of size 1000 from the mixture of two normal distributions (with means of -1 and 3 , respectively) is used as the synthetic data to illustrate the construction of the histogram. An important issue in constructing the histogram is the bin width (or number of bins). While the interval of the bin is too wide, details of the distribution properties of the data may be masked. Some efforts have been devoted to determining an optimal number of bins, such as the square root of the number of observations (Montgomery and Runger 2010). The histogram with the frequency on the y-axis is shown in Fig. 3a, for which the bin number is 32 (\approx square root of 1000). The histogram with the approximated density function with different numbers of bins (10, 30, 60) is shown in Fig. 3b-d, from which generally the bimodal distribution of the data can be shown.

There are several drawbacks associated with the histogram, such as the inefficiency in the high dimension (the bin number increases exponentially with the dimension). As such, the histogram is commonly used for simple visualization of the density property for one or two dimension problems.

Kernel Density Estimation

An alternative approach to fit the probability density function of hydroclimatic variables is the kernel density estimation, which can be regarded as the extension of the histogram and can be applied in high dimensions (Hao and Singh 2016; Lall 1995). The kernel density estimation does not assume a specific form of the distribution of the data but estimates the density function directly from the data based on smoothing with kernels on each of the data points. The potential drawbacks of this method is that large bias may occur in density estimation near the end points of the support of the density or boundaries (i.e., or the boundary effect).

For a random variable X with n observations x_1, x_2, \dots, x_n , the kernel density estimate of the probability density function $f(x)$ can be expressed as (Silverman 1986):

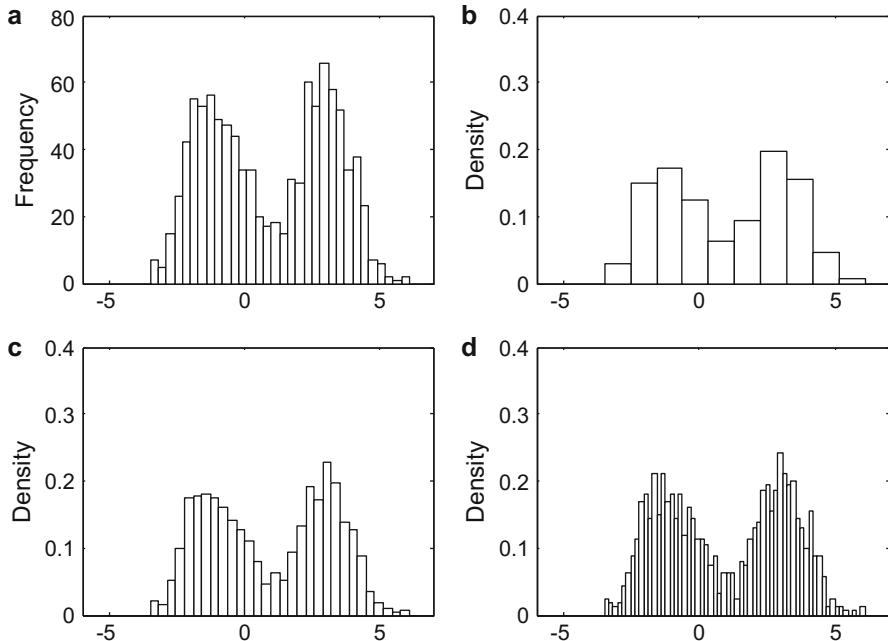


Fig. 3 Histogram of the synthetic data generated from two normal distributions. (a) y-axis is the frequency with bin number 32; (b) y-axis is the density with bin number 10; (c) y-axis is the density with bin number 30; (d) y-axis is the density with bin number 60

$$f(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x-x_i}{h}\right) \quad (29)$$

where h is the bandwidth (or smoothing parameter) and $K(\cdot)$ is a kernel function.

The kernel functions are nonnegative with unit area, i.e.:

$$\int K(x)dx = 1 \quad (30)$$

A range of kernel functions can be used to construct the kernel density functions, such as uniform, quadratic, quartic, normal, and others. For example, the Gaussian kernel can be expressed as:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right) \quad (31)$$

There are two parameters in the kernel density estimation, including the kernel K and the bandwidth h . The estimation of the optimal bandwidth h is generally more important than the choice of K (Sivakumar and Berndtsson 2010; Wilks 2011). The role of the smoothing parameter h is similar to the bin width of the histogram. A

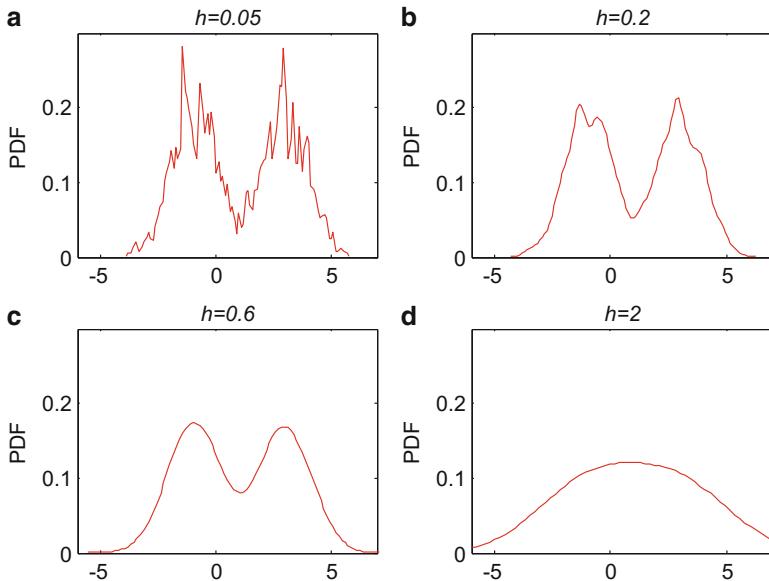


Fig. 4 Kernel density estimates for the synthetic data generated from two normal distributions with different bandwidth h with the Gaussian kernel. **(a)** $h = 0.1$; **(b)** $h = 0.2$; **(c)** $h = 0.6$; **(d)** $h = 2$

relatively large value of the smoothing parameter h may mask the details of density. To illustrate the property of the nonparametric density estimation, the sample from the mixture of two normal distributions is also used here to illustrate the property of kernel density estimation with different bandwidths h , as shown in Fig. 4. Here the Gaussian kernel is used to estimate the density. Generally, the kernel density estimation with the bandwidth $h = 0.05, 0.2$, or 0.6 clearly reveals the bimodal property of the data (Fig. 4a–c). The $h = 0.05$ seems to be too small, since there are too many spurious features of data (often referred to as under-smoothed). Meanwhile, the bandwidth $h = 2$ is too wide, since much of the underlying feature of the data, such as the bimodal property, is masked, as shown in Fig. 4d.

It can be seen that the bandwidth parameter is critical in the kernel density estimation. As such, substantial efforts have been devoted to the optimal selection of the bandwidth h , including rules of thumb, least squares cross-validation, and biased cross-validation (Jones et al. 1996; Wand and Jones 1994). The commonly used criteria for selecting the optimal bandwidth is based on the mean integrated squared error (MISE), which can be expressed as (Silverman 1986):

$$\text{MISE}(h) = E \int [f'(x) - f(x)]^2 dx \quad (32)$$

where $f'(x)$ is the estimator of density function $f(x)$ and E is the expected value with respect to samples.

When a Gaussian kernel is used, the optimal bandwidth h_{opt} can be obtained by minimizing the MISE as:

$$h_{\text{opt}} = 1.06 \sigma n^{-1/5} \quad (33)$$

where σ is the standard deviation of the sample and n is the length of the sample. For example, for the synthetic data used in plotting Fig. 4, the optimal bandwidth is estimated as 0.60.

Empirical Cumulative Distribution Function

The empirical cumulative distribution function is closely related to the histogram. For a real-valued variable X with observations x_1, x_2, \dots, x_n , the estimation of probability is based on the rank of data or order statistics (order statistics are defined by sorting the realizations of the random variables in increasing order). For many cases, the empirical cumulative distribution function of the i th order statistics $x_{(i)}$ can be expressed as:

$$P(x_{(i)}) = \frac{i + a}{n + b} \quad (34)$$

where n is the number of the sample values and a and b are constants.

In hydrology, the empirical distribution function is generally referred to a plotting position formula. Different values of constants a and b result in different empirical distribution functions. For example, when $a = 0$ and $b = 1$, the empirical distribution is referred to as the Weibull plotting position formula, while when $a = -0.44$ and $b = 0.12$, it leads to the Gringorten plotting position formula (Fuglem et al. 2013; Gringorten 1963; Hao and Singh 2013b; Makkonen 2006).

5.3 Mixed Variables and Distributions

Certain variables cannot be modeled with the distributions introduced above. For example, for daily precipitation, the occurrence and the amount of precipitation have to be modeled separately. In this case, a mixed discrete-continuous-type distribution with a discrete probability distribution of zero precipitation (precipitation occurrence) and a continuous probability distribution of nonzero precipitation (precipitation amount) has to be used for modeling daily precipitation. To model these mixed discrete-continuous variables, the mixed distribution has to be used:

$$f(x) = (1 - p)\delta(x) + pg(x) \quad (35)$$

where p is the probability of occurrence of the hydrologic event (e.g., precipitation) and δ is the one-dimensional Dirac delta function, which becomes infinity when x is 0 and becomes 0 otherwise. The mixed distribution is commonly used to model both the occurrence and intensity of hydroclimatic variables with both discrete and continuous components.

5.4 Parameter Estimation

Suppose samples are drawn from a population distribution $f(x|\theta)$ described by a parameter θ . The problem of parameter estimation seeks to find a good estimator of the parameter θ (often referred to as the point estimator) based on samples. There are several methods that can be used to estimate the parameter of a specific distribution, such as the methods of moment (MOM), maximum likelihood estimation (MLE), or probability weighted moments (PWM).

Method of Moment

The main idea of the method of moments for the parameter estimation is to equate the moment from the population distribution to the moment from samples. The moments of the population are expressed based on the expectations, which are a function of the unknown parameter θ . Specifically, the estimator of moments can be obtained by equating the first m moments of the population distribution to the first m sample moments, from which the unknown parameter can be obtained by solving the resulting equations.

Suppose $X_1, X_2 \dots, X_n$ are samples from a population of normal distribution $f(x|\mu, \sigma^2)$ with unknown parameters mean μ and variance σ^2 . Define the first two moments from the samples as:

$$m_1 = \frac{1}{n} \sum_{i=1}^n X_i \quad (36)$$

$$m_2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad (37)$$

The first two moments of the population moments can be expressed as:

$$E(X) = \int x \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = \mu \quad (38)$$

$$E(X^2) = \int x^2 \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right] dx = \mu^2 + \sigma^2 \quad (39)$$

Thus, one obtains:

$$\mu = \frac{1}{n} \sum_{i=1}^n X_i = \bar{X} \quad \mu^2 + \sigma^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 \quad (40)$$

By solving Eq. 40, one obtains the moment estimator of the two parameters:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \quad \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \quad (41)$$

Maximum Likelihood Estimation

Before introducing the MLE, the likelihood function of a probability distribution is first introduced. For a random variable X with probability density function $f(x|\theta_1, \dots, \theta_n)$ with the unknown parameters $\theta = (\theta_1, \dots, \theta_n)$, let x_1, x_2, \dots, x_n be realizations of a sample of size n from the population. Then the likelihood function is defined as:

$$L(\boldsymbol{\theta}|\mathbf{x}) = L(\theta_1, \dots, \theta_k | x_1, \dots, x_n) = \prod_{i=1}^n f(x_i | \theta_1, \dots, \theta_k) \quad (42)$$

The corresponding maximum likelihood estimator of the parameters is the one that maximizes the likelihood function $L(\boldsymbol{\theta}|\mathbf{x})$.

For example, suppose X_1, X_2, \dots, X_n are samples from a population of normal distribution $f(x|\mu, \sigma^2)$ with unknown mean μ but known variance σ^2 . The likelihood function can be expressed as:

$$L(\mu | x_1, \dots, x_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x_i - \mu)^2}{2\sigma^2}\right] \quad (43)$$

Taking the derivative of L with respect to μ and setting it to zero (note that this is only the necessary condition for the maximization problem), one obtains:

$$\frac{dL(\mu | x_1, \dots, x_n)}{d\mu} = 0 \quad (44)$$

Solving the equation above leads to the MLE estimator:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i \quad (45)$$

From Eq. 45, the MLE estimator of the mean is the sample mean.

5.5 Random Number Generation

Generating random samples from probabilities distribution functions is commonly used in hydrologic simulations (Hao and Singh 2011; Lall and Sharma 1996; Loucks et al. 2005; Prairie et al. 2006; Sharma et al. 1997; Sivakumar and Berndtsson 2010). Broadly speaking, the methods for generating random numbers from distributions include the direct method and indirect methods. The probability integral transformation plays an important role in this regard. Specifically, for a

random variable X with a continuous cumulative distribution function $F(x)$, the random variable $U = F(x)$ is uniformly distributed on $(0, 1)$ (i.e., $P(U \leq u) = u$) (Casella and Berger 2002). Based on a uniformly distributed variable v , the direct method generally applies to the case when there is a function $g(v)$ such that the distribution of the transformed variable $z = g(v)$ follows the desired distribution. For example, for the exponential distribution, the cumulative distribution function can be expressed as:

$$F(x|\lambda) = \begin{cases} 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (46)$$

Thus, for any $U \sim \text{uniform}(0, 1)$, the variable $z = F^{-1}(U) = -\lambda \log(1-U)$ is an exponentially distributed variable with parameter λ .

A closed form of the function $F^{-1}(U)$ may not exist. In reality, for most cases the indirect methods have to be resorted to, such as the accept/reject algorithm. In this case, a candidate random variable is generated but only accepted when certain conditions are satisfied.

6 Probability Distributions in the Multivariate Case

6.1 Methods for Constructing Joint Distributions

A variety of joint distribution functions to model multiple hydroclimatic variables, such as the parametric distribution, copula, entropy, and nonparametric methods (Balakrishnan and Lai 2009; Hao and Singh 2015, 2016; Kotz et al. 2000) have been developed in the past few decades for modeling the dependence among multivariate random variables, each having its own strengths and limitations. These methods provide different ways of modeling the dependence among different variables and may be selected, based on the property of data at hand. For example, for frequency analysis in hydrology, the nonparametric method generally may not be suitable due to the deficiency of extrapolation in the tail or extreme region. In this section, we mainly focus on the multivariate parametric distribution function and the copula.

6.2 Multivariate Parametric Distributions

The commonly used multivariate parametric distribution function is generally based on the extension of the distribution forms in the univariate case, such as normal, gamma, or exponential distribution. The commonly used multivariate parametric distributions include the bivariate (multivariate) normal distribution (Goel et al. 1998; Nadarajah 2007; Sackl and Bergmann 1987; Yue 1999), bivariate t distribution (Ghizzoni et al. 2010; Shaw and Lee 2008), bivariate lognormal distribution (Yue 2000), bivariate exponential distribution (Bacchi et al. 1994; Choulakian

et al. 1990; Singh and Singh 1991), bivariate gamma distribution (Yue et al. 2001), bivariate Pearson distribution, and bivariate meta-Gaussian distribution (Kelly and Krzysztofowicz 1997). Among the multivariate parametric distributions, the multivariate normal (MVN) distribution is the most commonly used in classical statistics, which is easy to implement and flexible to be extended to higher dimensions.

The bivariate Gaussian distribution function of two random variables X and Y can be expressed as (Wilks 2011):

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)} \left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - 2\rho \frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} \right] \right\} \quad (47)$$

where μ_x and μ_y are the mean values of random variables X and Y , respectively; σ_x and σ_y are the standard deviations of X and Y , respectively; and ρ is the correlation.

There are totally five parameters of the bivariate normal distribution, including μ_x , μ_y , σ_x , σ_y , and ρ , which can be simply estimated based on samples. For example, the correlation is essentially the Pearson product-moment correlation coefficient. The marginal distributions of X and Y are also Gaussian distributions. Note that generally the normal distribution of the individual variable X and Y does not imply that their joint distribution is bivariate normal distribution.

Though hydroclimatic variables may not be Gaussian distributed, there are different techniques that can be used to transform the non-Gaussian data to be approximately Gaussian distributed, such as power transformation, log transformation, trigonometric transformation, and normal quantile transformation (NQT) (Bogner et al. 2012; Kelly and Krzysztofowicz 1997; Montanari and Brath 2004; Wilks 2011). The multivariate normal distributions (MVN) possess interesting distribution properties, such as that the conditional distribution of MVN random variables is given other subsets and the distribution of the sum of MVN random variables is still normal distributions. We show these properties in the bivariate case with a bivariate normal distribution $f(x, y)$.

First, a linear combination of $aX + bY$ is still the normal distribution, i.e.:

$$aX + bY \sim N(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2) \quad (48)$$

The property that the sum of normal variables is normally distributed results from the central limit theorem (CLT) introduced before.

In addition, the conditional distribution of Y given $X = x$ is also normally distributed, which can be expressed as:

$$Y | X \sim N\left(\mu_y + \rho \frac{\sigma_y}{\sigma_x}(X - \mu_x), \sigma_y^2(1 - \rho^2)\right) \quad (49)$$

For the bivariate case, the conditional mean of Y given X depends on the mean of X , but the conditional variance does not depend on X .

6.3 Copulas

As stated before, the joint probability distribution function is required to characterize the joint behavior of two or more events that are statistically dependent. In hydroeteorology, the past decade has witnessed a flurry of developments and applications of copula methods to model multiple variables, such as precipitation, temperature, streamflow, or flood/drought properties, at single or multiple locations.

The advantage of copula in constructing the joint distribution of random variables is that the modeling of dependence is independent of the marginal distributions. Due to the flexible property in dependence modeling of multivariate random variables, it has been widely used in a wide array of applications, such as frequency analysis (Favre et al. 2004; Hao and Singh 2013a; Salvadori and De Michele 2010), streamflow or rainfall simulation (Hao and Singh 2013c, 2015; Li et al. 2013), downscaling (Laux et al. 2011; Van den Berg et al. 2011; Verhoest et al. 2015), and bias correction (Piani and Haerter 2012; Vogl et al. 2012).

Assume $F_X(x)$ (denoted as U) and $F_Y(y)$ (denoted as V) the marginal distributions of a continuous random vector (X, Y) . The copula of random vector (X, Y) can be expressed as (Genest and Favre 2007; Nelsen 2006):

$$P(X \leq x, Y \leq y) = C[F_X(x), F_Y(y); \theta] \quad (50)$$

where θ is the parameter of copula. The copula C maps the marginal into the joint distribution. A variety of copula families, such as meta-elliptical copula (Gaussian, t), Archimedean copula (Clayton, Frank, Gumbel), extreme-value copula, entropy copula, and vine copula, can be used to construct the copula C . The one-parameter Archimedean copulas are among the most commonly used copulas, for which the Clayton copula, Frank copula and Gumbel copula are expressed as:

$$\text{Clayton} \quad C(u,v) = (u^{-\theta} + v^{-\theta} - 1)^{-1/\theta} \quad (51)$$

$$\text{Frank} \quad C(u,v) = -\frac{1}{\theta} \ln \left[1 + \frac{(e^{-\theta u} - 1)(e^{-\theta v} - 1)}{e^{-\theta} - 1} \right] \quad (52)$$

$$\text{Gumbel} \quad C(u,v) = \exp \left\{ - \left[(-\log u)^{-\theta} + (-\log v)^{-\theta} \right]^{-1/\theta} \right\} \quad (53)$$

where θ is the parameter that controls the dependence structure of the marginals.

A copula function suitable for the data at hand has to be selected, which can be performed based on the goodness of fit test statistics, such as the Cramér – von Mises statistic (S_n) and Kolmogorov-Smirnov statistic (T_n) (Genest and Favre 2007). For the estimation of parameter θ , two commonly used methods are the exact maximum

likelihood (EML) method, which estimates parameters of the marginal distribution and copula function simultaneously, and the inference functions for marginal (IFM), for which parameters of marginal distributions and the copula function are estimated separately by maximizing the respective likelihood functions (Joe 1997; Nelsen 2006). In addition, for the bivariate case, the copula parameter can be estimated based on the inversion of Kendall's tau or Spearman's rho (Genest and Favre 2007; Genest et al. 2007).

The commonly used copulas generally differ in modeling the dependence properties of the data. To illustrate this point, random numbers of size 1000 from different copulas are shown in Fig. 5. It can be seen from Fig. 5 that random variables from both the Gaussian and frank copula exhibit symmetric dependence, while those from the Clayton and Gumbel copula exhibit asymmetric dependence (Hao and Singh 2016; Trivedi and Zimmer 2005). It can be seen that the dependence of Clayton copula is strong in the left tail but weak in the right tail. In contrast, the dependence of Gumbel copula is strong in the right tail and relatively weak in the left tail.

Apart from the commonly used parametric copulas, there are other copulas, such as empirical copula, vine copula, entropy copula, and extreme-value copula (Genest and Favre 2007; Hao and Singh 2016). The empirical copula can be employed in modeling the dependence of multivariate random variables

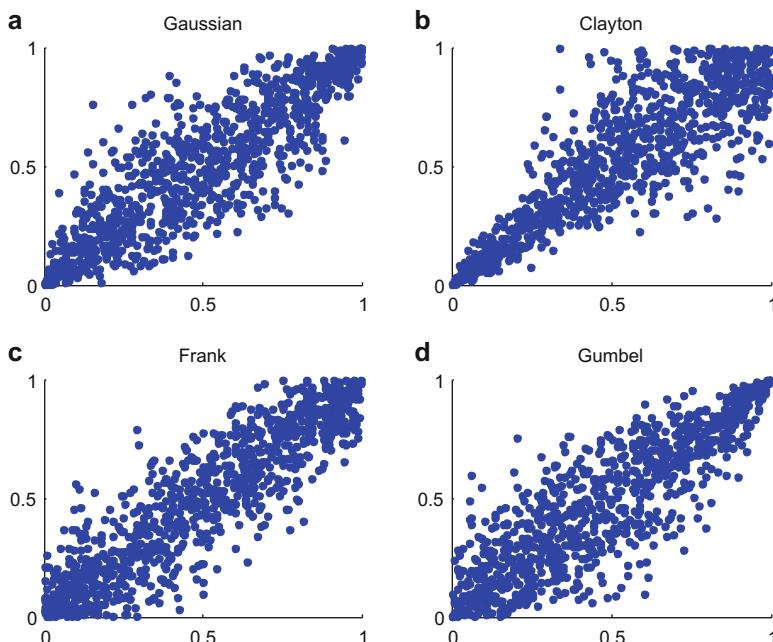


Fig. 5 Scatter plots of random samples from different copulas. (a) Gaussian, (b) Clayton, (c) Frank, (d) Gumbel

without making assumptions about the distribution form of the copula and is attractive in a variety of applications, especially when a large sample size is available. The general form of the empirical copula C_n can be expressed as (Nelsen 2006):

$$C_n\left(\frac{i}{n}, \frac{j}{n}\right) = \frac{\#\left(x \leq x_{(i)}, y \leq y_{(j)}\right)}{n} \quad (54)$$

where n is the sample size and $x_{(i)}$ and $x_{(j)}$ are order statistics with $1 \leq i, j \leq n$.

Generally, the commonly used copula falls short in modeling the dependence in high dimensions. The vine copula (Bedford and Cooke 2002) has attracted much attention recently, since it enables the flexible dependence modeling in high dimensions (Joe 2014; Kurowicka and Joe 2011). The main concept of the vine copula of d -dimension in dependence modeling is to decompose it into $d(d-1)/2$ bivariate copulas according to a tree structure of d nodes. For example, for the trivariate case, the joint density function $f(x_1, x_2, x_3)$ can be expressed with the vine copulas as (Aas et al. 2009):

$$f(x_1, x_2, x_3) = f_1(x_1) \times f_2(x_2) \times f_3(x_3) \times c_{12}[F_1(x_1), F_2(x_2)] \\ \times c_{23}[F_2(x_2), F_3(x_3)] \times c_{13|2}[F(x_1|x_2), F(x_3|x_2)] \quad (55)$$

where $c_{12}[F_1(x_1), F_2(x_2)]$, $c_{23}[F_2(x_2), F_3(x_3)]$ and $c_{13|2}[F(x_1|x_2), F(x_3|x_2)]$ are the joint copula density functions.

Due to the advantage of flexible dependence modeling (including the tail dependence) in high dimension, the vine copula has emerged to be a powerful tool to be used in a variety of applications in hydrometeorology, including the statistical forecasting.

7 Hypothesis Testing

A hypothesis is a statement about the population parameter. The statistical hypothesis test (also known as significance testing) is used to decide which of two complementary hypotheses (*null hypothesis* H_0 and *alternative hypothesis* H_1) is true, based on the evidence from data. Let θ be a population parameter. The general form of the hypothesis testing is to test the *null hypothesis*:

$$H_0 : \theta \in \Theta \quad (56)$$

against the *alternative hypothesis*:

$$H_1 : \theta \notin \Theta \quad (57)$$

where Θ is a subset of the parameter space.

Generally there are two types of hypothesis tests: the parametric hypothesis test and nonparametric hypothesis test, based on whether the assumption about a particular distribution form of the data has been made (Wilks 2011). A statistic is a numerical measure of the characteristic of the sample. In hypothesis testing, the statistic is generally regarded as a random variable, which can be characterized by a probability distribution based on the data (the sampling distribution). A statistic that is of particular interest to the data and question at hand is generally called a test statistic.

A hypothesis testing generally proceeds with the following procedure (Helsel and Hirsch 1992). After choosing a test statistic (e.g., the mean of a distribution), a null hypothesis (and the corresponding alternative hypothesis) is then defined (e.g., the mean equals zero). Based on a chosen significance level or acceptable error rate α (e.g., 5%), the sampling distribution of the test statistic is then derived (a null distribution) either based on a parametric or nonparametric distribution, from which one can compute the p -value (the probability that the observed test statistic will occur from the sampling distribution). The null hypothesis is rejected if the p -value is lower than or equal to the significance level (e.g., 0.05).

The chi-square distribution with n degrees of freedom is the distribution of the sum of n squared independent standard normal variates. Suppose X_1, \dots, X_n are independent and standard normal random variables. Denote Z the sum of the squares of X , i.e.:

$$Z = \sum_{i=1}^n x_i^2 \quad (58)$$

Then the distribution of Z is the chi-square distribution with n degree of freedom and is extensively used in hypothesis test mainly due to its relationship to the normal distribution. In the hypothesis test, a sum of squared errors (or the difference between observed and expected frequencies) is generally constructed, and the sample distribution of the test statistic follows the chi-square distribution. The chi-square test is a commonly used statistical hypothesis test to determine whether a significant association or relationship exists between two random variables and the goodness of fit of observations to hypothetical distributions.

8 Bayesian Inference

The methods and tools introduced before for the statistical inference are mainly the classical approach (or the frequency approach) that interprets the probability as relative frequencies, which only use the information in the data. The Bayesian approach is another approach for the statistical inference that combines information from both the samples and other information previous to the collection of samples (Montgomery and Runger 2010).

Bayesian and classical (or frequentist) approaches are two branches in the statistical inferences with different interpretation of the meaning of probability. For the frequentist approach, the probability is defined as the frequency

of events to occur from repeated experiments or trials. For the Bayesian approach, a probability is defined as a measure of the degree of belief in an event based on the information available. The parameter is also interpreted differently by two approaches. For the frequentist approach, parameters are fixed but unknown, while from a Bayesian viewpoint, parameters are random variables (Casella and Berger 2002). The difference in the interpretation of the probability definition and parameters results in differences in the framework for statistical inference. For a 95% confidence interval, the interpretation from the frequentist is that 95% of intervals will cover the true parameter based on repeated experiments, while the Bayesian concludes that the parameter falls within the interval with 95% probability based on data analysis. The two approaches highlight different aspects of statistical inferences and understanding of the underlying theories and interpretations would be essential for proper practices in hydrometeorological studies.

In this section, we briefly introduce the Bayesian approach for statistical inference of parameters. Assume that a random variable X_1, X_2, \dots, X_n has the probability density function f with parameter θ . The probability density function is written as $f(\mathbf{x}|\theta)$, indicating that the distribution is conditional on the values of θ . Note that in the Bayesian approach, parameter θ is regarded as a random variable described by a probability distribution. This distribution is based on the subjective belief and is formulated before the samples are collected. As such, the distribution of the parameter is generally called the prior distribution, denoted as $\pi(\theta)$, which can be updated with the sample information. Specifically, the posterior (or conditional) distribution of θ given the sample \mathbf{x} , denoted as $f(\theta|\mathbf{x})$, can be expressed as (Casella and Berger 2002; Montgomery and Runger 2010):

$$f(\theta|\mathbf{x}) = \frac{f(\mathbf{x}, \theta)}{p(\mathbf{x})} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{p(\mathbf{x})} \quad (59)$$

where $f(\mathbf{x}, \theta)$ is the joint distribution of the sample and the parameter, $f(\mathbf{x}|\theta)$ is the conditional distribution of the sample, and $p(\mathbf{x})$ is the marginal distribution of \mathbf{x} that can be expressed as:

$$p(\mathbf{x}) = \int f(\mathbf{x}|\theta)\pi(\theta)d\theta \quad (60)$$

The posterior distribution can then be used for the statistical inference of parameter θ . Generally the Bayesian estimator of parameter θ corresponds to the mean of the posterior distribution $f(\theta|\mathbf{x})$.

9 Hydrometeorological Forecasting with Statistical Methods

Statistical forecasting method has been applied to predict hydroclimatic variables of interested based on relationships of the historical observations. In addition, these methods have also been commonly used to post-process the short-term weather

forecast or long-term climate forecast (e.g., model output statistics, MOS (Wilks 2011)). The probabilistic forecast is generally desirable to reflect the inherent uncertainty, and the statistical forecasting plays an important role in this regard through the inclusion of probability. A brief introduction of the statistical methods for hydrometeorological forecasting is introduced in this section.

9.1 Linear Methods

Linear Regression

Regression method is among the most commonly used methods for forecasting. The concept of the regression for the statistical forecasting is to establish the relationship between two variables. Given the observations of x and y , the simple linear regression for forecasting can be expressed as:

$$y = \alpha + \beta x + \varepsilon \quad (61)$$

where α and β are the parameters; Y is the variable to be predicted (response variable or predictand); X is the independent variable, explanatory or predictor; and ε is the error term assumed to be Gaussian distributed. In the modeling setting, X can be multiple predictors (termed as multiple linear regressions), and the variables Y can be a vector (termed as multivariate linear regression).

Based on observation of the dependent variables Y and independent variables X , the parameter in the equation can be estimated (often referred to as “model estimation”). Usually, the regression model chooses the line with the least error for the predictions of y given values of x . Since the most commonly used criteria of the error is the sum of the squared errors, the regression is commonly termed least square regression. After the parameter estimation, the forecast of y can be achieved based on the new values of x .

Autoregressive (AR) Model

The autoregressive (AR) is commonly used for the forecasting of hydrometeorological time series, which is a special case of the autoregressive integrated moving average (ARIMA) model. The main concept of the AR model for forecasting is that the predictand depends linearly on its own previous values (and on a stochastic component). The AR model of order p for the time series z_t can be expressed as:

$$z_t = \alpha + \sum_{i=1}^p \beta_i z_{t-i} + \varepsilon_t \quad (62)$$

where α is the constant, β is the parameters, and ε_t is the white noise. This model is generally referred to as the AR(p) model and can be extended for the forecasting of multivariate time series (or the vector autoregressive models, VAR).

9.2 Nonlinear Models

In reality, the relationship between different variables among the climatic system and hydrologic cycle is complicated and may not be linear. In these cases, the linear models may not be suitable since the relationship between the predictand and predictor is generally assumed to be linear. Nonlinear models, such as logistic regression, artificial neural networks (ANN), support vector machines (SVM), genetic programming (GP), fuzzy logic (FL), and wavelet transforms (WT), have been developed to overcome some of the limitations of linear models for forecasting purposes in hydrology and hydrometeorology (Deka 2014; Fahimi et al. 2016; Yaseen et al. 2015). In this section, we mainly focus on the logistic regression, ANN and SVM methods.

Logistic Regression

When the relation between the predictand and the prediction is not linear or the assumption of normal residuals is not valid, the nonlinear regression can be used to model the relationship between two variables. For example, when the predictand is binary with values 0 and 1 only (e.g., the occurrence of drought or the occurrence of streamflow $>1000 \text{ m}^3/\text{s}$), the residuals are not Gaussian. In this case, the logistic regression is a useful tool for modeling and forecasting purposes with respect to multiple predictors. It has been widely used in the hydroclimatic forecasting, such as precipitation, streamflow, or drought (Hamill et al. 2008; Hamill and Whitaker 2006; Hao et al. 2016; Regonda et al. 2006; Wilks and Hamill 2007).

Let p be the probability of the occurrence of a specified event (i.e., $p = P(Y = 1)$). The general form of the logistic regression of a binary predictand Y can be expressed as:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta x \quad (63)$$

Equation 63 leads to:

$$p = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \quad (64)$$

It should be noted that the logistic regression results in the probability of the binary predictand (between 0 and 1) and thus facilitates the probabilistic regression.

Both the linear regression and the logistic regression are special cases of the generalized linear models (GLM), which is a synthesis and extension of regression models. The GLM consists of three components (Dunteman and Ho 2006; McCullagh and Nelder 1989): (1) A random component that specifies the conditional distribution of the response variable Y , which the distribution is generally a member of the exponential family, such as the Gaussian, binomial, and Poisson distribution; (2) a systematic component (or linear predictor) that is a linear function of

explanatory variables, which can be expressed as $\alpha + \beta x$; and (3) a link function (e.g., $g(\mu)$) that specifies the link between random component and systematic components. Specifically, it links the expected value of the response variable $\mu = E(Y)$ and the linear predictor of explanatory variables X , i.e.:

$$g(\mu) = g(E(Y)) = \alpha + \beta x \quad (65)$$

Some commonly used link function includes the identity, logit, or log. For example, the logit link function can be expressed as:

$$\text{logit}(p) = \frac{p}{1 - p} \quad (66)$$

It can be seen that the linear regression is the GLM with identity link function (and the Gaussian distribution of the response variable), while the logistic regression is the GLM with logit link function (and Bernoulli distribution of the response variable). Note that the GLM does not assume a linear relationship between the response variable Y and the explanatory variables. However, the assumption of a linear relationship between the transformed response variables and the explanatory variables is made in terms of the link function.

Artificial Neural Networks

Due to the complexity of real hydroclimatic systems, it is sometimes difficult to specify the specific relationships between inputs (e.g., predictors) and outputs (e.g., predictand). In this case, some traditional statistical forecasting methods may not perform well due to the assumption about the particular forms of the relationship (e.g., linear relationship between predictors and (transformed) predictand). The artificial neural networks (ANN) method has been applied for forecasting and is a nonlinear model for modeling the complex relationships among hydroclimatic processes (ASCE 2000; Dawson and Wilby 2001; Govindaraju and Rao 2013). The ANN is a soft computing technique and is data driven with only a few assumptions about models for problems at hand. It is generally based on learning and pattern recognition from observations and establishes the complicated relationship between inputs and outputs that are generally hard to describe.

After training with observations of inputs and outputs of the model, the ANN can then be applied for forecasting purposes to infer the future behavior of outputs given new inputs. The ANN has been widely used for hydrometeorological forecasting, such as streamflow prediction based on a variety of inputs, including rainfall, temperature, or snow at different time scales (Govindaraju and Rao 2013; Hsu et al. 1995; Tokar and Johnson 1999).

Support Vector Machine

The support vector machine (SVM) (or its extension support vector regression (SVR)) is one of the soft computing techniques and has been widely used in hydrological and environmental forecasting problems (Hsieh 2009). The basic idea

of a SVM method for the regression or forecasting problem is to map input data into a high-dimensional feature space by a nonlinear mapping, in which the linear regression is then performed (Shabri and Suhartono 2012; Smola and Schölkopf 2004). The SVM is effective in high-dimensional spaces. To some extent, it is similar to the ANN model in that the input is mapped nonlinearly to a hidden space and then to the outputs in which an error function is minimized for training the model (Bourdin et al. 2012). The advantage of SVM is that it can be applied with a small training dataset and the global minimum is easy to obtain, since the training in SVMs is equivalent to solving a linearly constrained quadratic programming problem (Behzad et al. 2009; Cao 2003). Recently, it has been used for the hydrological forecasting of streamflow or drought (Asefa et al. 2006; Deka 2014; Ganguli and Reddy 2014; Kalra and Ahmad 2009; Shabri and Suhartono 2012).

9.3 Ensemble Forecast

Ensemble forecast is the recent advance in hydrometeorological forecasting that generates an ensemble of forecasts for the target period. Instead of relying on a single deterministic prediction, the ensemble forecast utilizes the forecast from multi-models and provides probabilistic forecasts for operational water resources planning and management. In past decades, the ensemble forecast has been commonly used in the short-term weather forecast (Krishnamurti et al. 2000; Wilks and Hamill 2007), seasonal climate forecast (Kharin and Zwiers 2002; Palmer et al. 2004), and hydrologic forecast through different land surface models (Duan et al. 2007; Luo and Wood 2008). For example, the ensemble forecast of streamflow can be generated through a hydrological model driven by numerical weather prediction (NWP) products from the ensemble prediction system (Cloke et al. 2013; Cuo et al. 2011; Schaake et al. 2007). Meanwhile, the ensemble forecast can also be achieved through the ensemble streamflow prediction (ESP) method (Day 1985; Twedd et al. 1977). The ESP method uses the resampled historical meteorological data that are assumed to occur in the future with equal probability to drive the hydrological model to produce ensembles of streamflow, which has been operational used by the National Weather Service River Forecast System (NWSRFS) (Demargne et al. 2014; Werner et al. 2005).

The statistical methods have been commonly used in the ensemble forecast to post-process the dynamic forecast from different models, which is also referred to as the hybrid forecast. With a suite of ensemble members from dynamic models, a key question is how to obtain a superior forecast and quantify the uncertainty of forecast. Intuitively, the forecast from multiple models can be combined linearly with different weighting methods, either based on the equal weight or optimal weights determined by regression models (Duan et al. 2007). Extensive studies have shown that the multi-model ensemble generally performs better than the forecast from the individual member (Georgakakos et al. 2004; Krishnamurti et al. 2000; Shamseldin et al. 1997). A variety of methods have been developed for multi-model averaging with forecasts from a variety of models (Arsenault et al. 2015), such as simple arithmetic mean and Bayesian model averaging.

Due to the complexity of the climate systems and hydrologic systems and uncertainties in the inputs, parameterizations, and model configurations, it is important to quantify the uncertainty from various sources in the weather, climate, and hydrologic forecasting systems. Recently, extensive efforts have been devoted to the ensemble forecasting based on multi-model weather, climate, and hydrological forecast to aid informed decision-makers (Demargne et al. 2014; Krishnamurti et al. 2016; Yuan et al. 2015b). Certain challenges still exist in the ensemble forecast, such as how to combine short-term and long-term forecasts in a seamless manner and how to provide the forecast information that is easily understandable and useful to end users (Demargne et al. 2014). Improving the hydrometeorological forecast skill through multi-model ensemble forecast is an endless and sustained endeavor, and efforts are still needed to address technical and institutional challenges to aid water resources management and to cope with natural hazards.

10 Summary

For statistical inference, hydroclimatic variables are generally regarded as random variables to facilitate a variety of applications. For risk analysis of floods, extreme rainfall, and droughts, a probability distribution is the basis, and a variety of distribution forms are available to this end, such as parametric distribution, nonparametric distribution, and mixed distribution. The selection and application of the distribution function should be suitable for the problem at hand. The parametric distribution is easy to implement but may be subject to the limitations that a specified form of distribution has to be specified. The nonparametric method allows for the probability estimation without the assumption of the distribution form but may fall short in the statistical interpolation due to the boundary problem. After a suitable selection and estimation of the distribution function, a variety of applications can be achieved, such as random simulations from the estimated distribution for hydroclimatic simulations.

It has been recognized that different variables are mutually correlated, such as precipitation and streamflow. Moreover, properties of different hydroclimatic events are also dependent on each other, such as rainfall duration and severity and flood volume and peak, to name a few. Based on the concept of statistical dependence, the joint distribution is a powerful tool to describe the joint behavior of multiple variables or properties. Recent decades have witnessed fast development of the methods for dependence modeling of multiple variables, and there is a surge in the development and application of copula in various applications of hydrometeorology. Several copulas, such as empirical copula, parametric copula, and vine copula, are introduced with distinct properties in modeling the dependence. The vine copula, due to its advantage in flexible dependence modeling even in high dimensions, has attracted much attention recently and is expected to be widely applied in hydrometeorological applications including forecasting.

Most of the concepts and methods in this chapter are introduced from the viewpoint of classical statistics. An alternative and attractive approach is the Bayesian inference, which interpolates the meaning of probability and parameter

differently from classical statistics. Instead of viewing the parameter as fixed, the Bayesian approach regards the distribution parameter as a random variable, and the estimation of the distribution of parameter is based on both the prior belief and information from the sample. Different statistical methods for hydrometeorological forecasting are reviewed, which can be broadly classified as the linear model, such as linear regression and autoregressive (AR) model and nonlinear model, such as the logistic regression, artificial neural networks (ANN), and support vector machines (SVM). The ensemble forecast method based on multiple forecast members has attracted much attention in hydrometeorological forecasting in the past decade, which provides the probabilistic forecasting with uncertainty estimation and is important for decision-making.

References

- K. Aas, C. Czado, A. Frigessi, et al., Pair-copula constructions of multiple dependence. *Insur. Math. Econ.* **44**(2), 182–198 (2009)
- R. Arsenault, P. Gatien, B. Renaud, et al., A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. *J. Hydrol.* **529**, 754–767 (2015)
- ASCE, Artificial neural networks in hydrology. II: hydrologic applications. *J. Hydrol. Eng.* **5**(2), 124–137 (2000)
- T. Asefa, M. Kembrowski, M. McKee, et al., Multi-time scale stream flow predictions: the support vector machines approach. *J. Hydrol.* **318**(1), 7–16 (2006)
- B. Bacchi, G. Becciu, N.T. Kottekoda, Bivariate exponential model applied to intensities and durations of extreme rainfall. *J. Hydrol.* **155**(1), 225–236 (1994)
- N. Balakrishnan, C. Lai, *Continuous Bivariate Distributions* (Springer, New York, 2009)
- T. Bedford, R.M. Cooke, Vines – a new graphical model for dependent random variables. *Ann. Stat.* **30**(4), 1031–1068 (2002)
- M. Behzad, K. Asghari, M. Eazi, et al., Generalization performance of support vector machines and neural networks in runoff modeling. *Expert Syst. Appl.* **36**(4), 7624–7629 (2009)
- K.J. Beven, *Rainfall-Runoff Modelling: The Primer* (Wiley, New York, 2011)
- K. Bogner, F. Pappenberger, H. Cloke, et al., Technical note: the normal quantile transformation and its application in a flood forecasting system. *Hydrol. Earth Syst. Sci.* **16**(4), 1085–1094 (2012)
- D.R. Bourdin, S.W. Fleming, R.B. Stull, Streamflow modelling: a primer on applications, approaches and challenges. *Atmosphere-Ocean* **50**(4), 507–536 (2012)
- L. Cao, Support vector machines experts for time series forecasting. *Neurocomputing* **51**, 321–339 (2003)
- G. Casella, R.L. Berger, *Statistical Inference* (Duxbury, Pacific Grove, 2002)
- V. Choulakian, N. El-Jabi, J. Moussi, On the distribution of flood volume in partial duration series analysis of flood phenomena. *Stoch. Hydrol. Hydraul.* **4**(3), 217–226 (1990)
- H. Cloke, F. Pappenberger, Ensemble flood forecasting: a review. *J. Hydrol.* **375**(3), 613–626 (2009)
- H.L. Cloke, F. Pappenberger, S.J. van Andel, et al., Hydrological ensemble prediction systems preface. *Hydrol. Process.* **27**, 1–4 (2013)
- L. Cuo, T.C. Pagano, Q. Wang, A review of quantitative precipitation forecasts and their use in short-to medium-range streamflow forecasting. *J. Hydrometeorol.* **12**(5), 713–728 (2011)
- C. Dawson, R. Wilby, Hydrological modelling using artificial neural networks. *Prog. Phys. Geogr.* **25**(1), 80–108 (2001)
- G.N. Day, Extended streamflow forecasting using NWSRFS. *J. Water Resour. Plan. Manag.* **111**(2), 157–170 (1985)
- P.C. Deka, Support vector machine applications in the field of hydrology: a review. *Appl. Soft Comput.* **19**, 372–386 (2014)

- J. Demargne, L. Wu, S.K. Regonda, et al., The science of NOAA's operational hydrologic ensemble forecast service. *Bull. Am. Meteorol. Soc.* **95**(1), 79–98 (2014)
- Q. Duan, N.K. Ajami, X. Gao, et al., Multi-model ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.* **30**(5), 1371–1386 (2007)
- G.H. Duntzman, M.-H.R. Ho, *An Introduction to Generalized Linear Models* (CRC Press, Boca Raton, 2006)
- F. Fahimi, Z.M. Yaseen, A. El-shafie, Application of soft computing based hybrid models in hydrological variables modeling: a comprehensive review. *Theor. Appl. Climatol.* **128**(3–4), 875–903 (2016)
- A.C. Favre, S. El Adlouni, L. Perreault, et al., Multivariate hydrological frequency analysis using copulas. *Water Resour. Res.* **40**(1), W01101 (2004). <https://doi.org/10.1029/02003WR002456>
- M. Fuglem, G. Parr, I. Jordaan, Plotting positions for fitting distributions and extreme value analysis. *Can. J. Civ. Eng.* **40**(2), 130–139 (2013)
- P. Ganguli, M.J. Reddy, Ensemble prediction of regional droughts using climate inputs and the SVM–copula approach. *Hydrol. Process.* **28**(19), 4989–5009 (2014)
- C. Genest, A.-C. Favre, Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydrol. Eng.* **12**(4), 347–368 (2007)
- C. Genest, A.C. Favre, J. Béliveau, et al., Metaelliptical copulas and their use in frequency analysis of multivariate hydrological data. *Water Resour. Res.* **43**, W09401 (2007). <https://doi.org/10.1029/02006WR005275>
- K.P. Georgakakos, D.-J. Seo, H. Gupta, et al., Towards the characterization of streamflow simulation uncertainty through multimodel ensembles. *J. Hydrol.* **298**(1), 222–241 (2004)
- T. Ghizzoni, G. Roth, R. Rudari, Multivariate skew-t approach to the design of accumulation risk scenarios for the flooding hazard. *Adv. Water Resour.* **33**(10), 1243–1255 (2010)
- N. Goel, S. Seth, S. Chandra, Multivariate modeling of flood flows. *J. Hydraul. Eng.* **124**(2), 146–155 (1998)
- R.S. Govindaraju, A.R. Rao, *Artificial Neural Networks in Hydrology* (Springer Science & Business Media, Heidelberg, 2013)
- I.I. Gringorten, A plotting rule for extreme probability paper. *J. Geophys. Res.* **68**(3), 813–814 (1963)
- T.M. Hamill, J.S. Whitaker, Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Weather Rev.* **134**(11), 3209–3229 (2006)
- T.M. Hamill, R. Hagedorn, J.S. Whitaker, Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part II: precipitation. *Mon. Weather Rev.* **136**(7), 2620–2632 (2008)
- Z. Hao, V.P. Singh, Single-site monthly streamflow simulation using entropy theory. *Water Resour. Res.* **47**(9), W09528 (2011). <https://doi.org/10.1029/02010WR010208>
- Z. Hao, V.P. Singh, Entropy-based method for bivariate drought analysis. *J. Hydrol. Eng.* **18**(7), 780–786 (2013a)
- Z. Hao, V.P. Singh, Entropy-based method for extreme rainfall analysis in Texas. *J. Geophys. Res. Atmos.* **118**(2), 263–273 (2013b)
- Z. Hao, V.P. Singh, Modeling multi-site streamflow dependence with maximum entropy copula. *Water Resour. Res.* **49** (2013c). <https://doi.org/10.1002/wrcr.20523>
- Z. Hao, V.P. Singh, Integrating entropy and copula theories for hydrologic modeling and analysis. *Entropy* **17**(4), 2253–2280 (2015)
- Z. Hao, V.P. Singh, Review of dependence modeling in hydrology and water resources. *Prog. Phys. Geogr.* **40**, 549 (2016)
- Z. Hao, F. Hao, Y. Xia, et al., A statistical method for categorical drought prediction based on NLDAS-2. *J. Appl. Meteorol. Climatol.* **55**, 1049 (2016)
- D.R. Helsel, R.M. Hirsch, *Statistical Methods in Water Resources* (Elsevier, Amsterdam, 1992)
- J.A. Hoeting, D. Madigan, A.E. Raftery, et al., Bayesian model averaging: a tutorial. *Stat. Sci.* **14**(4), 382–401 (1999)
- W.W. Hsieh, *Machine Learning Methods in the Environmental Sciences: Neural Networks and Kernels* (Cambridge University Press, Cambridge, UK, 2009)
- K.l. Hsu, H.V. Gupta, S. Sorooshian, Artificial neural network modeling of the rainfall-runoff process. *Water Resour. Res.* **31**(10), 2517–2530 (1995)

- H. Joe, *Multivariate Models and Dependence Concepts* (Chapman & Hall, London, 1997)
- H. Joe, *Dependence Modeling with Copulas* (CRC Press, Boca Raton, 2014)
- M.C. Jones, J.S. Marron, S.J. Sheather, A brief survey of bandwidth selection for density estimation. *J. Am. Stat. Assoc.* **91**(433), 401–407 (1996)
- A. Kalra, S. Ahmad, Using oceanic-atmospheric oscillations for long lead time streamflow forecasting. *Water Resour. Res.* **45**(3), W03413 (2009)
- K. Kelly, R. Krzysztofowicz, A bivariate meta-Gaussian density for use in hydrology. *Stoch. Hydraul. Hydraul.* **11**(1), 17–31 (1997)
- V.V. Kharin, F.W. Zwiers, Climate predictions with multimodel ensembles. *J. Clim.* **15**(7), 793–799 (2002)
- B. Kirtman, D. Min, J. Infanti, et al., The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Am. Meteorol. Soc.* **95**(4), 585–601 (2014)
- S. Kotz, N. Balakrishnan, N.L. Johnson, *Continuous Multivariate Distributions: Models and Applications* (Wiley, New York, 2000)
- T.N. Krishnamurti, C. Kishtawal, Z. Zhang, et al., Multimodel ensemble forecasts for weather and seasonal climate. *J. Clim.* **13**(23), 4196–4216 (2000)
- T. Krishnamurti, V. Kumar, A. Simon, et al., A review of multimodel superensemble forecasting for weather, seasonal climate, and hurricanes. *Rev. Geophys.* **54**, 336 (2016)
- D. Kurowicka, H. Joe, *Dependence Modeling: Vine Copula Handbook* (World Scientific, Singapore, 2011)
- U. Lall, Recent advances in nonparametric function estimation: hydrologic applications. *Rev. Geophys.* **33**(S2), 1093–1102 (1995)
- U. Lall, A. Sharma, A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.* **32**(3), 679–693 (1996). <https://doi.org/10.1029/1095WR02966>
- P. Laux, S. Vogl, W. Qiu, et al., Copula-based statistical refinement of precipitation in RCM simulations over complex terrain. *Hydrol. Earth Syst. Sci.* **15**(7), 2401–2419 (2011)
- C. Li, V.P. Singh, A.K. Mishra, A bivariate mixed distribution with a heavy-tailed component and its application to single-site daily rainfall simulation. *Water Resour. Res.* **49**(2), 767–789 (2013)
- D.P. Loucks, E. Van Beek, J.R. Stedinger, et al., *Water Resources Systems Planning and Management: An Introduction to Methods, Models and Applications* (UNESCO, Paris, 2005)
- L. Luo, E.F. Wood, Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. *J. Hydrometeorol.* **9**(5), 866–884 (2008)
- L. Luo, E.F. Wood, M. Pan, Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.* **112**(D10), D10102 (2007)
- L. Makkonen, Plotting positions in extreme value analysis. *J. Appl. Meteorol. Climatol.* **45**(2), 334–340 (2006)
- P. McCullagh, J.A. Nelder, *Generalized Linear Models* (CRC Press, Boca Raton, 1989)
- J. McEnergy, J. Ingram, Q. Duan, et al., NOAA's advanced hydrologic prediction service. *Bull. Am. Meteorol. Soc.* **86**(3), 375 (2005)
- A. Montanari, A. Brath, A stochastic approach for assessing the uncertainty of rainfall-runoff simulations. *Water Resour. Res.* **40**(1), W01106 (2004)
- D.C. Montgomery, G.C. Runger, *Applied Statistics and Probability for Engineers* (Wiley, New York, 2010)
- S. Nadarajah, A bivariate gamma model for drought. *Water Resour. Res.* **43**(8), W08501 (2007). <https://doi.org/10.1029/02006WR005641>
- R.B. Nelsen, *An Introduction to Copulas* (Springer, New York, 2006)
- T. Palmer, F. Doblas-Reyes, R. Hagedorn, et al., Development of a European multimodel ensemble system for seasonal-to-interannual prediction (DEMIETER). *Bull. Am. Meteorol. Soc.* **85**(6), 853–872 (2004)
- C. Piani, J. Haerter, Two dimensional bias correction of temperature and precipitation copulas in climate models. *Geophys. Res. Lett.* **39**(20), L20401 (2012)

- W. Pozzi, J. Sheffield, R. Stefanski, et al., Towards global drought early warning capability: expanding international cooperation for the development of a framework for global drought monitoring and forecasting. *Bull. Am. Meteorol. Soc.* **94**(6), 776–785 (2013)
- J. Prairie, B. Rajagopalan, T. Fulp, et al., Modified K-NN model for stochastic streamflow simulation. *J. Hydrol. Eng.* **11**(4), 371–378 (2006)
- A.E. Raftery, T. Gneiting, F. Balabdaoui, et al., Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Weather Rev.* **133**(5), 1155–1174 (2005)
- S.K. Regonda, B. Rajagopalan, M. Clark, A new method to produce categorical streamflow forecasts. *Water Resour. Res.* **42**(9), W09501 (2006)
- B. Sackl, H. Bergmann, A bivariate flood model and its application, in *Hydrologic Frequency Modeling. Proceedings of the International Symposium on Flood Frequency and Risk Analyses, Baton Rouge USA, 1986*, ed. by V.P. Sing (Springer, Dordrecht, The Netherlands, 1987), pp. 571–582
- G. Salvadori, C. De Michele, Multivariate multiparameter extreme value models and return periods: a copula approach. *Water Resour. Res.* **46**(10), W10501 (2010)
- J.C. Schaake, T.M. Hamill, R. Buizza, et al., HEPEX: the hydrological ensemble prediction experiment. *Bull. Am. Meteorol. Soc.* **88**(10), 1541 (2007)
- A. Shabri, Suhartono, Streamflow forecasting using least-squares support vector machines. *Hydrol. Sci. J.* **57**(7), 1275–1293 (2012)
- A.Y. Shamseldin, K.M. O'Connor, G. Liang, Methods for combining the outputs of different rainfall-runoff models. *J. Hydrol.* **197**(1–4), 203–229 (1997)
- A. Sharma, D. Tarboton, U. Lall, Streamflow simulation: a nonparametric approach. *Water Resour. Res.* **33**(2), 291–308 (1997)
- W. Shaw, K. Lee, Bivariate Student t distributions with variable marginal degrees of freedom and independence. *J. Multivar. Anal.* **99**(6), 1276–1287 (2008)
- B. Silverman, *Density Estimation for Statistics and Data Analysis* (Chapman & Hall/CRC, New York, 1986)
- K. Singh, V. Singh, Derivation of bivariate probability density functions with exponential marginals. *Stoch. Hydraul.* **5**(1), 55–68 (1991)
- V.P. Singh, D.A. Woolhiser, Mathematical modeling of watershed hydrology. *J. Hydrol. Eng.* **7**(4), 270–292 (2002)
- B. Sivakumar, R. Berndtsson, *Advances in Data-Based Approaches for Hydrologic Modeling and Forecasting* (World Scientific, Hackensack, 2010)
- A.J. Smola, B. Schölkopf, A tutorial on support vector regression. *Stat. Comput.* **14**(3), 199–222 (2004)
- J. Thielen, J. Schaake, R. Hartman, et al., Aims, challenges and progress of the Hydrological Ensemble Prediction Experiment (HEPEX) following the third HEPEX workshop held in Stresa 27 to 29 June 2007. *Atmos. Sci. Lett.* **9**(2), 29–35 (2008)
- A.S. Tokar, P.A. Johnson, Rainfall-runoff modeling using artificial neural networks. *J. Hydrol. Eng.* **4**(3), 232–239 (1999)
- P.K. Trivedi, D.M. Zimmer, Copula modeling: an introduction for practitioners. *Found. Trends Econom.* **1**(1), 1–111 (2005)
- T.M. Twedt, J.C. Schaake Jr., E.L. Peck, National Weather Service extended streamflow prediction, in *Proceedings of the 45th Annual Western Snow Conference*, Albuquerque, 1977
- M. Van den Berg, S. Vandenberghe, B.D. Baets, et al., Copula-based downscaling of spatial rainfall: a proof of concept. *Hydrol. Earth Syst. Sci.* **15**(5), 1445–1457 (2011)
- N.E. Verhoest, M.J. van den Berg, B. Martens, et al., Copula-based downscaling of coarse-scale soil moisture observations with implicit bias correction. *IEEE Trans. Geosci. Remote Sens.* **53**(6), 3507–3521 (2015)
- S. Vogl, P. Laux, W. Qiu, et al., Copula-based assimilation of radar and gauge information to derive bias-corrected precipitation fields. *Hydrol. Earth Syst. Sci.* **16**(7), 2311–2328 (2012)
- M.P. Wand, M.C. Jones, *Kernel Smoothing*. (CRC Press, Boca Raton, 1994)
- K. Werner, D. Brandon, M. Clark, et al., Incorporating medium-range numerical weather model output into the ensemble streamflow prediction system of the National Weather Service. *J. Hydrometeorol.* **6**(2), 101–114 (2005)

- D.S. Wilks, *Statistical Methods in the Atmospheric Sciences* (Academic, San Diego, 2011)
- D.S. Wilks, T.M. Hamill, Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Weather Rev.* **135**(6), 2379–2390 (2007)
- E.F. Wood, S.D. Schubert, A.W. Wood, et al., Prospects for advancing drought understanding, monitoring and prediction. *J. Hydrometeorol.* **16**(4), 1636–1657 (2015)
- Z.M. Yaseen, A. El-shafie, O. Jaafar, et al., Artificial intelligence based models for stream-flow forecasting: 2000–2015. *J. Hydrol.* **530**, 829–844 (2015)
- X. Yuan, E.F. Wood, On the clustering of climate models in ensemble seasonal forecasting. *Geophys. Res. Lett.* **39**(18), L18701 (2012)
- X. Yuan, E.F. Wood, Z. Ma, A review on climate-model-based seasonal hydrologic forecasting: physical understanding and system development. *WIREs Water* **2**, 523–536 (2015a)
- X. Yuan, J.K. Roundy, E.F. Wood, et al., Seasonal forecasting of global hydrologic extremes: system development and evaluation over GEWEX basins. *Bull. Am. Meteorol. Soc.* **96**, 1895 (2015b)
- S. Yue, Applying bivariate normal distribution to flood frequency analysis. *Water Int.* **24**(3), 248–254 (1999)
- S. Yue, The bivariate lognormal distribution to model a multivariate flood episode. *Hydrol. Process.* **14**(14), 2575–2588 (2000)
- S. Yue, T. Ouarda, B. Bobée, A review of bivariate gamma distributions for hydrological application. *J. Hydrol.* **246**(1), 1–18 (2001)



Estimation of Probability Distributions for Hydrometeorological Applications

Grey S. Nearing

Contents

1	Introduction	1464
1.1	Uncertainty and Probability in Hydrometeorological Forecasting	1464
1.2	The General Forecasting Problem	1465
1.3	The Role of the Ensemble	1466
2	Estimating Probability Distributions	1466
2.1	Density Estimation from iid Samples	1466
2.2	Estimating Distributions over Models and Their Parameters	1477
3	Summary	1483
	References	1484

Abstract

Hydrometeorologists use imperfect (i.e., incomplete and/or partially erroneous) measurements and imperfect models to make predictions, both for forecasting and to support scientific inference. Because no models or data are ever perfect, forecasting and hypothesis testing must account for uncertainty. The probability calculus is unarguably the most common quantitative framework used for this purpose. This article presents probabilistic methods for estimating and reducing uncertainty that are common in hydrometeorological applications. The major focus is on Bayesian methods and approximations of those methods based on ensembles (i.e., Monte Carlo methods). The article includes a brief overview of both parametric and nonparametric methods, a brief introduction to inverse methods, and a brief introduction to data assimilation from a Bayesian perspective. It is important to caution that although the probability calculus can be used to

G. S. Nearing (✉)

Department of Geological Sciences, University of Alabama, Tuscaloosa, AL, USA

e-mail: gsnearing@ua.edu

estimate predictive uncertainty and to aid scientific reasoning, all applications of probabilistic reasoning necessarily contain some amount of subjectivity.

Keywords

Probability · Bayes · Markov Chain Monte Carlo · Data assimilation · Nonparametric methods

1 Introduction

1.1 Uncertainty and Probability in Hydrometeorological Forecasting

Models of the hydrosphere translate inputs (e.g., forcing data, parameters, initial states) into predictions. Models of biogeophysical processes are generally deterministic in the sense that given a particular set of inputs, they provide a deterministic value for the predicted variable. Such deterministic predictions might be loosely interpreted as Dirac delta distributions in the sense that they place infinite probability on a single outcome (unit probability when there are a finite number of possible outcomes) and zero probability on all other outcomes. Since no models or data are ever perfect, predictions are always subject to some amount of uncertainty, and our presumed objective is to estimate this uncertainty as accurately as possible given information available from models and data.

The most common tool for dealing with uncertainty is the probability calculus. In this article, I take the perspective that the probability calculus measures relative beliefs of a rational agent (Jaynes 2003). This interpretation of probability theory is grounded in Cox' (1946) theorem, which lays the groundwork for a demonstration that any calculus of standard logical reasoning must be isomorphic with probability theory (Van Horn 2003).

The central idea is that a probability density function f_x (or probability mass function in the discrete case) places higher weights (higher density) on events that are deemed more likely to occur. Let $X \in \Omega^{d_x}$ be a random vector of dimension d_x representing some hydrometeorological scenario (This article uses the following conventions: (1) uppercase letters represent random variables and lowercase letters represent realizations of those random variables, (2) lowercase Greek symbols represent data, (3) uppercase Greek symbols represent a collection of repeated random samples of some probability distribution, (4) boldface letters and symbols represent vectors or matrices, and (5) lowercase “d” with a subscript indicates the dimension of the vector in the subscript, with the exception of vectors that are concatenated random samples, which have dimension equal to the sample size.), then the probability density function (pdf) f_x yields a measure that is proportional to our belief that the true value of X lies in some subset $\mathcal{R} \subset \Omega$, called a *probability*, according to:

$$P(X \in \mathcal{R}) = \int_{\Omega} 1_{x \in \mathcal{R}} f_x(x) dx. \quad (1.1)$$

If $X \in \mathcal{R} \subset \mathbb{R}$ is scalar and \mathcal{R} is continuous, then this reduces to:

$$P(X \in [a, b]) = \int_a^b f_x(x) dx \quad (1.2)$$

If the event that we want to represent with the random variable X is repeatable in some sense, then it is sufficient to interpret probabilities as the limiting frequency of each possible outcome relative to the set of all possible outcomes. If, however, we are interested in a forecasting a one-off event, then any probabilities we assign are epistemic – they are “*related in part to [our] ignorance, in part to our knowledge*” (Laplace 1902). In reality, no event is truly repeatable, and the frequentist perspective applies only to situations where our a priori information or a priori understanding of the relevant context does not change between forecasts.

1.2 The General Forecasting Problem

A model m translates a vector of input data v into a probability distribution over hydrometeorological predictions X , and the objective of a forecaster is to estimate the distribution f_x given uncertainties related to the model and its inputs. The approach that is consistent with Kolmogorov’s axioms (Kolmogorov 1956), which provide the set-theoretic foundation of probability theory, is to estimate the prediction distribution f_x conditional on model input data u by marginalizing over a random variable M that represents a choice of model:

$$f_x(X | v) = \int m(X | v) f_m(m) dm. \quad (2.1)$$

This integration accounts for the fact that we always have some uncertainty related to our choice of model, and we represent that uncertainty by placing a pdf, f_m , over different models or model structures. Any particular model $M = m$ may be empirical, semi-empirical, or based on approximations of first principles, and it may be deterministic or stochastic (where the former is simply a special case of the latter, as mentioned above).

Hidden in (2.1) is a possibility to include uncertainty about model inputs, either due to incomplete measurements or to measurement error. Ultimately, given some measurement data v , any model m may include a probabilistic measurement distribution, f_u , and therefore make probabilistic predictions about X that account for uncertainty in the model inputs. That is, it is always coherent to include the measurement model in any model m that we might want to test, and so Eq. (2.1) is general in this sense; however, we can also notate input data uncertainty explicitly as:

$$f_x(X | v) = \int f_u(u | v) \int m(X | u) f_m(m) dm du \quad (2.2)$$

1.3 The Role of the Ensemble

It will be quite rare for either (2.1) or (2.2) to admit analytic solution. A common numerical approximation is by Monte Carlo integration (Metropolis 1987), whereby independent and identically distributed (iid) samples of $\mathbf{U} \mid \mathbf{v} \sim f_u$ and $\mathcal{M} \sim f_m$ are drawn so that f_x is sampled repeatedly by $x_{i,j} \sim m_i(X \mid \mathbf{u}_j)$, where i is the sample index over models drawn from f_m and j is the sample index of model input realizations drawn from f_u . The challenges related to ensemble forecasting are: (1) estimating and sampling the f_m distribution over different potential models in a way that reasonably approximates our uncertainty about how to represent the governing biogeophysical processes in the hydrometeorological system, (2) estimating and sampling the f_u distribution over models inputs (e.g., parameters, forcings, and initial states) in a way that reasonably approximates our uncertainty about our measured data, and (3) estimating f_x from N Monte Carlo samples from Eq. (2.2), $\{\mathbf{x}_i\}_{i=1:N}$.

Challenge (1) is in the domain of science, that is, it is related to developing representations of dynamic natural systems. The way to improve our representation of biogeophysical processes in forecasting models is to develop and test new mechanistic hypotheses about these systems. Challenge (2) might be a problem of either better representing our physical measurement device and its associated uncertainty, which might come from either testing new measurement models (as hypotheses), or developing new statistical models directly from data. Challenge (3) is a statistical sampling problem. It is important to understand that estimating these distributions is a comprehensive problem. There are no general strategies for estimating uncertainty distributions or probability distributions that are sufficient for every application, but there are generalizable ways to approach these challenges. The rest of this article lays out some of these generalizable strategies that are common in current hydrometeorological applications.

2 Estimating Probability Distributions

2.1 Density Estimation from iid Samples

The first and most basic problem that we address is as follows. Given N iid samples of a random continuous-valued population, $\mathbf{X} = \{\mathbf{x}_i\}_{i=1:N}$, we wish to estimate f_x . That is, if $\mathbf{x}_i \sim f_x$, we will obtain $\hat{f}_x \approx f_x$. Broadly speaking there are two approaches to estimating f_x from \mathbf{X} : parametrically and nonparametrically. Parametric methods assign an a priori functional form to \hat{f}_x , while nonparametric methods attempt to avoid strong a priori assumptions to the extent possible (notice that *nonparametric* does not mean “without parameters,” but instead means that we will make only weak assumptions about the functional form of the distribution).

2.1.1 Parametric Methods

Let's assume that for a particular application f_x may be assumed to have a particular functional form parameterized by $\mathbf{q} \in \mathbb{R}^{d_q}$. Table 1 describes several common parametric probability density functions of scalar random variables, along with their parameters. Note that parametric methods are not restricted to standard families of distributions like what appear in Table 1, and the scientist may hypothesize any pdf that integrates to unity over the range of the random variable. The choice of density function should always be informed by the particular science question or forecasting problem.

Once a density function is chosen, parameters of that distribution are treated as random variables, and Bayes' law can be used to summarize everything we know about \mathbf{Q} as follows:

$$f_{q|x}(\mathbf{Q} | \mathbf{X}) = \frac{f_x(\mathbf{X} | \mathbf{Q}) f_q(\mathbf{Q})}{\int f_x(\mathbf{X} | \mathbf{q}) f_q(\mathbf{q}) d\mathbf{q}}. \quad (3)$$

$f_x(\mathbf{X} | \mathbf{Q} = \mathbf{q})$ is the probability of sampling \mathbf{X} from distribution f_x with parameters \mathbf{q} . f_q is the *prior* distribution (more accurately, it is the *marginal* over parameters of the joint distribution over parameters \mathbf{Q} and data \mathbf{X}), f_x is the *likelihood*, and $f_{q|x}$ is the *posterior*.

Our job is to estimate \mathbf{Q} conditional on \mathbf{X} ; however, (3) is typically intractable depending on the choice of likelihood and marginal distributions. Further, we may choose either to select a single “best” set of parameter values, $\hat{\mathbf{q}} \sim f_{q|x}$, to obtain \hat{f}_x or to consider the entire posterior distribution. The remainder of this section will outline a few common strategies for dealing with these issues.

Efficient Estimators

Once a parametric form of \hat{f}_x has been chosen, we can define a *decision rule* (also called an *estimator*) in the form of a function \mathcal{d} that acts on the sample \mathbf{X} and returns a parameter estimate $\hat{\mathbf{q}}$; $\hat{\mathbf{q}} = \mathcal{d}(\mathbf{X})$. Estimators do not consider the full posterior distribution $f_{q|x}$, but instead return a nonrandom sample of this distribution with particular qualities determined by the decision rule.

An *efficient* estimator is one that minimizes (maximizes) some particular criteria. Efficiency criteria are based on a *loss function* (*utility function*), which determines the objectives of the estimation problem. In this section, we will present some generic loss functions that consider only the first one or two moments of the posterior distribution. Such loss functions are common in classical statistics; however, loss functions used in hydrometeorology often consider real-world applications.

The loss function acts on decision rule \mathcal{d} and also \mathbf{q}^* , which is a hypothetical set of “true” parameters, as: $\ell(\mathcal{d}, \mathbf{q}^*)$. A very basic example of a loss function is bias:

$$\ell_B(\mathcal{d}, \mathbf{q}^*) = \|B(\mathcal{d}, \mathbf{q}^*)\|, \text{ where} \quad (4.1)$$

Table 1 A list of some common parametric probability distributions over continuous random variables, their parameters, and the maximum likelihood estimates of those parameters

Name	Density function	Significance	Parameters	MLE
Gaussian (normal)	$f_x(x) = (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$	Results from the sum of many iid random variables from any other distribution	μ = mean, first moment σ^2 = variance, second moment	$\hat{\mu}_{mle} = N^{-1} \sum_{i=1}^N x_i$ $\hat{\sigma}_{mle} = N^{-1} \sum_{i=1}^N (x_i - \hat{\mu}_{mle})^2$
Lognormal	$f_x(x) = \begin{cases} x\sqrt{2\pi\sigma^2}^{-1} & \exp\left(-\frac{\ln(x)-\mu}{2\sigma^2}\right) \\ 0 & x > \theta \end{cases}$	Results from the product of many positive iid random variables The random variable $z = \ln(x)$ is Gaussian	μ = scale; mean of $\ln(x)$ σ^2 = shape; variance of $\ln(x)$	$\hat{\mu}_{mle} = N^{-1} \sum_{i=1}^N \ln(x_i)$ $\hat{\sigma}_{mle} = N^{-1} \sum_{i=1}^N (\ln(x_i) - \hat{\mu}_{mle})^2$
Wald (inverse Gaussian)	$f_x(x) = \left[\frac{\lambda}{2\pi x^3}\right]^{-\frac{1}{2}} \exp\left(-\frac{\lambda(x-\mu)}{2\mu^2 x}\right)$	Travel time for Brownian motion	μ = mean λ = shape	$\hat{\mu}_{mle} = N^{-1} \sum_{i=1}^N x_i$ $\frac{1}{\hat{\sigma}_{mle}} = (N-1)^{-1} \sum_{i=1}^N (x_i^{-1} - \hat{\mu}_{mle}^{-1})^2$
Pareto (power law)	$f_x(x) = \frac{\alpha x^\alpha}{x^{\alpha+1}}$	Describes many empirical data that do not cluster around a central tendency. The likelihood of	ν = minimum value, i.e., $x \in [l, \infty)$ α = shape	$\hat{\nu}_{mle} = \min_i x_i$ $\hat{\alpha}_{mle} = N \left(\ln \left(\frac{x_i}{\hat{\nu}_{mle}} \right) \right)^{-1}$

		an extreme event decreases proportionally to the log of the size of the event		
Cauchy	$(\pi\lambda)^{-1} \left[1 + \left(\frac{x-\theta}{\lambda} \right)^2 \right]^{-1}$	Example of a heavy-tailed distribution	$\theta = \text{location}$ $\lambda = \text{scale}$ $(\lambda > 0)$	Numerically maximize: $N\ln(\lambda) - \sum_{i=1}^N \ln(\lambda^2 + (x_i - \theta)^2)$
Gamma	$f_x(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} \exp(-\beta x)$	This family of distributions includes the two below (chi-squared and exponential) as special cases.	$\alpha = \text{shape}$ $(\alpha > 0)$ $\beta = \text{rate}$ $(\beta > 0)$	$\hat{\beta}_{mle} = \widehat{\alpha}_{mle} N \left(\sum_{i=1}^N x_i \right)^{-1}$ Numerically maximize: $(\alpha - 1) \sum_{i=1}^N \ln(x_i) - N\ln(\Gamma(\alpha)) - \alpha \ln \left(\sum_{i=1}^N x_i \right) + N\alpha(1 - \ln(\alpha))$
Chi-squared	$f_x(x) = 2^{-\frac{k}{2}} \Gamma(\frac{k}{2})^{-1} x^{\frac{k}{2}-1} \exp\left(-\frac{x}{2}\right)$	Sum of squares of iid Gaussian random variables with zero mean and unit variance	$k = \text{degrees of freedom} - \text{the number of squared iid Gaussians that are summed}$	This is not frequently used because we typically know how many iid Gaussians were summed
Exponential	$f_x(x) = \lambda \exp(-\lambda x)$	Waiting times between successful trials in a binary experiment	$\lambda = \text{rate}$	$\hat{\lambda}_{mle} = N \left(\sum_{i=1}^N x_i \right)^{-1}$

$$\mathcal{B}(\mathcal{D}, \mathbf{q}^*) = E[\hat{\mathbf{q}}] - \mathbf{q}^* \text{ and} \quad (4.2)$$

$$E[\hat{\mathbf{q}}] = \int_{\Omega_N} d(\mathbf{X}) f_{q|x}(d(\mathbf{X})) d\mathbf{X}. \quad (4.3)$$

Often, there are several estimators that are efficient according to the bias criteria. For example, it is often the case that we may find several unbiased estimators for a particular parametric distribution f_x given sample \mathbf{X} . An unbiased estimator simply means that if the estimator acts on random samples of f_x parameterized by any \mathbf{q}^* , then the expected value of $\hat{\mathbf{q}}$ is always \mathbf{q}^* . The sample mean and sample variance are unbiased estimators of the true mean and true variance for Gaussian distributions.

Since biasedness reflects accuracy, we might additionally consider the second moment of $f_{q|x}$, which reflects precision of $\hat{\mathbf{q}}$. For instance, if we minimize the second moment of $f_{q|x}$:

$$\ell_V(\mathcal{D}, \mathbf{q}^*) = \text{trace}[var(\hat{\mathbf{q}})] = \text{trace}\left[E\left[(\hat{\mathbf{q}} - E[\hat{\mathbf{q}}])^2\right]\right] \quad (5)$$

under an unbiased constraint:

$$\ell_V(\mathcal{D}, \mathbf{q}^*) = \text{trace}\left[E\left[(\hat{\mathbf{q}} - \mathbf{q}^*)^2\right]\right] \quad (6)$$

we obtain a *minimum variance unbiased estimator* (MVUE), $\hat{\mathbf{q}}_{mvue}$. MVUE do not always exist because different estimators may have different variances depending on the true parameter \mathbf{q}^* , so that one estimator may have the lowest variance of all possible estimators for only some particular values of \mathbf{q}^* . Even if an MVUE exists, it is often difficult to find, and the concept is not especially practical for most hydrometeorological applications.

A more practical approach is to minimize an efficiency criterion that *balances* accuracy and precision. For example, the mean-squared error (MSE) considers bias and variance of $\hat{\mathbf{q}}$, but does not require the decision rule to be an unbiased estimator:

$$\begin{aligned} \ell_{mse}(\mathcal{D}, \boldsymbol{\theta}^*) &= E\left[\text{trace}\left[\sum_{i=1}^{d_\theta} (\hat{\mathbf{q}}_i - \mathbf{q}^*)(\hat{\mathbf{q}}_i - \mathbf{q}_i^*)^T\right]\right] \\ &= \ell_V(\mathcal{D}, \mathbf{q}^*) + \ell_B(\mathcal{D}, \mathbf{q}^*)^2. \end{aligned} \quad (7)$$

An estimator that minimizes MSE and is also unbiased is an MVUE.

Strictly speaking, an estimator may be efficient by any chosen criteria (any loss function); however, without additional context, it is often safe to assume that an unbiased estimator is said to be efficient if its (co)variance achieves the Cramér-Rao bound:

$$\text{cov}(\hat{\mathbf{q}}_i) \geq \mathcal{I}(\mathbf{q})^{-1} \quad (8)$$

where $I(\mathbf{q})$ is the *Fisher information* of the parameter θ . The Fisher information matrix is such that the m, n^{th} entry is:

$$I(\mathbf{q})_{m,n} = -E\left[\frac{d^2 \ln f_x(\mathbf{X} | \mathbf{q})}{dq_n dq_m}\right] \quad (9)$$

The intuition behind this is that the second derivative, which represents curvature, of the likelihood gives us some idea of how accurate our parameter estimate may be – likelihood functions with greater curvature are harder to maximize. Any unbiased estimator that achieves the Cramér-Rao bound for all values of \mathbf{q}^* is a MVUE.

Maximum a Posteriori Estimators

In practice, it is rare for hydrometeorological applications to look for efficient estimators like what are described in Sect. 2.1.1.1. A more intuitive efficiency criterion, and one that is sometimes used in real applications, is to maximize the probability of the distribution parameters:

$$\hat{\mathbf{q}}_i = \underset{\mathbf{q}}{\operatorname{argmax}} f_{q|x}(\mathbf{q} | \mathbf{X}). \quad (10)$$

This is called *maximum a posteriori* (MAP) estimation. The MAP estimate $\hat{\mathbf{q}}$ is typically found by numerical methods using Eq. (3). Notice that the integral in the denominator of (3) is constant with respect to the unknown parameters so that for all \mathbf{q} :

$$f_{q|x}(\mathbf{Q} | \mathbf{X}) \propto f_x(\mathbf{X} | \mathbf{Q}) f_q(\mathbf{Q}), \quad (11)$$

This greatly simplifies the process of finding MAP parameters.

The most common MAP estimators are *maximum likelihood estimators* (MLE), $\hat{\mathbf{q}}_{\text{MLE}}$, and since we began by assuming that all of our samples $\{\mathbf{x}_i\}$ are independent, we may treat the probability of the full sample X as a product of the probability of each of the samples \mathbf{x}_i :

$$f_x(\mathbf{X} | \mathbf{q}) = \prod_{i=1}^N f_x(\mathbf{x}_i | \mathbf{q}). \quad (12)$$

If we assume that the prior f_q is uniform, then the posterior distribution $f_{q|x}$ is proportional to the likelihood f_x . In this case, to find the MAP estimate it is sufficient to find the parameter set $\hat{\mathbf{q}}_{\text{MLE}}$ that maximizes $f_x(\mathbf{X} | \mathbf{Q})$.

For certain f_x distributions, the MLE can be obtained analytically; Table 1 lists several examples of MLEs for distributions over scalar random variables. In most cases, however, the MLE must be approximated by maximizing f_x numerically. Either way, the method to find the MLE is typically to maximize the log-likelihood, which allows us to use the sum over individual samples x_i , rather than the product:

$$\hat{\mathbf{q}}_{mle} = \underset{\theta}{\operatorname{argmin}} (\mathcal{L}(\mathbf{q}, \mathbf{X})), \quad (13.1)$$

$$\mathcal{L}(\mathbf{q}, \mathbf{X}) = - \sum_{i=1}^N \ln[f_x(\mathbf{x}_i | \mathbf{q})]. \quad (13.2)$$

This is equivalent to maximizing (11) when \mathbf{X} are iid and f_q is uniform. Analytical solutions, when possible, are obtained by zeroing the gradient of $\mathcal{L}(\mathbf{q}, \mathbf{X})$ w.r.t \mathbf{q} .

MLEs are *consistent*, meaning that as the number of samples N grows, $\hat{\mathbf{q}}_{mle}$ converges to \mathbf{q}^* (assuming that f_x takes the prescribed parametric form. Stated formally, for any $\varepsilon > 0$:

$$\lim_{N \rightarrow \infty} (P(|\hat{\mathbf{q}}_{mle} - \mathbf{q}^*| \geq \varepsilon)) = 0. \quad (14)$$

Additionally, as N grows the distribution over $\hat{\mathbf{q}}_{mle}$ (typically) tends to a normal distribution with variance given by the Cramér-Rao bound. Therefore, the MLE is said to be *asymptotically efficient*. It is important to note, however, that the MLE is not Cramér-Rao efficient for any finite sample size. In particular, the distribution over the MLE converges according to:

$$\sqrt{N} (\hat{\mathbf{q}}_{mle} - \mathbf{q}^*) \rightarrow \mathcal{N}(0, I^{-1}) \quad (15)$$

where $\mathcal{N}(0, I^{-1})$ is the normal (Gaussian) distribution with mean zero and variance given by the inverse of the Fisher information. This means that the MLE converges to the true parameter estimate at a rate of \sqrt{N} .

Markov Chain Monte Carlo

So far, we have discussed how to find a single parameter set with certain properties from the posterior distribution $\hat{\mathbf{q}} \sim f_{q|x}$. In many cases, we might want to consider the entire posterior distribution over distribution parameters $f_{q|x}$, instead of just a single estimator from that distribution. Again, in certain cases, analytical solutions to (3) are possible, depending on the parametric form of the likelihood and prior – these are called *conjugate families*. Fink (1997) gave a thorough overview of the concept of conjugates, including several practical examples. In most cases, however, the posterior must be estimated numerically. Markov Chain Monte Carlo (MCMC) is by far the most common and powerful approach to numerically approximate Bayes' law.

Any feature, say g , of the posterior distribution that we might be interested in will have the property:

$$E[g(\mathbf{Q}) | \mathbf{X}] = \int g(\mathbf{q}) f_{q|x}(\mathbf{q} | \mathbf{X}) d\mathbf{q}, \quad (16)$$

which may be estimated by Monte Carlo integration:

$$E[g(\mathbf{Q})| \mathbf{X}] \approx N_q^{-1} \sum_{i=1}^{N_q} g(\mathbf{q}_i), \quad (17)$$

The *law of large numbers* states that given an iid sample $\boldsymbol{\Theta} = \{\mathbf{q}_i\}_{i=1, \dots, N_q}$ of $f_{q|x}$ (17) converges to the true expected value:

$$P\left(\lim_{N_q \rightarrow \infty} \left(\sum_{i=1}^{N_q} g(\mathbf{q}_i) \right) = N_q E[g(\mathbf{Q})| \mathbf{X}] \right) = 1 \quad (18)$$

This is important because it means that if we sample $f_{q|x}(\mathbf{q}| \mathbf{X})$ enough, we can estimate any feature of the parameter space. The idea is to approximate the parameter sample $\boldsymbol{\Theta}$ using a *Markov chain* in the sense that we may obtain $\hat{\boldsymbol{\Theta}}$ that contains samples of \mathbf{Q} in proportions that are approximately correct according to $f_{q|x}$. A Markov chain is an ordered sequence of random variables (called states), say $\{\mathbf{Q}_0, \mathbf{Q}_1, \mathbf{Q}_2, \dots\}$, where each \mathbf{Q}_t depends on the previous \mathbf{Q}_{t-1} according to some stationary (i.e., the distribution does not change with t) transition density $\pi(\mathbf{Q}_t | \mathbf{Q}_{t-1})$. The dependence of each state on the initial state is denoted $\pi^{(t)}(\mathbf{Q}_t | \mathbf{Q}_0)$, which, given a long enough sequence, converges to a stationary distribution: $\lim_{t \rightarrow \infty} \pi^{(t)} = \pi^*$. In this case, an ergodic average can be substituted for (17) as long as $\pi^* = f_{q|x}$.

The *Metropolis-Hastings* algorithm provides one of the earliest and most general ways to do this. This algorithm is such that a candidate for \mathbf{Q}_{t+1} , say \mathbf{q}_{t+1} is drawn randomly from a *proposal density* $\pi'(\mathbf{q}_{t+1} | \mathbf{q}_t)$, and this state is accepted (i.e., $\mathbf{Q}_{t+1} = \mathbf{q}_{t+1}$) with probability:

$$\alpha(\mathbf{q}_{t+1} | \mathbf{q}_t) = \min\left(1, \frac{f_{q|x}(\mathbf{q}_{t+1} | \mathbf{X}) \pi'(\mathbf{q}_t | \mathbf{q}_{t+1})}{f_{q|x}(\mathbf{q}_t | \mathbf{X}) \pi'(\mathbf{q}_{t+1} | \mathbf{q}_t)}\right). \quad (19)$$

If \mathbf{q}_{t+1} is not accepted, then the Markov state is not changed and $\mathbf{Q}_{t+1} = \mathbf{q}_t$. It turns out that any proposal density will result in $\pi^* = f_{q|x}$, so that the Metropolis-Hastings algorithm will always return a sample $\boldsymbol{\Theta}$ that can be applied to estimate the expected value of any property of $f_{q|x}$ according to (17). Pseudo-code for the Metropolis Hastings algorithm is as follows:

```

INITIALIZE  $\mathbf{Q}_0 = \mathbf{q}_0$ 
REPEAT incrementing  $t$ :
    SAMPLE  $\mathbf{q}' \sim \pi'(\mathbf{Q} | \mathbf{q}_{t-1})$ 
    SAMPLE  $U$  (scalar) from a uniform distribution on  $[0, 1]$ 
    IF  $U \leq \alpha$  assign  $\mathbf{q}_t = \mathbf{q}'$ 
    OTHERWISE assign  $\mathbf{q}_t = \mathbf{q}_{t-1}$ 
END

```

The rate of convergence to a sample of the target distribution $f_{q|x}$ is greatly influenced by the choice of π' . Of course, the best proposal distribution is always $\pi' = f_q$; however, this is impossible, since by definition we do not know f_q .

The primary factor to consider when choosing π' is the *acceptance rate*, which is the frequency at which the Markov chain changes state. A high acceptance occurs when π' is narrow, and means that it will take a long time to sample the tails of $f_{q|x}$. A low acceptance rate reduces independence between samples and therefore also requires longer Markov chains for reliable ergodic approximations. A good rule of thumb for Gaussian proposal distributions is to target an acceptance rate of approximately 25% (Gelman et al. 1996).

There have been numerous adaptations of the Metropolis-Hastings algorithm for various MCMC problems, most of which are interested in increasing efficiency, especially for high dimensional problems (Neal 1993). For example, if π' is symmetric (e.g., a multivariate normal distribution) so that $\pi'(\mathbf{q}_t | \mathbf{q}_{t+1}) = \pi'(\mathbf{q}_{t+1} | \mathbf{q}_t)$, then α reduces to the (bounded) ratio of $f_{q|x}(\mathbf{q}_{t+1} | \mathbf{X}) / f_{\theta|x}(\mathbf{q}_t | \mathbf{X})$. This is called the Metropolis sampler, and works well when π' is a good approximation of f_q .

To deal with large dimensional problems, it is not necessary to treat a multivariate \mathbf{Q} jointly. In this case, an acceptance probability is computed for each element of \mathbf{q}_t individually:

$$\alpha_i(q_{t+1,i} | \mathbf{q}_t) = \min\left(1, \frac{f_{q|x}(q_{t+1,i} | \mathbf{q}_{t,\sim i}, \xi) \pi'(q_{t,i} | \mathbf{q}_{t+1}, \mathbf{q}_{t,\sim i})}{f_{q|x}(q_{t,i} | \mathbf{q}_{t,\sim i}, \xi) \pi'(q_{t+1,i} | \mathbf{q}_t)}\right), \quad (20)$$

where $\mathbf{q}_{t,\sim i}$ denotes all except the i^{th} element of \mathbf{q}_t . This is called *single-component* Metropolis Hastings, and it especially effective when the various conditional distributions of $f_{q|x}$ are calculable. The advantage of this method is that state changes in the Markov chain are induced much more frequently than when the entire joint distribution is considered.

The *Gibbs sampler*, one particularly popular implementation of single-component MCMC, uses the proposal distribution $\pi'(q_{t+1,i} | \mathbf{q}_t) = f_{q|x}(q_{t,i} | \mathbf{q}_{t,\sim i}, \mathbf{X})$, which results in an acceptance probability of 1 always. Gibbs sampling consists of sampling purely from the conditional distributions of $f_{q|x}$.

MCMC is the most general method for estimating posterior distributions according to (3). Although somewhat computationally expensive, these methods can be applied to almost any density estimation problem, and they return a random sample of the entire posterior distribution, rather than a single nonrandom sample like other methods described in this section. There are many types of MCMC algorithms – a modern and comprehensive implementation-oriented overview is given by Brooks et al. (2011), and application-ready code is described by Vrugt (2016).

2.1.2 Nonparametric Methods

It is possible to estimate f_x without specifying an a priori parametric form. There are a few ways to do this (e.g., maximum entropy methods); however, here we will discuss two of the most common: *empirical distributions* and *kernel density estimation*.

Empirical Distributions and Copulas

The *cumulative distribution function* (CDF) of a random variable \mathbf{X} is the integral of the density function:

$$F_x(\mathbf{X}) = \int_{-\infty}^{\mathbf{X}} f_x(\mathbf{x}) d\mathbf{x}, \quad (21)$$

and defines a probability may be defined on a half-space:

$$F_x(\mathbf{x}) = P(\mathbf{X} \leq \mathbf{x}). \quad (22)$$

The distribution f_x may be retrieved as the derivative of the CDF. An *empirical distribution* is estimated directly from \mathbf{X} by counting the number of samples whose value lies in the half-space:

$$\hat{F}_x(\mathbf{X}) = N^{-1} \sum_{i=1}^N \mathbf{1}_{x_i \leq \mathbf{x}}. \quad (23)$$

The Glivenko-Cantelli theorem states that the \hat{F}_x converges uniformly to F_x :

$$P\left(\lim_{N \rightarrow \infty} \left(\sup_{\mathbf{x} \in \Omega} \left(|\hat{F}_x(\mathbf{x}) - F_x(\mathbf{x})| \right) \right) = 0\right) = 1. \quad (24)$$

Empirical distributions are relatively straightforward to estimate in low dimensions; however, they are incredibly hard to estimate when d_x is large, due to what is called the *curse of dimensionality* (Bellman 2003); it is simply impractical to sample the entirety of a high-dimensional space densely enough to estimate an empirical density function.

One common work-around is to estimate the marginal distributions individually over the dimensions of \mathbf{X} and then use parametric methods to estimate the relationship between these marginal distributions. This is called a *copula*. The CDF of any random variable is itself a uniform random variable (this is called the *probability integral transform*):

$$P(a \leq F_x(X) \leq b) = b - a. \quad (25)$$

Remember that F_x is always bounded between 0 and 1, and so a and b must also be similarly bounded. A copula defines the joint distribution (usually, but not always, using parametric methods) between the uniformly distributed $F_{x_j}(X_j)$, where j indexes the d_x dimensions of the random variable \mathbf{X} . Very often a Gaussian copula is used. This represents a compromise between parametric and nonparametric density estimation.

Kernel Density Estimation and Histograms

The idea behind kernel density estimation is to assign finite density to a small window around each sample, so that at any given value of $\mathbf{X} = \mathbf{x}$:

$$\hat{f}_x(\mathbf{X} = \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N K(\mathbf{x}, \mathbf{x}_i). \quad (26)$$

The function K is called a *kernel* and must take some parametric form. The kernel function is parameterized by *hyperparameters*, \mathbf{h} , which are typically estimated using a MAP approach.

The most basic kernel density estimator is the sum of Kronecker delta functions. This discretizes the random variable such that only sampled values are assigned positive probability:

$$\hat{f}_x(\mathbf{X} = \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \delta(\mathbf{x}, \mathbf{x}_i) \text{ where} \quad (27.1)$$

$$\delta(\mathbf{x}, \mathbf{x}_i) = \begin{cases} 1; & \mathbf{x} = \mathbf{x}_i \\ 0; & \mathbf{x} \neq \mathbf{x}_i \end{cases}. \quad (27.2)$$

The Kronecker sum has no hyperparameters.

A *histogram* discretizes the random variable using a kernel function \tilde{h} . The \tilde{h} kernel partitions the domain of the random variable into M rectangles with edges $\mathbf{a}_1, \dots, \mathbf{a}_{M+1}$, and then counts the number of samples that fall into each rectangle:

$$\hat{f}_x(\mathbf{X} = \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \tilde{h}(\mathbf{x}, \mathbf{x}_i) \text{ where} \quad (28.1)$$

$$\tilde{h}(\mathbf{x}, \mathbf{x}_i) = \frac{1}{N} \sum_{j=1}^M \sum_{i=1}^N \mathbf{1}_{\mathbf{x}_i \in [\mathbf{a}_j, \mathbf{a}_{j+1}]} \mathbf{1}_{\mathbf{x} \in [\mathbf{a}_j, \mathbf{a}_{j+1}]} \quad (28.2)$$

There are several ways to estimate the set of rectangle edges $\mathbf{h} = \{\mathbf{a}_i\}_{i=1, \dots, M}$, which are the histogram hyperparameters. For example, we may choose an efficiency criterion for the hyperparameter representing the width of the (scalar) histogram intervals. One popular efficiency criterion is the *mean integrated squared error* (MISE):

$$MISE(a_i - a_j) = E \left[\int (\hat{f}_x - f_x)^2 dx \right], \quad (29)$$

where the expected value is over the unknown true distribution f_x . If the random variable is scalar and comes from a normal distribution, then minimizing the MISE yields an interval width of $a_i - a_j = 3.5 \sqrt{\text{var}(X)} N^{-\frac{1}{3}}$; this is called Scott's rule (Scott 2004).

Alternatively, continuous density estimates may be obtained by centering each (symmetric) kernel k at a sample so that:

$$\hat{f}_x(\mathbf{X} = \mathbf{x}) = \frac{1}{N} \sum_{i=1}^N \hat{k}(\mathbf{x} - \mathbf{x}_i). \quad (30)$$

This allows for smooth density estimates of continuous random variables. Under some mild constraints on \hat{k} and f_x (Silverman 1978), \hat{f}_x converges uniformly almost surely to f_x :

$$P\left(\lim_{N \rightarrow \infty} \left(\sup_{\mathbf{x} \in \Omega} |\hat{f}_x(\mathbf{x}) - f_x(\mathbf{x})| \right) = 0\right) = 1. \quad (31)$$

Perhaps most commonly, a Gaussian kernel with mean x_i and variance σ_k^2 is used. In this case, σ_k – called the kernel *bandwidth* – is the hyperparameter and is estimated either numerically by MAP methods or using some ad hoc efficiency criteria like MISE. For example, if the true underlying distribution is Gaussian, bandwidth which minimizes the MISE is $\sigma_k = \left(\frac{4}{3}\left(\sqrt{\text{var}(X)}\right)^5 N^{-1}\right)^{\frac{1}{5}}$.

2.2 Estimating Distributions over Models and Their Parameters

So far, we have discussed how to estimate parameters and hyperparameters of parametric and nonparametric probability distributions, respectively, given data representing samples of the distribution we want to estimate. These methods are directly applicable to problems of estimating parameters of uncertainty distributions over data, including model inputs \mathbf{U} . It is important to notice that these methods can also be generalized directly to the problem of estimating parameters of any process model or dynamical systems model \mathbf{m} (e.g., biogeophysical, ecohydrological, hydrometeorological) by simply noting that $\mathbf{m}(\mathbf{X}|\mathbf{u}, \mathbf{q})$ is itself a distribution. If we want to estimate parameters \mathbf{q} of model \mathbf{m} , then we simply use $\mathbf{m}(\mathbf{X}|\mathbf{u}, \mathbf{q})$ as the likelihood. This is probably the most common way that methods – especially MAP and MCMC methods – described above are used in hydrometeorological applications. We will return to a discussion of estimating parameters dynamical systems models in Sect. 2.2.2, but first will discuss estimating the distribution over models notated f_m in Eq. (3).

2.2.1 Model Structures

Suppose that we have two or more models of a particular dynamical system: $\{\mathbf{m}_i\}_{i=1, \dots, M}$. This family of model might come from a set of competing hypotheses about the system, from a set of existing models built by different modeling groups, or the scientist might assume some continuous distribution over response surfaces (e.g., Rasmussen and Williams 2006). The probability calculus then allows us to associate each model with some level of *relative* belief. Probability calculus does not provide a measure of *absolute* belief, because we can never test distributions over all

possible models, simply because we do not ever have access to all potential or all possible models of a dynamical system.

Again, Bayes' law provides the foundation for assigning probabilities to any family of candidate models given some set of observation data. Without loss of generality let's consider a case consisting of model inputs \mathbf{v} and observations of system responses ξ corresponding to model predictions x . Then Bayes' law gives us, up to a proportionality constant (remember, we are only able to quantify relative belief, or relative support over a family of models), the following expression for a distribution over models:

$$f_{m|\mathbf{v},\xi}(m_i|\mathbf{v},\xi) \propto f_m(m_i)m_i(\xi|\mathbf{v}). \quad (32.1)$$

In this expression, each model m_i provides a distribution over the value of the observed response data given the observed input data. As mentioned above, there may be error in observing model inputs, and it is necessary (Liu and Gupta 2007) to explicitly consider a distribution over random variable inputs \mathbf{u} :

$$f_{m|\mathbf{v},\xi}(m_i|\mathbf{v},\xi) \propto f_m(m_i) \int m_i(x|\mathbf{u})f_u(u|\mathbf{v})d\mathbf{u}. \quad (32.2)$$

Error in observations of model output or model response cannot be treated similarly because our goal is to estimate the true system response, not necessarily to predict error-affected observations. Therefore, we must explicitly notate a distribution over errors in data ξ as f_ξ :

$$f_{m|\mathbf{v},\xi}(m_i|\mathbf{v},\xi) \propto f_m(m_i) \int f_\xi(\xi|x) \int m_i(x|\mathbf{u})f_u(u|\mathbf{v})d\mathbf{u}dx. \quad (32.3)$$

Here, x represents model predictions corresponding to data ξ , but under the assumption that data ξ is affected by observation error according to distribution f_ξ . We might use techniques similar to those outlined in Sect. 2.1 to infer distributions like f_ξ and f_u .

The practice of science, writ large, as I see it is to improve the posterior distribution over potential models of complex systems, $f_{m|\mathbf{v},\xi}$, in Eq. (32.3) by collecting new data and by proposing (sampling) new models. The job of a forecaster, as I see it, is to use science-informed models to estimate predictive distributions like f_x from Eq. (3).

2.2.2 Model Parameters

Often in hydrometeorology and hydrology applications, we not only want a distribution over potential models (or model structures), but also a distribution over the parameters of those models. Following Liu and Gupta (2007), we can separate uncertainty in models from uncertainty in their parameters using a distribution over model parameters q , f_q . Notice that we used the random variable Q above to represent parameters of a family of probability distributions, and here we are

using the same random variable to represent parameters of a biogeophysical or hydrometeorological process model – the intent is to emphasize that the underlying estimation problem, underlying estimation theory, and applicable statistical techniques are essentially the same in both cases. In the case of calibrating a process model or dynamical systems model, that model m simply takes the role of a parametric form of a probability distribution over system responses: $\mathbf{x} \sim m(\mathbf{X} | \boldsymbol{\nu}, \boldsymbol{q})$. Again, the process model m might be deterministic, but we can never use m with real observation data without also using a data error model like f_ξ , so that $\boldsymbol{\xi} \sim \int f_\xi(\boldsymbol{\xi} | \mathbf{x}, \boldsymbol{q}) m(\mathbf{x} | \boldsymbol{\nu}, \boldsymbol{q}) d\mathbf{x}$. Thus, because of data uncertainty, even deterministic models yield probability distributions over observed response data – this is true whether or not we also consider uncertainty in input data.

In this case, we may use any of the methods outlined in Sect. 2.1 to estimate model parameters. Just as in the previous section, MCMC is the most general, as it returns the full posterior parameter distribution with the fewest assumptions. Also common are efficient estimators (including MAP and MLE), where the loss function is minimized (maximized) numerically.

Perhaps the most common example is to use a MSE loss function. The procedure is to sample the distribution $\mathbf{x} \sim f_x(\mathbf{X} | \boldsymbol{\nu}, \boldsymbol{q})$ for some parameter set \boldsymbol{q} and then measure loss as:

$$\ell_{mse}(\mathbf{x}, \boldsymbol{\xi}) = E \left[\text{trace} \left[\sum_{i=1}^{d_x} (\mathbf{x}_i - \boldsymbol{\xi}_i)(\mathbf{x}_i - \boldsymbol{\xi}_i)^T \right] \right]. \quad (33)$$

The samples \mathbf{x} can be either samples or some other estimator of f_x . Most often m_i is deterministic, ℓ_{mse} substitutes for the f_ξ error distribution, and uncertainty in inputs (via f_u) is not considered, so that f_x is a Dirac distribution. Notice that minimizing the ℓ_{mse} loss function has an equivalent MAP interpretation where f_ξ is a squared exponential (e.g., Gaussian). The function ℓ_{mse} is almost always minimized using numerical methods.

For random variables that can take on k discrete values, $\{\mathbf{a}_i\}_{i=1, \dots, k}$, we may use a squared error criterion called the *Brier score*:

$$\ell_{bs}(\mathbf{x}, \boldsymbol{\xi}) = \sum_{j=1}^k \sum_{i=1}^N \left[(f_x(\boldsymbol{\xi}_i = \mathbf{a}_j | \mathbf{m}, \boldsymbol{\nu}) - \mathbf{1}_{\boldsymbol{\xi}_i = \mathbf{a}_j})^2 \right]. \quad (34)$$

Again, squared errors like MSE and the Brier score are simply examples of loss functions that compare observations with samples from the likelihood.

Another possibility is to use a more general MAP procedure. In the general situation, we would minimize (numerically):

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{\xi}) = -\ln \left(f_q(\boldsymbol{q}) \int f_\epsilon(\boldsymbol{\xi} | \mathbf{x}) m_i(\mathbf{x} | \boldsymbol{\nu}, \boldsymbol{q}) d\mathbf{x} \right). \quad (35.1)$$

Usually, we might assume a uniform prior distribution over model parameters:

$$\mathcal{L}(\boldsymbol{q}, \boldsymbol{\xi}) = -\ln \left(\int f_o(\boldsymbol{\xi} | \boldsymbol{x}) m_i(\boldsymbol{x} | \boldsymbol{\nu}, \boldsymbol{q}) d\boldsymbol{x} \right). \quad (35.2)$$

2.2.3 Bayesian Model Averaging

When there are many candidate models (either because we propose a finite number of models a priori or because we have sampled a continuous distribution over models a finite number of times), then – to reiterate from Sect. 1 – $f_{m|\boldsymbol{\nu}, \boldsymbol{\xi}}$ from (32.1, 32.2, 32.3) may act as model weights that are proportional to our *relative beliefs* about individual candidate models. Once we have calculated these model weights for all candidate models, the posterior distribution may be used to make forecasts:

$$f_{x|\boldsymbol{\nu}, \boldsymbol{\xi}}(\boldsymbol{X} | \boldsymbol{v}, \boldsymbol{\xi}, \boldsymbol{v}^*) = \sum_{i=1}^M m_i(\boldsymbol{X} | \boldsymbol{v}^*) f_{m|\boldsymbol{\nu}, \boldsymbol{\xi}}(m_i | \boldsymbol{v}, \boldsymbol{\xi}). \quad (36)$$

Here \boldsymbol{v} represents inputs used for estimating the $f_{m|\boldsymbol{\nu}, \boldsymbol{\xi}}$ distribution, and \boldsymbol{v}^* represents new model input data used to make forecasts outside of the period of calibration/estimation. Equation (36) can deal with diversity and weighting of model parameters, simply by noting that we can notate each different parameter set \boldsymbol{q}_i as a different model m_i , and in fact, the notation m_i can represent any combined sampling of model structures and model parameters.

As long as the true model is in the set of candidates, this will always outperform any individual model given enough training data. Of course, the true model will never be in our set of candidates; however, experience in many types of applications shows that the $f_{x | \boldsymbol{\nu}, \boldsymbol{\xi}}$ distribution often provides better predictions than using individual models (Hoeting et al. 1999).

2.2.4 Akaike Information Criteria

When using Eqs. (32.1, 32.2, 32.3) to derive the model weights (i.e., the probabilistic relative beliefs over different candidate models as $f_{m|\boldsymbol{\xi}, \boldsymbol{v}}$), we have some latitude in choosing a prior distribution f_m . There is a particular school of thought that places a priori preference on “simpler” models over more complex ones. Often this is justified by calling on an Ockham’s Razor type principle, which is really just a statement for this preference.

One way to conceptualize this preference is to prefer models with fewer parameters. Models with more free parameters will almost always do a better job of fitting the data (return a higher likelihood value) than models with fewer parameters. To help avoid over-fitting, and to account for any a priori philosophical preference, the prior can be used to assign some degree of preference to models with fewer parameters. In fact, a prior distribution that considers *only* model parsimony will always lead to the best possible posterior distribution over models using (32.1, 32.2, 32.3) as $N \rightarrow \infty$ (Solomonoff 1964).

One very simple way to account for parsimony is the *Akaike information criteria* (AIC). If we assign a prior distribution over models that is proportional to the exponential of the number of model parameters (K_i) and take the log-transform of (30), we obtain:

$$AIC(m_i) = 2K_i - 2\ln\left(\int f_\epsilon(\xi | \mathbf{x}) m_i(\mathbf{x} | \nu) d\mathbf{x}\right). \quad (37)$$

Predictive distributions like Eq. (36) may be obtained by weighting different models according to their respective AIC.

Equation (37) was originally derived in the context of information theory (Akaike 1974), however, that is beyond the scope of the present discussion. The intention here is to point out that Bayesian model weights do not have to be calculated using a uniform prior.

2.2.5 Data Assimilation

Perhaps the most common application of Bayesian model weighting is in the context of *data assimilation*. Data assimilation is a process of conditioning predictive distributions from dynamic systems models on observation data. Reichle (2008) provided an early review of data assimilation in the Earth sciences, and a more up-to-date treatment is given in the *Fundamentals of Data Assimilation and Theoretical Advances* chapter of this book.

Data assimilation starts by representing a dynamic system using a stochastic differential equation (SDE). A discrete-time solution to this SDE, given by m_{ds} (ds stands for dynamic system) and time-dependent model inputs (e.g., boundary conditions), so that the Markov state \mathbf{Y} at each time step is:

$$f_y(\mathbf{Y}_t | \mathbf{v}_{1:t}) = \int m_{ds}(\mathbf{Y}_t | \mathbf{y}_{t-1}, \mathbf{v}_t) f_y(\mathbf{y}_{t-1} | \mathbf{v}_{1:t-1}) d\mathbf{y}_{t-1}. \quad (38)$$

An observation operator, denoted \tilde{h}_{ds} , translates the model state Y into estimates of one or more time-dependent observed variables ξ_t :

$$f_x(\xi_t | \mathbf{v}_{1:t}) = \int \tilde{h}_{ds}(\xi_t | \mathbf{y}_t, \mathbf{v}_t) f_y(\mathbf{y}_t | \mathbf{v}_{1:t}) d\mathbf{y}_t. \quad (39)$$

\tilde{h}_{ds} accounts for the f_ϵ observation noise, as well as the relationship between the model state Y and the observable variable X . Equations (38, 39) constitute a *hidden Markov model*.

Given some time-indexed observations $\{\xi_t\}_{t \in \mathcal{S}}$ where \mathcal{S} is some subset of τ total simulation time steps, the objective is to condition the posterior distribution over model states through time $f_y(\mathbf{Y}_1 : \tau)$. This is given by another application of Bayes' law:

$$f_{y|x}(\mathbf{Y}_t | \mathbf{v}_{1:t}, \xi_t) \propto \tilde{h}_{ds}(\xi_t | \mathbf{Y}_t, \mathbf{v}_t) \int m_{ds}(\mathbf{Y}_t | \mathbf{y}_{t-1}, \mathbf{v}_t) f_{y|\zeta}(\mathbf{y}_{t-1} | \mathbf{v}_{1:t-1}, \xi_{1:t-1}) d\mathbf{y}_{t-1}. \quad (40)$$

The smoother is typically intractable due to the fact that $\mathbf{Y}_{1:\tau}$ is of very high dimension (number of time steps multiplied by the dimension of the simulator state), and it is necessary to make some simplifying approximations. By far, the most common approximation is to estimate \mathbf{Y}_t independently of future observations $\zeta_{t+1:\tau}$ (this is called a *filter*):

$$f_{y|x}(\mathbf{Y}_{1:\tau} | \mathbf{v}_{1:\tau}, \boldsymbol{\xi}_{1:\tau}) \propto h_{ds}(\boldsymbol{\xi}_{1:\tau} | \mathbf{Y}_{1:\tau}, \mathbf{v}_{1:\tau}) m_{ds}(\mathbf{Y}_{1:\tau} | \mathbf{v}_{1:\tau}). \quad (41)$$

The filter supplies time-dependent marginals over \mathbf{Y}_t rather than the joint distribution over the whole state trajectory $\mathbf{Y}_{1:\tau}$. It is, however, still generally intractable for the same reason as Eq. (3) (analytical solutions only exist in special cases), and MCMC is too expensive to apply recursively at every time step to estimate all of the τ marginals. To simplify Eq. (41) further, we can again use either parametric or nonparametric approximations. The most common parametric and nonparametric filter approximations are described in the following two subsections.

The Ensemble Kalman Filter

The parametric filter approximation used most commonly in hydrometeorology is the *ensemble Kalman filter* (EnKF) (Evensen 2003). If m_{ds} and h_{ds} are both linear and Gaussian then $f_{y|x}$ is Gaussian with an analytic mean and variance. Of course, most hydrometeorologic models are not linear, and nonlinearity in m_{ds} can be accounted for using a Monte Carlo approximation of the integral in (41). In fact, the EnKF approximates the posterior distribution from Eq. (41) as Gaussian, so that the Monte Carlo sample is used to estimate a mean and variance of $f_{y|x}$ according to (17).

The procedure is to sample $m_{ds}(\mathbf{Y}_t | \mathbf{y}_{t-1}, \mathbf{v}_t)$ independently N times; this is called the *background* sample and notated $\mathbf{Y}_t^b \in \mathbb{R}^{d_y, N}$, where d_y is the dimension of the model state. Each background sample is used to predict a value of the observation \mathbf{x}_t by sampling $h_{ds}(\mathbf{x}_t | \mathbf{Y}_t, \mathbf{v}_t)$ (which is Gaussian with covariance \mathbf{R}), which are similarly collectively stored in a background matrix $\mathbf{X}_t^b \in \mathbb{R}^{d_x, N}$. The cross-covariance of the background state and observation is estimated as:

$$\mathbf{P}_t = \frac{1}{N-1} \left(\mathbf{Y}_t^b - \overline{\mathbf{Y}}_t^b \right) \left(\mathbf{X}_t^b - \overline{\mathbf{X}}_t^b \right)', \quad (42.1)$$

where $\overline{\mathbf{Y}}_t^b$ is the sample mean of the background state. The covariance of the observation is estimated as:

$$\mathbf{Q}_t = \frac{1}{N-1} \left(\mathbf{Y}_t^b - \overline{\mathbf{Y}}_t^b \right) \left(\mathbf{Y}_t^b - \overline{\mathbf{Y}}_t^b \right)'. \quad (42.2)$$

N MLEs of \mathbf{Y}_t , called an *analysis* sample and notated $\mathbf{Y}_t^a \in \mathbb{R}^{d_y, N}$, are found by taking each of the N background samples individually to be the mean of the background distribution:

$$\mathbf{Y}_t^a = \mathbf{Y}_t^b \mathbf{K}_t. \quad (42.3)$$

$$\mathbf{K}_t = \mathbf{I}_k + \mathbf{P}_t (\mathbf{Q}_t + \mathbf{R})^{-1} (\boldsymbol{\xi}_t - \mathbf{X}_t^b). \quad (42.4)$$

where $\boldsymbol{\xi}_t$ are N samples from the (Gaussian) observation distribution \mathcal{h}_{ds} : $\boldsymbol{\xi}_t \sim \mathcal{N}[\mathbf{X}_t, \mathbf{R}]$. \mathbf{Y}_t^a constitutes an iid sample of $f_{y|x}$.

INITIALIZE \mathbf{Y}_0 as a sample of initial states

REPEAT incrementing t over the simulation period $1, \dots, \tau$:

SAMPLE the boundary condition $\mathbf{v}_{t,i} \sim f_u$ for $i = 1, \dots, N$

SAMPLE the background state $\mathbf{y}_{t,i}^b \sim m_{ds}(\mathbf{Y}_t | \mathbf{y}_{t-1,i}^a, \mathbf{v}_{t,i})$ by running the model N times

SAMPLE the observation $\mathbf{x}_{t,i}^b \sim \mathcal{h}_{ds}(\mathbf{X}_t | \mathbf{y}_{t-1,i}^b, \mathbf{v}_{t,i})$ N times

STORE $\{\mathbf{y}_{t,i}^b\}$ in \mathbf{Y}_t^b

SET $\mathbf{P}_t, \mathbf{Q}_t, \mathbf{K}_t$

SET $\mathbf{Y}_t^a = \mathbf{Y}_t^b \mathbf{K}_t$

END

The EnKF described by Eq. (42.1), (42.2), (42.3), (42.4) is by far the most common data assimilation algorithm used in the terrestrial earth sciences. It is extremely simple and efficient to implement and can handle nonlinear state-space models in an approximate manner. Further implementation details and also descriptions of several other common data assimilation methods are elsewhere in this handbook.

3 Summary

Science is a process of adapting our beliefs about descriptions of the natural world. Probability calculus is one tool for measuring states or degrees of beliefs about different events and propositions (e.g., data, states of the world, hypotheses). The probability calculus also provides, via Bayes' law, the foundation for innumerable procedures and methods for adapting beliefs states about different hypotheses in the presence of new observation data. It is important to note, however, that no model (probabilistic or otherwise) will ever be correct. The best we can strive for is to quantify what we currently know and what we currently know that we do not know about hydrometeorological systems. Probability calculus is the most common tool for these sorts of quantitative efforts.

Further, in applied sciences like hydrometeorology, models facilitated by science are used to make forecasts that aid decision makers. The probability calculus can also be used to quantify our beliefs about the outcome of a forecasted event. Again, no forecast will ever be correct, so it is important not to assert that the probability distributions that underlie probabilistic forecasts represent a real property of the

forecasted event – in fact, at best they represent a real property of ourselves and our models – specifically, about our own limited knowledge of the potential explanations and/or outcomes of dynamical systems.

The advantage of framing both science and forecasting problems in the context of probability calculus is that there are strong rules for inference. In particular, the probability calculus provides the *only* formal calculus of inference that does not violate the standard rules of logic (Van Horn 2003). This logical system provides a straightforward (at least in theory) way to merge hypotheses (models) with observations for both induction and prediction. Although most of the time, Bayes' law is intractable, modern computers make general approximate solutions, via Monte Carlo methods.

References

- H. Akaike, A new look at the statistical model identification. *IEEE Trans. Autom. Control* **19**, 716–723 (1974)
- R. Bellman, *Dynamic Programming* (Dover Publications, Mineola, 2003)
- S. Brooks, A. Gelman, G. Jones, X.-L. Meng, *Handbook of Markov Chain Monte Carlo* (Taylor & Francis CRC Press, Boca Raton, 2011)
- R.T. Cox, Probability, frequency and reasonable expectation. *Am. J. Phys.* **14**, 1–13 (1946)
- G. Evensen, The ensemble Kalman filter: theoretical formulation and practical implementation. *Ocean Dyn.* **53**, 343–367 (2003)
- D. Fink. A compendium of conjugate priors. Tech Report, 46 pp (1997). <https://www.johndcook.com/CompendiumOfConjugatePriors.pdf>
- A. Gelman, G. Roberts, W. Gilks, Efficient metropolis jumping rules, in *Bayesian Statistics*, 5, ed. by J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith (Oxford University Press, New York, 1996), pp. 599–608
- J.A. Hoeting, D. Madigan, A.E. Raftery, C.T. Volinsky, Bayesian model averaging: a tutorial. *Stat. Sci.* **14**, 382–401 (1999)
- E.T. Jaynes, *Probability Theory: The Logic of Science* (Cambridge University Press, New York, 2003)
- A.N. Kolmogorov, *Foundations of the Theory of Probability* (Chelsea, New York, 1956)
- P.S. Laplace, *A Philosophical Essay on Probabilities* (Chapman & Hall, London, 1902)
- Y.Q. Liu, H.V. Gupta, Uncertainty in hydrologic modeling: toward an integrated data assimilation framework. *Water Resour. Res.* **43**(7), W07401 (2007)
- N. Metropolis, The beginning of the Monte Carlo method. *Los Alamos Sci.* **15**, 125–130 (1987)
- R. M. Neal, Probabilistic inference using Markov Chain Monte Carlo methods. Technical Report CRG-TR-93-1. Department of Computer Science, University of Toronto, Toronto, 1993
- C. Rasmussen, C. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006)
- R.H. Reichle, Data assimilation methods in the Earth sciences. *Adv. Water Resour.* **31**(11), 1411–1418 (2008)
- D. W. Scott, Multivariate density estimation and visualization, in *Handbook of Computational Statistics: Concepts and Methods*, ed. by J. E. Gentle, W. Haerdle, Y. Mori (Springer, New York, 2004), pp. 517–538,
- B.W. Silverman, Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Ann. Stat.* **6**, 177–184 (1978)
- R.J. Solomonoff, A formal theory of inductive inference. Part I. *Inf. Control.* **7**, 1–22 (1964)
- K.S. Van Horn, Constructing a logic of plausible inference: a guide to cox's theorem. *Int. J. Approx. Reason.* **34**, 3–24 (2003)
- J.A. Vrugt, Markov chain Monte Carlo simulation using the DREAM software package: theory, concepts, and MATLAB implementation. *Environ. Model. Softw.* **75**, 273–316 (2016)



Regression Techniques Used in Hydrometeorology

Wei Gong

Contents

1	Introduction	1486
2	Linear Regression	1487
2.1	Single Variable Linear Regression	1487
2.2	Multivariate Linear Regression	1491
2.3	Ridge Regression	1492
2.4	Quantile Regression	1493
3	Nonlinear Regression	1494
3.1	Logistic Regression	1495
3.2	Poisson Regression	1497
4	Machine Learning Methods for Regression	1498
4.1	Artificial Neural Network (ANN)	1498
4.2	Support Vector Machine (SVM)	1501
4.3	Regression Tree and Random Forests	1503
4.4	Multivariate Adaptive Regression Splines (MARS)	1505
4.5	Gaussian Processes Regression	1506
5	Summary	1510
	References	1510

Abstract

Regression methods play an important role in ensemble forecasting. The atmosphere-land-ocean system is complex and dynamical, which makes it difficult to predict the state of hydrometeorological variables deterministically. Consequently, stochastic approaches become useful for hydrometeorological

W. Gong (✉)

State Key Laboratory of Earth Surface Processes and Resource Ecology, Faculty of Geographical Science, Beijing Normal University, Beijing, China

Institute of Land Surface System and Sustainable Development, Faculty of Geographical Science, Beijing Normal University, Beijing, China

e-mail: gongwei2012@bnu.edu.cn

forecasting. As forecast uncertainty is inevitable, it is of key importance to use regression approaches to extract useful information from raw observational data and forecasts from dynamical models while providing an appropriate estimation of the confidence level of the forecasts. Regression methods are usually used in two ways in ensemble forecasting. One is used as a statistical forecasting model, which accounts for the relationships between predictors and historical observation data. Another is used as a post-processor for the forecasts from dynamical models in order to correct various biases in them and to improve their reliability and skill scores. If the statistical relationships between the dynamical forecasts and the observation data exist, the systematic bias and ensemble distribution errors can be corrected, and associated uncertainty can be reduced. The two means of applying regression approaches share a common statistical foundation. This chapter will give a brief introduction to various common linear/nonlinear regression approaches that have been used or can be potentially applicable in ensemble forecasting.

Keywords

Regression · ANOVA · Ridge regression · Quantile regression · Logistic regression · Poisson regression · Gaussian processes regression · Kriging · Regression tree · Multivariate adaptive regression splines · Support vector machine · Artificial neural network

1 Introduction

This chapter provides an overview of the regression methods, which play an important role in hydrometeorological ensemble forecasting. In short, regression is a means for estimating the parameters of parametric models which describe the relationships between independent variable(s), or predictor(s) \mathbf{x} (\mathbf{x} can be a single variable or a vector consisting of multiple variables), and dependent variable or predictand y .

$$y = f(\mathbf{x}, \boldsymbol{\theta}) + e \quad (1)$$

The parametric model $f(\mathbf{x}, \boldsymbol{\theta})$ represents our prior knowledge or assumptions about the nonrandom relationship between \mathbf{x} and y , and e represents the random error that also influences the predictand y . The output of parametric model $f(\mathbf{x}, \boldsymbol{\theta})$ depends not only on the predictor \mathbf{x} but also on parameter vector $\boldsymbol{\theta}$. The error term e can be homoscedastic (independent of \mathbf{x}), or heteroscedastic (depends on \mathbf{x}), and the assumptions about error term are also called error model. The process of regression is to determine the values of $\boldsymbol{\theta}$ with various approaches under various assumptions about $f(\mathbf{x}, \boldsymbol{\theta})$ and e .

In this chapter, a brief introduction to the various regression methods is presented. We begin with the simplest single variable linear regression, together with some basic concepts about analysis of variance (ANOVA), and then extend it

as multivariate linear regression model. Two specific kinds of linear regression called ridge regression and quantile regression are also introduced. After that, the linear regression model will be extended to generalized linear model (GLM), which represents the predictand as a nonlinear function of a linear regression equation.

Besides the classical regression methods, a plethora of the so-called machine learning methods can also be regarded as regression methods. They have more sophisticated assumptions about $f(x, \theta)$ and sometimes have better ability to fit complicated hydrometeorological data. In this chapter we give an introduction about some of these machine learning methods that have been well known in the community of ensemble forecasting: Gaussian processes regression (GPR), regression tree and random forests, multivariate adaptive regression spline (MARS), support vector machine (SVM), and artificial neural network (ANN). The connections between classical regression methods and machine learning methods are presented in the Sect. 5.

2 Linear Regression

2.1 Single Variable Linear Regression

Let's begin with the simplest single variable linear regression, which is also the most frequently used approach in statistical weather forecasting. The single variable linear regression describes the linear relationship between two variables, x , the *independent variable or predictor*, and y , the *dependent variable or predictand*.

The regression procedure seeks a line producing least error for predictand y given predictor x . The most frequently used error criterion is the *sum of squared errors*, so this kind of regression is also called least-squares regression. The expression of the straight line is:

$$\hat{y} = a + bx \quad (2)$$

where \hat{y} means the predicted value of y , a is the *intercept*, and b is the *slope*. a and b are called regression parameters, which are to be estimated via regression. The errors, or residuals, are defined as:

$$e_i = y_i - \hat{y}(x_i) \quad (3)$$

where e_i is the residual of \hat{y} for the data pair (x_i, y_i) . Each e_i is assumed to be an independent random variable with zero mean and constant variance. The constant variance assumption is usually referred as homoscedasticity. If the variance of residuals changes with predictor, it is called heteroscedasticity. Heteroscedastic data sets usually need special manipulation, such as log transformation or Box-Cox transformation, to transform them to homoscedastic data for further analysis. Combining the above two equations yields the regression equation:

$$y_i = \hat{y}(x_i) + e_i = a + bx_i + e_i \quad (4)$$

Consequently, the sum of squared errors can be expressed as:

$$\sum_{i=1}^n (e_i)^2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - [a + bx_i])^2 \quad (5)$$

To estimate a and b , we calculate the minimum value of $\sum_{i=1}^n (e_i)^2$, which can be obtained by setting the derivatives of Eq. (5) with respect to a and b to zero.

$$\left\{ \begin{array}{l} \frac{\partial \sum_{i=1}^n (e_i)^2}{\partial a} = \frac{\partial \sum_{i=1}^n (y_i - a - bx_i)^2}{\partial a} = -2 \sum_{i=1}^n (y_i - a - bx_i) = 0 \quad (\text{a}) \\ \frac{\partial \sum_{i=1}^n (e_i)^2}{\partial b} = \frac{\partial \sum_{i=1}^n (y_i - a - bx_i)^2}{\partial b} = -2 \sum_{i=1}^n x_i(y_i - a - bx_i) = 0 \quad (\text{b}) \end{array} \right. \quad (6)$$

Equation (6) can be rearranged like this:

$$\left\{ \begin{array}{l} \sum_{i=1}^n y_i = na + b \sum_{i=1}^n x_i \quad (\text{a}) \\ \sum_{i=1}^n x_i y_i = a \sum_{i=1}^n x_i + b \sum_{i=1}^n (x_i)^2 \quad (\text{b}) \end{array} \right. \quad (7)$$

Rearranging Eq. (7a), we obtain $a = \frac{1}{n} \left(\sum_{i=1}^n y_i - b \sum_{i=1}^n x_i \right) = \bar{y} - b\bar{x}$. Substituting it into Eq. (7b), we obtain:

$$b = \frac{\sum_{i=1}^n [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n (x_i)^2 - \left(\sum_{i=1}^n x_i \right)^2} \quad (8)$$

In this way, the least square solution for a and b is obtained for Eq. (2).

2.1.1 Analysis of Variance (ANOVA) in Single Variable Linear Regression

Linear regression can provide a lot of interesting information with the help of *analysis of variance* (ANOVA). The statistical foundation of ANOVA is the decomposition of variance. The total variation of predictand y (SST) can be decomposed to

the portion represented by the regression (SSR) and the unrepresented portion due to the residuals (SSE), as shown below:

$$SST = SSR + SSE \quad (9)$$

SST is the abbreviation for “sum of squares, total”. It is defined as the sum of squared deviations of observed y_i around their mean \bar{y} :

$$SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - ny\bar{y}^2 \quad (10)$$

SST represents the overall variability of the observed predictand y_i . Actually, SST is proportional to sample variance, which is defined as $V = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$. The term SSR represents the variability explained by the regression equation, namely, the sum of squared differences between the predictions $\hat{y}(x_i)$ and the sample mean \bar{y} :

$$SSR = \sum_{i=1}^n [\hat{y}(x_i) - \bar{y}]^2 \quad (11)$$

Substituting the regression equation into Eq. (11), we obtain:

$$SSR = b^2 \sum_{i=1}^n (x_i - \bar{x})^2 = b^2 \left[\sum_{i=1}^n x_i^2 - n\bar{x}^2 \right] \quad (12)$$

Equation (12) indicates that a regression line with a small slope will produce small SSR, while a large slope will lead to large SSR. The term SSE is defined as the sum of squared residuals:

$$SSE = \sum_{i=1}^n e_i^2 \quad (13)$$

A typical ANOVA table, like Table 1, provides the following information: the degree of freedom, the sum of squares, the mean squares, and the F ratio. The degree of freedom means the number of independent random variables. The *degree of freedom* of the single variable regression is 1 because it has only one independent variable. Given n totally independently sampled points, the total sum of squares has $n-1$ degrees of freedom because the mean value has been fixed. For regression, the freedom of residuals is $n-2$ because two parameters, a and b , are fixed in the regression equation. The three sums of squares, SST, SSR, and SSE, satisfy the decomposition of variance. The mean squares are defined as sum of squares per degree of freedom. The mean squared error, MSE, is a most commonly used measure of goodness of fit, since it represents the variability of observed y around the predicted regression line and has the same unit as the predictand y . MSE is equal to 0 if the linear relationship is perfect, and SST is equal to SSR since the regression

Table 1 Analysis of variance (ANOVA) table of single variable linear regression

	Degree of freedom	Sum of squares	Mean squares	F ratio
Total	n-1	SST		
Regression	1	SSR	MSR=SSR/1	MSR/MSE
Residual	n-2	SSE	MSE=SSE/(n-2)	

equation explains all of the variation. In the opposite end, there is absolutely no linear relationship, SSR is equal to 0, and MSE is very close to the sample variation.

The *F ratio* defined as MSR/MSE is a frequently used metric of the strength of the regression. The F ratio increases with the strength of the regression because a stronger relationship between x and y leads to larger MSR and relatively smaller MSE. With the assumption that the residuals are independent and follow the Gaussian distribution, the distribution of F ratio follows F distribution (also known as Fisher-Snedecor distribution), and the significant level of linear regression can be qualified with a hypothesis test.

2.1.2 Regression Coefficients

The *coefficient of determination* is also a frequently used measure of goodness of fit. It is defined as the ratio of SSR and SST, as follows:

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST} \quad (14)$$

R^2 can be interpreted as the portion of total variance accounted for the regression equation. $R^2 = 1$ means perfect regression while $R^2 = 0$ means absolutely no linear relationship. The squared root of the determination coefficient, namely, R , is also called (the absolute value of) Pearson correlation coefficient.

The sampling distribution of the regression coefficients, intercept a and slope b , can be obtained from the mean squares of residuals. Suppose that we have different batches of size n from the same data-generating process, then we will obtain different pairs of intercept a and slope b , and the sampling distribution of a and b represent their expectations and variabilities. Under the i.i.d (independent and identically distributed) Gaussian assumptions of residuals, as listed previously, intercept a and slope b are also Gaussian. The expectation and standard deviation of intercept a are:

$$\left\{ \begin{array}{l} \mu_a = a \\ \sigma_a = \text{RMSE} \cdot \sqrt{\frac{\sum_{i=1}^n x_i^2}{n \sum_{i=1}^n (x_i - \bar{x})^2}} \end{array} \right. \quad (15)$$

where RMSE is the abbreviation of root mean squared error, which equals to the squared root of MSE. The expectation and standard deviation of slope b are:

$$\left\{ \begin{array}{l} \mu_b = b \\ \sigma_b = \frac{\text{RMSE}}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2}} \end{array} \right. \quad (16)$$

As indicated by Eqs. (15) and (16), the coefficients estimated by least-square regression are unbiased, and their standard deviation is both dependent on the RMSE. For the slope b , we can carry out a t-test to evaluate the significance of regression. With the null assumption of $b = 0$, if the t-ratio (estimated b to its standard deviation) is not significantly large, the true value of b could plausibly be zero, and the correlation relationship between x and y is plausibly not existing.

2.2 Multivariate Linear Regression

The simplest single variable linear regression can be easily extended to multivariate linear regression as follows:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m \quad (17)$$

where \hat{y} denotes the predicted value of y , b_i denotes the correlation coefficient of each predictor x_i , and m is the number of predictors. For convenience, the intercept is denoted as b_0 . The regression equation can be written in the vector form:

$$\hat{y} = \mathbf{x}^T \mathbf{b} \quad (18)$$

where $\mathbf{x} = (1, x_1, x_2, \dots, x_m)^T$ is the predictor vector and $\mathbf{b} = (b_0, b_1, b_2, \dots, b_m)^T$ is the coefficient vector. Similarly, the coefficients of multivariate regression can also be calculated from minimizing the squared residuals $\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$, where $\|\cdot\|$ is the Euclidean norm. By setting the derivatives to zero, we obtain the equation for regression coefficients:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (19)$$

where \mathbf{X} is the observed data matrix (called design matrix) in which each row is an observation of the predictor vector \mathbf{x} and \mathbf{y} is a column vector of observations of predictand y , as follows:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (20)$$

and n is the number of observations.

The analysis of variance (ANOVA) can also be extended to multivariate linear regression. The SST, SSR, and SSE are defined as:

Table 2 Analysis of variance (ANOVA) table of multivariate linear regression

	Degree of freedom	Sum of squares	Mean squares	F ratio
Total	n-1	SST		
Regression	m	SSR	MSR=SSR/m	MSR/MSE
Residual	n-m-1	SSE	MSE=SSE/(n-m-1)	

$$\begin{aligned} SST &= \sum_{i=1}^n (y_i - \bar{y})^2 \\ SSR &= \sum_{i=1}^n [\hat{y}(\mathbf{x}_i) - \bar{y}]^2 \\ SSE &= \sum_{i=1}^n [y_i - \hat{y}(\mathbf{x}_i)]^2 \end{aligned} \quad (21)$$

The ANOVA table of multivariate regression is shown as Table 2. The degree of freedom of regression is equal to m, the number of predictors, whereas that of the residuals changes to n-m-1. Similarly, the F-ratio follows F(m,n-m-1) distribution, and the strength of regression can be evaluated with a F-test. The definition of determination coefficient $R^2 = \frac{SSR}{SST}$ is also valid for multivariate regression.

Furthermore, each term of the multivariate linear regression can be replaced by nonlinear functions of factors, such as:

$$\hat{y} = b_1 g_1(x_1, \dots, x_s) + b_2 g_2(x_1, \dots, x_s) + \dots + b_m g_m(x_1, \dots, x_s) \quad (22)$$

where x_1, \dots, x_s are the s factors, g_1, \dots, g_m are the m functions of x_1, \dots, x_s , and their weighted sum is equal to the predictand y . The coefficients b_1, \dots, b_m can be calculated from Eq. (19). Equation (22) can represent many special cases, including linear regression like Eq. (17), quadratic regression as follows:

$$\hat{y} = b_0 + \sum_{i=1}^m b_i x_i + \sum_{i=1}^m \sum_{j=1}^m b_{ij} x_i x_j \quad (23)$$

and regression with nonlinear terms:

$$\hat{y} = b_0 + b_1 \log(x_1) + b_2 \sin(x_1) + b_3 \sin(x_2) \cos(x_3) \quad (24)$$

Note that different transformations of the same factor, such as $\log(x_1)$ and $\sin(x_1)$, are treated as different predictors in the regression equation as well as in the analysis of variance.

2.3 Ridge Regression

In the multivariate linear regression, if the correlation matrix $\mathbf{X}^T \mathbf{X}$ is close to unit matrix, the estimation of \mathbf{b} can be unbiased and have minimum variance. However, if

$\mathbf{X}^T \mathbf{X}$ is very far from unit matrix, the calculation of inverse matrix becomes unstable, and the estimated \mathbf{b} does not make sense under the context of physical background. This kind of problem usually happens when two or more predictors are strongly correlated, which is called *multicollinearity*.

To mitigate this problem, Hoerl and Kennard (1970) proposed *ridge regression*, which penalizes the magnitude of the regression coefficients by minimizing the squared residuals plus a scaled sum of squared regression coefficients

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \lambda \|\mathbf{b}\|^2 \quad (25)$$

The solution of ridge regression problem is:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y} \quad (26)$$

where λ is a penalty factor. In the ridge regression, the diagonal of matrix $\mathbf{X}^T \mathbf{X}$ is strengthened so that the process of matrix inversion becomes more stable. If $\lambda \rightarrow 0$, the result of \mathbf{b} will be close to that of classical linear regression, whereas if $\lambda \rightarrow \infty$, \mathbf{b} will be close to zero. Applying the penalty factor λ has a shrinking effect that can correct the abnormally large value of \mathbf{b} , but it can also reduce the contribution of sensitive predictors. The value of λ should be carefully selected in order to give an appropriate estimation of \mathbf{b} . For practical problems, the value of λ is usually determined by a try-and-fail method. First, we should assign the value of λ and estimate \mathbf{b} and then evaluate the prediction of the regression model with another independently generated dataset or use cross-validation, leave-one-out method, etc. The value of λ which minimizes the prediction error is accordingly adopted.

More generally, the ridge regression can be extended to *Tikhonov regularization*, which replaces the penalty factor with a matrix:

$$\|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 + \|\Lambda\mathbf{b}\|^2 \quad (27)$$

The solution of Tikhonov regularization is:

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X} + \Lambda^T \Lambda)^{-1} \mathbf{X}^T \mathbf{y} \quad (28)$$

where Λ is the Tikhonov matrix. For regression problems minimizing the sum of squared residuals, the most common choice of Λ is the scaled identity matrix $\lambda \mathbf{I}$. The ridge regression can be understood as a special case of Tikhonov regularization that is commonly used in regression.

2.4 Quantile Regression

The classical linear regression can give the value of conditional mean value of the predictand given the values of the predictors. Is it possible to give more uncertainty information, such as the conditional quantiles? Quantile regression proposed by

Koenker and Bassett (1978) is able to give the value of any quantiles, or percentiles, with the condition of predictors.

The quantile regression (QR) can predict the conditional quantile values of predictand y on predictor \mathbf{x} . The formula of the QR for a specified quantile p ($0 < p < 1$) is as follows:

$$Q^{(p)}(y|\mathbf{x}) = b_0^{(p)} + b_1^{(p)}x_1 + b_2^{(p)}x_2 + \dots + b_m^{(p)}x_m \quad (29)$$

As the conditional quantiles (such as medians, 25%, 75%, 95%, 5% quantiles) are used instead of conditional means, the quantile regression is robust with regard to outliers. Furthermore, quantile regression can provide a full spectrum of the uncertainty bound of the predictor-predictand relationship, making it a sharp tool in the community of ensemble forecast (López López et al. 2014).

Similar with defining the conditional mean as the solution to the problem of minimizing a sum of squared residuals, the conditional median can also be defined as the solution to the problem of minimizing a sum of absolute residuals. Let \hat{y} denote the predicted value of y and $e_i = y_i - \hat{y}(x_i)$ denote prediction error, namely, the residual of \hat{y} to y_i in the data pair (x_i, y_i) , as we have done in Eq. (3). The loss function $L(e(x)) = L(y - \hat{y}(x))$ is associated with the prediction errors. If $L = e^2$, as having been shown in Eq. (5), the loss function is called squared error loss, and the optimal predictor is the least squares, as we have done in classical linear regression. If $L = |e|$, the optimal predictor is the conditional median, and the optimal predictor is obtained by minimizing the least absolute deviation (LAD) function $\sum_i |y_i - \mathbf{x}_i^T \mathbf{b}|$,

where $\mathbf{b} = (b_0, b_1, b_2, \dots, b_m)^T$ is the coefficient vector.

Furthermore, any conditional quantiles can be obtained by minimizing asymmetrically weighted absolute residuals, as shown below:

$$\min_{\mathbf{b} \in R^m} \left[\sum_{i \in \{i:y_i \geq \mathbf{x}_i^T \mathbf{b}\}} p|y_i - \mathbf{x}_i^T \mathbf{b}| + \sum_{i \in \{i:y_i < \mathbf{x}_i^T \mathbf{b}\}} (1-p)|y_i - \mathbf{x}_i^T \mathbf{b}|\right] \quad (30)$$

where i is the index of predictors/predictand pairs (\mathbf{x}_i, y_i) and p means the p -th quantile ($0 < p < 1$). The median regression is a special case of quantile regression if setting $p = 0.5$. This objective function is non-differentiable, so it cannot be solved by gradient-based methods, such as calculating the derivatives, as we have done in the classical linear regression. This optimization problem can be efficiently solved by linear programming method, such as simplex method, which is guaranteed to yield a solution in a finite number of iterations.

3 Nonlinear Regression

Linear regression models, such as multivariate linear regression, quadratic regression, and linear regression with nonlinear term, are appropriate if the assumption of Gaussian residuals with constant variance is tenable. It is necessary to use nonlinear regression if the

Gaussian assumption is inappropriate. This section introduces a class of regression models usually called generalized linear models (GLMs), which represent the predictand as a nonlinear function of a linear regression equation. A typical GLM looks like:

$$g(\hat{y}) = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m \quad (31)$$

here the link function $g(\hat{y})$ is an invertible function of \hat{y} . The linear regression is a special case of nonlinear regression if $g(\hat{y}) = \hat{y}$. Note that nonlinear regression refers to the models with nonlinear predictand term, while those models with nonlinear terms on the predictor side are still linear regression (with nonlinear terms).

3.1 Logistic Regression

Sometimes we need to predict the probability of an event, such as the probability of rain or no rain. In such a case the observed, the predictand takes two values: 0 (no rain) and 1 (rain). Consequently the residual does not follow Gaussian distribution but Bernoulli. Logistic regression is a frequently used regression model for binary predictand with Bernoulli residuals. By setting a log-odd link function, the logistic regression equation can be written as:

$$\ln\left(\frac{\hat{p}}{1 - \hat{p}}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m \quad (32)$$

where \hat{p} is the event probability, such as the probability of rain. The equation can be rearranged like this:

$$\hat{p} = \frac{\exp(b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m)}{1 + \exp(b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m)} \quad (33)$$

Logistic regression is not the only choice for binary predictand. An alternative choice is probit regression, which uses the inverse Gaussian CDF as the link function, such as:

$$\hat{p} = \Phi(b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m) \quad (34)$$

Equation (19) for regression coefficients is not applicable for logistic regression, because the event probability p_i is not observable. The observable predictand is the event happens ($y_i = 1$) or not ($y_i = 0$). The regression coefficients can be calculated with maximum likelihood method. The likelihood function is a function of probability p_i , fixing the predictand y_i and predictors x_1, \dots, x_m , as follows:

$$\Lambda = \begin{cases} p_i, & y_i = 1 \\ 1 - p_i, & y_i = 0 \end{cases} \quad (35)$$

Substituting Eq. (32) into (35), we can obtain a function of b_1, \dots, b_m :

$$\Lambda(\mathbf{b}) = \prod_{i=1}^n \frac{y_i \exp\left(b_0 + \sum_{j=1}^m b_j x_j\right) + (1 - y_i)}{1 + \exp\left(b_0 + \sum_{j=1}^m b_j x_j\right)} \quad (36)$$

And the log-likelihood function is:

$$\begin{aligned} L(\mathbf{b}) &= \ln[\Lambda(\mathbf{b})] \\ &= \sum_{i=1}^n \left\{ y_i \left(b_0 + \sum_{j=1}^m b_j x_j \right) - \ln \left[1 + \exp \left(b_0 + \sum_{j=1}^m b_j x_j \right) \right] \right\} \end{aligned} \quad (37)$$

The regression coefficients b_1, \dots, b_m can be obtained by maximizing the log-likelihood function.

ANOVA is also applicable for logistic regression. But for evaluating multiple alternative logistic regression models, instead of hypothesis test of F-ratio, the information criteria, such as the Bayesian information criterion (BIC):

$$BIC = -2L(\mathbf{b}) + (m + 1)\ln(n) \quad (38)$$

or the Akaike information criterion (AIC):

$$AIC = -2L(\mathbf{b}) + 2(m + 1) \quad (39)$$

are preferred. Both AIC and BIC consist of a negative log-likelihood term plus a penalty term of the number of predictors, and the preferred regression will be the one giving smaller information criterion. BIC also considers the number of observations and it is preferred for large n problems. For small n problems, BIC always chooses the simpler model instead of the suitable one so that AIC is preferred in this case.

Generally speaking, the information criteria, such as AIC and BIC, can also be applied to other regression models to mitigate the over-fitting problem. Due to the statistical learning theory (Vapnik 1982), a model with more free parameters is almost always better at reducing empirical risk (better at fitting training data), whereas might suffering larger structural risk (worse at fitting test data). The statistical learning theory provides a fundamental support to the traditional Occam's razor: unnecessary assumptions should be shaved away. In order to mitigate the over-fitting problem, we need some methods to balance the empirical risk and structural risk. The information criteria, consisting of a likelihood term representing empirical risk and a penalty term representing the complexity of the regression model, are frequently used to control the over-fitting of not only nonlinear regression but also other kinds of data-driven or even dynamical models.

3.2 Poisson Regression

Residual distribution of nonnegative integers, such as the number of hydrometeorology events like floods, storms, hurricanes, or tornados in a calendar year, is poorly represented by a Gaussian distribution. Therefore, for a predictand which is a nonnegative integer and tends to be a small number (e.g., the number of typhoons landed in specific location within a year), we know that the distribution of the residuals would be usually asymmetric and a regression model that treats it as a continuous value may lead to very large errors in its predictions.

A more appropriate probability model for predicting the number of events is a Poisson distribution. A conventional regression function can be interpreted as a conditional mean of the predictand given the values of the predictors. In Poisson regression, instead of predicting the number of events directly, it chooses the Poisson parameter μ as the predictand in the regression equation.

The Poisson distribution describes random events occurring sequentially. The random variable in Poisson distribution only takes nonnegative integer values. Each Poisson event happens randomly and independently, but the average happening rate over a time period is a constant value. Take the number of typhoons as an example, if the number of typhoons landed is a random variable Y , then the probability of $Y = y$ is

$$P(Y = y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \dots \quad (40)$$

where μ is the only parameter of Poisson distribution specifying the average number of events per unit time. Parameter μ is also called intensity, and a higher μ means the events happen more intensively while smaller μ less intensively.

With the link function $g(\mu) = \ln(\mu)$, we can construct a regression equation about the nonlinear relationship between Poisson intensity and various predictors. With a generalized linear model (GLM), we use the log-link function $g(\mu) = \ln(\mu)$ to define a Poisson regression:

$$\ln(\mu) = b_0 + b_1x_1 + b_2x_2 + \dots + b_mx_m \quad (41)$$

The logarithm link function is not the only choice, but it makes the regression more tractable and ensures the Poisson mean to be nonnegative.

Similar to logistic regression, with the assumption that all n data points are generated independently, the regression coefficients b_1, \dots, b_m can be obtained by maximizing the log-likelihood function.

$$L(\mathbf{b}) = \sum_{i=1}^n \left\{ y_i \left(b_0 + \sum_{j=1}^m b_j x_j \right) - \exp \left(b_0 + \sum_{j=1}^m b_j x_j \right) \right\} \quad (42)$$

where the term involving $y!$ from denominator of Eq. (40) has been omitted, because this term has no influence to the regression parameters. The regression coefficients

can be computed by algorithms such as Newton-Raphson algorithm, the expectation-maximization, or the SCE-UA (Duan et al. 1993). The methods for evaluating the strength of regression, such as ANOVA, AIC, and BIC, are also applicable to Poisson regression.

4 Machine Learning Methods for Regression

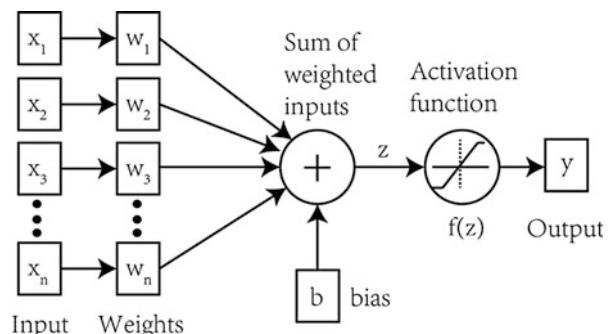
Besides classical linear/nonlinear regression methods, there are a plethora of methods developed in the community of machine learning that can be used as regression methods. In this section, various machine learning methods for regression are introduced, including artificial neural network (ANN), support vector machine (SVM), regression tree and random forests, multivariate adaptive regression splines (MARS), and Gaussian processes regression (GPR). These methods are developed based on different statistical foundations, but all of them are widely used as regression methods in various applications, including hydrometeorological forecasting problems. In this section, a brief introduction about the theoretical foundations and method procedures for each method is provided.

4.1 Artificial Neural Network (ANN)

Artificial neural network (ANN) is a classical machine learning method for classification and regression. ANN can, although imperfectly, mimic the mechanics of the brain of human beings. If properly trained, ANN can effectively extract complex nonlinear relationships from input/output data. Typically, a neural network is constructed by numerous network layers, and each layer has multiple neurons, or perceptrons, which can be mathematically abstracted as a linear weighing function and a nonlinear activation function, as shown in Fig. 1.

Where x_i is the i -th input value, w_i is the weight factor, b is the bias, and z is the weighted sum of the neuron. $f(z)$ is the activation function and its value is the output of the neuron. The neuron is active if $f(z)$ is large enough and inactive if it is

Fig. 1 The structure of a neuron



small enough. As an analog of a biological nerve cell, the input of a neuron is like the stimulus from the previous neuron (presynaptic cell), and the weight factor is like the strength of synapse. The signal from another neuron is received by dendrite, a short-branched extension of a nerve cell, and then transmitted to the cell body of the presynaptic cell. A larger weight factor means a stronger synapse, and the stimulus can be amplified through a strong synapse, and it will be more likely to make the neuron activated. If the total stimulus received by all of the dendrites of a neuron is larger than a threshold, then the neuron becomes activated and the signal is transmitted through the axon, the long threadlike part of a nerve cell along which impulses are conducted from the cell body to other cells. At the end of the axon, there are many synapses and the impulses can be transmitted to the next nerve cell. A neuron can receive impulses from many neurons and can transmit the impulse to many neurons.

The weighted sum of a neuron's input can also be represented as the following equation.

$$z = \sum_{i=1}^n w_i x_i + b \quad (43)$$

There are many kinds of activation functions, as shown in the following equation.

$$f(z) = \begin{cases} \frac{1}{1 + \exp(-z)}, & \text{sigmoid} \\ \tan h(z), & \tan h \\ \begin{cases} 0, & \text{for } z < 0 \\ z, & \text{for } z \geq 0 \end{cases}, & \text{ReLU} \\ \frac{e^{z_i}}{\sum_{j=1}^J e^{z_j}} & \text{for } j = 1, 2, \dots, J, \quad \text{softmax} \\ \max_i z_i, & \text{maxout} \end{cases} \quad (44)$$

The output of the previous layer will be the input of the next layer so that we can construct a multilayer neural network. Typically, we use three layers, with one layer using the raw input data, one hidden layer in the middle, and one output layer. This is usually called three-layer, or single hidden layer, feed-forward network, as shown in Fig. 2. In this network, the impulse propagates forward from the input side to the output side. Both input and output variables can contain multiple variables so that the single hidden layer network can be used as a multivariate multi-objective regression method. Furthermore, we can add multiple hidden layers in the network, as shown in Fig. 3. This kind of network with multiple hidden layers is usually called a deep neural network.

Fig. 2 The structure of three-layer (single hidden layer) neural network

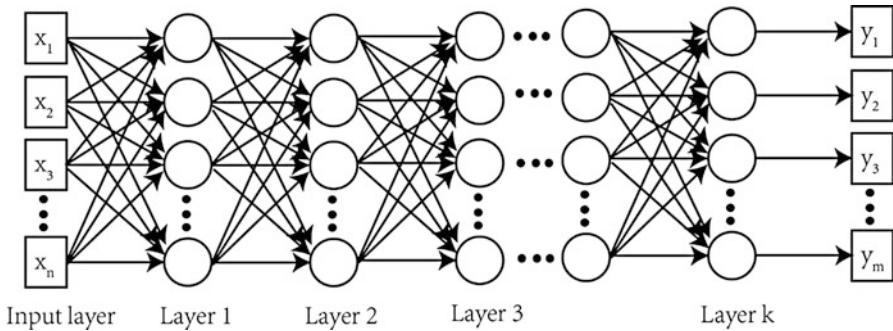
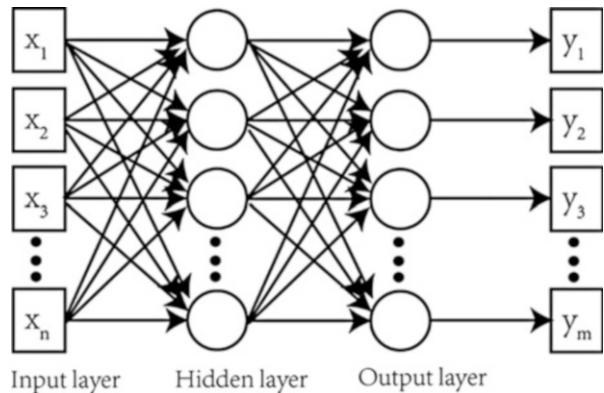


Fig. 3 The structure of deep (multiple hidden layer) neural network

In the literature, single layer network has been demonstrated to be unable to solve nonlinear problems (Minsky and Papert 1969), but multilayer networks have the ability to infinitely approximate any nonlinear function (Cybenko 1989) (the universal approximation theorem). Training an ANN is a large-scale optimization problem to find the best values of weights w_i . Because the neurons between each layer are fully connected, the total number of weights can be a very huge number. Take a single hidden layer network as an example. If there are 10 variables in the input layer, 15 neurons in the hidden layer, and 3 output variables, then the total number of weights is $10 \times 15 \times 3 = 450$. Simultaneously optimizing 450 parameters is a very challenging problem, because there are too many parameters to adjust, and there might be many local optima.

Gradient-based optimization methods are frequently used in training neural networks. Suppose we have a loss function, L , which can measure the difference between the neural network outputs and the true values of output variables, then we can use the gradient of loss function of each weight factor, $\frac{\partial L}{\partial w_i}$, to adjust the weight factor. The weight factor is adjusted using the following equation:

$$\hat{w}_i = w_i - \eta \frac{\partial L}{\partial w_i} \quad (45)$$

where \hat{w}_i means the updated weight factor, while w_i means the weight factor before update, and η is the learning rate. The value of learning rate represents the “speed” of learning. A large learning rate means the value updating has a large stride, while smaller learning rate means the weight factor is only updated a small step along the gradient direction. The learning rate is a tunable parameter in neural network construction. We usually tune the value with experience or adaptively change the value of η in the training process.

The value of gradient $\frac{\partial L}{\partial w_i}$ can be determined with backpropagation method. In each layer the value of the gradient can be automatically derived with backpropagation. The name of backpropagation comes from the backward propagation of errors: the error is calculated at the output end of network and distributed through the network layers toward the input end. Once the gradient has been determined, the weight factors can be updated with many gradient-based optimization methods, such as the Newton downhill algorithm, L-BFGS, Levenberg-Marquardt (LM) (Marquardt 1963) algorithm, etc. Backpropagation is a special case of an older and more general technique called automatic differentiation. Some tools like Tensorflow, Theano, Torch, CNTK, Caffe, among others, can automatically derive the value of gradient of very complex, deep neural networks. With the development of automatic differentiation, training more complex, deep neural networks become possible, and the research and application of deep learning have become more and more popular in the recent years. Some useful information about recent development of deep learning can be found in the review papers (Guo et al. 2016; LeCun et al. 2015; Schmidhuber 2015).

In addition to the multilayer perceptrons, various types of ANNs, such as radial basis function nets, self-organizing maps, Hopfield network, etc., have been developed. A very brief introduction about various types of ANNs can be found in Jain et al. (1996) and also Hastie et al. (2009). A very helpful and textbook-like reference about the recent breakthroughs of neural networks and deep learning is the online book *Neural Networks and Deep Learning* (<http://neuralnetworksanddeeplearning.com/index.html>).

4.2 Support Vector Machine (SVM)

Support vector machine (SVM) is a machine learning method based on statistical learning theory (Vapnik 1982, 2002). It can be applied to both classification and regression problems. The most significant advantage of SVM is that it can automatically avoid the over-fitting problem by balancing the structure risk and empirical risk.

Given the N points training set (\mathbf{x}_i, y_i) , $i = 1, 2, \dots, N$, the linear regression function $f(\mathbf{x})$ can be written as:

$$f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b \quad (46)$$

where \mathbf{w} is the weighting factor or regression coefficients, b is the intercept, and \mathbf{x} is the n-dimensional input vector. Furthermore, nonlinear problems can be transformed to linear problems by applying a mapping function $\phi(\mathbf{x})$, which transforms a low-dimensional nonlinear problem to a high-dimensional linear problem, and the regression function becomes:

$$f(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \quad (47)$$

The kernel function $K(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x})^T \phi(\mathbf{x}')$ is defined as the inner product of the nonlinear mapping function. Various kernel functions can be applied to SVM, such as linear kernel function, polynomial, sigmoid, and radial basis function (RBF), as shown below:

$$\text{Linear : } K(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}' \quad (48a)$$

$$\text{Polynomial : } K(\mathbf{x}, \mathbf{x}') = (\gamma \mathbf{x}^T \mathbf{x}' + r)^d, \gamma > 0 \quad (48b)$$

$$\text{Sigmoid : } K(\mathbf{x}, \mathbf{x}') = \tanh(\gamma \mathbf{x}^T \mathbf{x}' + r) \quad (48c)$$

$$\text{RBF : } K(\mathbf{x}, \mathbf{x}') = \exp(-\gamma |\mathbf{x} - \mathbf{x}'|^2), \gamma > 0 \quad (48d)$$

For regression function $f(\mathbf{x})$, a loss function be defined as:

$$|y - f(\mathbf{x})|_\varepsilon = \begin{cases} 0, & \text{if } |y - f(\mathbf{x})| \leq \varepsilon \\ |y - f(\mathbf{x})| - \varepsilon, & \text{otherwise} \end{cases} \quad (49)$$

The loss function qualifies the “risk,” namely, the amount of penalty for regression error. Residuals less than ε are not penalized, and this penalty free zone is called ε -tube. Training a SVM is to include the training points in the ε -tube as many as possible by minimizing the value of penalty with a relative small ε -tube, and finally the regression function is determined by a small set of training points that support the ε -tube. These supporting points are called support vectors, and the name of SVM comes from that. The optimization problem can be written in the form of quadratic regression:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \xi^*} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i + C \sum_{i=1}^n \xi_i^* \\ \text{subject to} \quad & \mathbf{w}^T \phi(\mathbf{x}_i) + b - y_i \leq \varepsilon + \xi_i \\ & y_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b \leq \varepsilon + \xi_i^* \\ & \xi_i, \xi_i^* \geq 0, i = 1, 2, \dots, n \end{aligned} \quad (50)$$

By transforming it to the dual problem:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi, \xi^*} \quad & \frac{1}{2} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*)^T \mathbf{K} (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) + \varepsilon \sum_{i=1}^n (\alpha_i + \alpha_i^*) + \sum_{i=1}^n y_i (\alpha_i - \alpha_i^*) \\ \text{subject to} \quad & \mathbf{e}^T (\boldsymbol{\alpha} - \boldsymbol{\alpha}^*) = 0 \\ & y_i - \mathbf{w}^T \phi(\mathbf{x}_i) - b \leq \varepsilon + \xi_i^* \\ & 0 \leq \alpha_i, \alpha_i^* \leq C, i = 1, 2, \dots, n \end{aligned} \quad (51)$$

(where \mathbf{K} is a matrix of $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$), we can obtain the predictive function:

$$f(\mathbf{x}) = \sum_{i=1}^n (-\alpha_i + \alpha_i^*) K(\mathbf{x}_i, \mathbf{x}) + b \quad (52)$$

where $(\boldsymbol{\alpha}^* - \boldsymbol{\alpha})$ are the support vectors.

4.3 Regression Tree and Random Forests

The classification and regression tree (CART) (Breiman et al. 1984) is a type of machine learning method that describes the input-output relationship by iteratively splitting the feature space (namely, the input space) into a set of rectangles. The sample points falling in one rectangle is classified into one class (for classification problems) or averaged to a constant value (for regression problems), and the tree is constructed by minimizing the error of misclassification (for classification problems) or fitting error (for regression problems). In this chapter, we focus on the regression tree. The expression of regression tree is usually written as:

$$\hat{f}(\mathbf{x}) = \sum_{m=1}^M c_m I(\mathbf{x} \in R_m) \quad (53)$$

$$I(\mathbf{x} \in R_m) = \begin{cases} 1, & \mathbf{x} \in R_m \\ 0, & \mathbf{x} \notin R_m \end{cases} \quad (54)$$

where $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is the n-dimensional input predictor, $\hat{f}(\mathbf{x})$ is the value of regression tree output, M is the total number of rectangles, m is the index of rectangle, R_m is the corresponding region of number m rectangle, and c_m is the mean value of the sample points falling in R_m . The feature space is split iteratively by adding an additional boundary, and the feature to split as well as the position of the boundary is determined by a greedy algorithm. With the greedy binary partitioning, a regression tree can be constructed very fast even for very large dataset.

Obviously, a very large regression tree might over-fit training data, while a very small tree might lose some important details. So, it is necessary to appropriately control the tree size that maintains the fitting ability while avoiding over-fitting. The

most preferred strategy is first growing a very large tree and then pruning it with a cost-complexity pruning method. Define the cost-complexity criterion as:

$$C_\alpha(T) = \sum_{m=1}^{|T|} N_m Q_m(T) + \alpha |T| \quad (55)$$

where α is the pruning parameter, $|T|$ is the number of nodes of tree T , N_m is the number of points in the rectangle R_m , and $Q_m(T)$ is the fitting error, defined as:

$$\begin{aligned} N_m &= No\{x_i \in R_m\} \\ \hat{y}_m &= \frac{1}{N_m} \sum_{x_i \in R_m} y_i \\ Q_m(T) &= \frac{1}{N_m} \sum_{x_i \in R_m} (y_i - \hat{y}_m)^2 \end{aligned} \quad (56)$$

where $No\{x_i \in R_m\}$ is defined as the number of points x_i falling in the rectangular region R_m . The pruning process is to remove tree nodes one by one by minimizing $C_\alpha(T)$. The appropriate value of α can be found by cross-validation. More details about the regression tree can be found in Breiman et al. (1984). A brief introduction about various types of tree-based methods can be found in Hastie et al. (2009).

However, a single tree has many drawbacks, such as a lack of smoothness, high regression residuals, and a lack of uncertainty information. Random forests (Breiman 2001) is an ensemble of classification and regression trees (CART) that can mitigate these disadvantages. A random forest contains many trees constructed from randomly selected predictors of random sample points, and the value of predictand is computed by averaging the outputs of all the trees. A random forest has two parameters: the total number of trees t and the number of selected features \hat{m} . A random forest can be trained with the following steps:

1. Randomly select sample points one by one with replacement from the sample pool (\mathbf{x}_i, y_i) , $i = 1, 2, \dots, N$. Repeat N times to construct a resample set of N points. This data set is called a bootstrap replication, and we can construct a regression tree from each bootstrap replication. If the total number of trees is t , we then need t bootstrap replications. This step is also called bootstrap aggregating (bagging).
2. Construct the regression tree with each bootstrap replication. First randomly select \hat{m} features from the total M features, try each feature and select the best fitting one to split, then move on to next iteration, selecting \hat{m} features, and split again. In each iteration, the \hat{m} features are selected randomly and independently.
3. Averaging the output of t trees to get the value of predictand.

$$\hat{f}_{rf}(\mathbf{x}) = \frac{1}{t} \sum_{j=1}^t \hat{f}_j(\mathbf{x}) \quad (57)$$

Random forests are very popular in very high-dimensional problems, which usually have hundreds or thousands or even more input variables (features), but each feature can only contribute a little information. A single tree usually has very poor predictive skill because the feature space is very hollow. However, by combining the power of many trees, a random forest can sufficiently explore the hollowness of high-dimensional space with multiple individual regression trees to extract the little bit of information provided by each feature. Random forests, as well as other ensemble tree-based approaches, has become very popular for big-data problems because of their outstanding effectiveness and efficiency in dealing with the very high-dimensional feature space.

4.4 Multivariate Adaptive Regression Splines (MARS)

The multivariate adaptive regression splines (MARS) model is a hybrid model of classical linear regression and regression tree (Friedman 1991). The input space is first split into several subspaces, and the linear regression model is applied to each partition. The MARS model can be expressed as an expansion in product spline basis functions:

$$y = f(\mathbf{x}) = a_0 + \sum_{m=1}^M a_m \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})]_+ \quad (58)$$

where y is the predictand, $\mathbf{x} = (x_1, x_2, \dots, x_n)$ is the n-dimensional predictor, a_0 is the intercept, a_m is the regression coefficient (or weight) of each basis function, m is the index of basis function, and M is the total number of basis functions. For each basis function:

$$B_m(\mathbf{x}) = \prod_{k=1}^{K_m} [s_{km}(x_{v(k,m)} - t_{km})]_+ \quad (59)$$

k is the index of knot (the position of split boundary); K_m is the total number of knots of the m -th basis function; s_{km} equals to 1 or -1 , indicating the right or left direction of associated step function ($x_{v(k,m)} - t_{km}$); $v(k, m)$ is the input vector \mathbf{x} 's index of the k -th knot, m -th basis function; and t_{km} is the knot location of the k -th knot, m -th basis function.

MARS is constructed in a build-prune two-stage strategy. In the build stage, namely, the forward pass, an over-fitting model with all input variable is constructed, and in the prune stage, namely, the backward pass, the insensitive input variables are removed one at a time. With the two-stage strategy, MARS can be constructed

appropriately to extract the information contained in the data and also shield the influence of noise. To obtain the best number of terms λ for a MARS model, we can minimize the generalized cross-validation criterion:

$$GCV(\lambda) = \frac{\sum_{i=1}^N (y_i - \hat{f}_\lambda(x_i))^2}{\left(1 - \frac{M(\lambda)}{N}\right)^2} \quad (60)$$

where the numerator represents the fitting error and the denominator contains a complexity penalty factor $M(\lambda) = r + cK$ where r is the number of selected basis function and K is the number of selected nodes and the constant $c = 3$. More technical details about MARS can be found in the paper (Friedman 1991) and the book (Hastie et al. 2009).

4.5 Gaussian Processes Regression

Gaussian process regression (GPR) (Rasmussen and Williams 2006) is a kind of machine learning method based on Gaussian stochastic processes. It is also known as *kriging* in the community of geostatistics and meteorology. Given a random variable \mathbf{x} , if it changes with time t (or other factors), then the changing process $\mathbf{x}(t)$ is called a stochastic process. In GPR, the predictand y is supposed to be a Gaussian stochastic process changing with the predictors $\mathbf{x} = (1, x_1, x_2, \dots, x_m)^T$. The GPR can not only provide the expected mean value of predictand but also estimate its confidence interval. GPR is suitable for multivariate nonlinear problems and has been widely used in constructing surrogate models. The main drawback of GPR is that it is not suitable for big-data problems because its calculation is very time consuming (if it involves more than thousands of samples), and it is over-smoothed for high-dimensional problems (i.e., hundreds of factors).

There are two typical views of GPR, the weight-space view and the function-space view. Let's begin from the simpler weight-space view. Take the linear regression problem for an example:

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{w}, \quad y = f(\mathbf{x}) + \varepsilon \quad (61)$$

where $\mathbf{x} = (1, x_1, x_2, \dots, x_m)^T$ is the input vector, \mathbf{w} is the vector of weights (regression coefficients), $f(\mathbf{x})$ is the regression equation, and ε is i.i.d Gaussian residuals ($\varepsilon \sim N(0, \sigma_n^2)$).

Instead of setting the derivatives to zero, the goal of GPR is to derive the posterior distribution of weighing vector \mathbf{w} directly from the i.i.d Gaussian assumption and Bayes' rule. The likelihood function, namely, the conditional distribution of predictand y given predictors \mathbf{X} and weights vector \mathbf{w} , can be written as follows:

$$\begin{aligned}
p(\mathbf{y}|\mathbf{X}, \mathbf{w}) &= \prod_{i=1}^n p(y_i | \mathbf{x}_i, \mathbf{w}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left(-\frac{(y_i - \mathbf{x}_i^T \mathbf{w})^2}{2\sigma_n^2}\right) \\
&= \frac{1}{(2\pi\sigma_n^2)^{\frac{n}{2}}} \exp\left(-\frac{1}{2\sigma_n^2} |\mathbf{y} - \mathbf{X}^T \mathbf{w}|^2\right) \\
&= N(\mathbf{X}^T \mathbf{w}, \sigma_n^2 I)
\end{aligned} \tag{62}$$

where \mathbf{X} is the observed predictor matrix and \mathbf{y} is the observed predictand vector, defined as follows:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1m} \\ 1 & x_{21} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nm} \end{bmatrix}^T, \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \tag{63}$$

Note that the definition of \mathbf{X} is different from Eq. (20). By assuming the prior distribution of $\mathbf{w} \sim N(\mathbf{0}, \Sigma_p)$, the posterior distribution of \mathbf{w} can be derived from Bayes' rule:

$$\begin{aligned}
p(\mathbf{w}|\mathbf{X}, \mathbf{y}) &= \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|\mathbf{X})} \\
&\propto \exp\left(-\frac{1}{2\sigma_n^2} (\mathbf{y} - \mathbf{X}^T \mathbf{w})^T (\mathbf{y} - \mathbf{X}^T \mathbf{w})\right) \exp\left(-\frac{1}{2} \mathbf{w}^T \Sigma_p^{-1} \mathbf{w}\right) \\
&\propto \exp\left(-\frac{1}{2} (\mathbf{w} - \bar{\mathbf{w}})^T \mathbf{A} (\mathbf{w} - \bar{\mathbf{w}})\right)
\end{aligned} \tag{64}$$

where $\mathbf{A} = \sigma_n^{-2} \mathbf{X} \mathbf{X}^T + \Sigma_p^{-1}$ and $\bar{\mathbf{w}} = \sigma_n^{-2} \mathbf{A}^{-1} \mathbf{X} \mathbf{y}$. We recognize the form of posterior distribution as Gaussian with mean $\bar{\mathbf{w}}$ and covariance matrix \mathbf{A}^{-1} .

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) \sim N(\bar{\mathbf{w}}, \mathbf{A}^{-1}) \tag{65}$$

Furthermore, the predictive distribution of y_* given predictor \mathbf{x}_* can be derived from the distribution of \mathbf{w} :

$$p(y_* | \mathbf{x}_*, \mathbf{X}, \mathbf{y}) \sim N(\sigma_n^{-2} \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{X} \mathbf{y}, \mathbf{x}_*^T \mathbf{A}^{-1} \mathbf{x}_*) \tag{66}$$

In the function-space view, a GPR can be viewed as an ensemble of possible regression equations $f(\mathbf{x})$. The statistical characteristics of the trajectories of these functions can be completely specified by their mean and covariance function:

$$\begin{cases} m(\mathbf{x}) = E[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') = E[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))] \end{cases} \tag{67}$$

where $\mathbf{x} = (x_1, x_2, \dots, x_m)^T$ is the m-dimensional input vector of predictor, $m(\mathbf{x})$ is the mean function, and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function between two input vectors \mathbf{x} and \mathbf{x}' . Note that the definition of input vector here is different from that of weight-space view because in the function-space view, we have mean function instead of adding an additional term in the weight vector \mathbf{w} . If the predictor point \mathbf{x} is very far from other points \mathbf{x}' , the mean value of predictand $E[y]$ will be close to $m(\mathbf{x})$, and the variance will be very large. The mean function $m(\mathbf{x})$, and the covariance function $k(\mathbf{x}, \mathbf{x}')$ provides all of the information about regression, so for short the regression can be written as $f(\mathbf{x}) = GP(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$.

Suppose \mathbf{X} is the training input matrix (namely, the observed predictors), \mathbf{X}_* is the testing input matrix, \mathbf{y} is the training outputs (the observed predictand) and \mathbf{y}_* is the testing outputs, and $K(\mathbf{X}, \mathbf{X})$, $K(\mathbf{X}, \mathbf{X}_*)$, and $K(\mathbf{X}_*, \mathbf{X}_*)$ are covariance matrices of training and testing inputs. Each element of $K(., .)$ is equal to the covariance function $k(\mathbf{x}, \mathbf{x}')$ of input pairs. Assuming the mean function $m(\mathbf{x}) = 0$, the joint distribution of training and testing outputs is a joint Gaussian distribution like this:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim N\left(\mathbf{0}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix}\right) \quad (68)$$

Consequently the expression of mean and variance of the predictand \mathbf{y}_* can be derived from Bayes' rule:

$$\begin{cases} E[\mathbf{y}_*] = K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} \mathbf{y} \\ \text{cov}[\mathbf{y}_*] = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} K(\mathbf{X}, \mathbf{X}_*) \end{cases} \quad (69)$$

The deterministic mean function $m(\mathbf{x})$ can be easily introduced by applying a constant offset to the observed predictand. The mean of predictand becomes:

$$E[\mathbf{y}_*] = m(\mathbf{X}_*) + K(\mathbf{X}_*, \mathbf{X})[K(\mathbf{X}, \mathbf{X}) + \sigma_n^2 I]^{-1} (\mathbf{y} - m(\mathbf{X})) \quad (70)$$

and the predictand variance remains unchanged.

The covariance function, which encodes the prior assumption of the input-output relationship we wish to learn, is the most crucial ingredient of GPR. The covariance function is *stationary* if it only depends on the relative position $\mathbf{x} - \mathbf{x}'$ and it is *isotropic* if it is determined by the Euclidian distance between them, i.e., $k(\mathbf{x}, \mathbf{x}') = k(r)$, $r = |\mathbf{x} - \mathbf{x}'|$. As $k(r)$ is totally determined by r , this kind of covariance function is also called *radial basis function* (RBF). A simple example of stationary isotropic covariance function is the squared exponential (SE) cov-function:

$$k(r) = \exp\left(-\frac{r^2}{2l^2}\right) \quad (71)$$

In which l is characteristic length, a hyper-parameter of the cov-function. A smaller l leads to a rougher response surface. The SE cov-function is infinitely differentiable so that it is very smooth, sometimes over-smoothed for some practical problems.

A more general class of covariance function is Martérn covariance function:

$$k(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}r}{l} \right)^\nu K_\nu \left(\frac{\sqrt{2\nu}r}{l} \right) \quad (72)$$

where $K_\nu(\cdot)$ is a modified Bessel function and ν and l positive hyper parameters. ν is the shape factor, and l is the characteristic length. The Martérn covariance function is also isotropic, and the parameter ν controls the shape of Gaussian processes. The Martérn cov-function becomes smoother with larger ν , and if $\nu \rightarrow \infty$ it becomes the SE cov-function. To facilitate the computation, ν is usually set to half-integer $\nu = p + 1/2$ so that the Martérn cov-function becomes a product of a polynomial and an exponential. The most commonly used cases are $\nu = 3/2$ and $\nu = 5/2$:

$$k_{\nu=3/2}(r) = \left(1 + \frac{\sqrt{3}r}{l} \right) \exp \left(-\frac{\sqrt{3}r}{l} \right) \quad (73)$$

$$k_{\nu=5/2}(r) = \left(1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2} \right) \exp \left(-\frac{\sqrt{5}r}{l} \right) \quad (74)$$

Other kinds of covariance functions are also applicable, such as rational quadratic covariance function which is also isotropic:

$$k(r) = \left(1 + \frac{r^2}{2al^2} \right)^{-\alpha} \quad (75)$$

Polynomial covariance function, which is nonstationary and non-isotropic:

$$k(r) = (\mathbf{x} \cdot \mathbf{x}' + \sigma_0^2)^p \quad (76)$$

Neural network covariance function, which is also non-stationary and non-isotropic:

$$k(r) = \sin^{-1} \left(\frac{2\tilde{\mathbf{x}}^T \Sigma \tilde{\mathbf{x}}'}{\sqrt{(1 + 2\tilde{\mathbf{x}}^T \Sigma \tilde{\mathbf{x}})(1 + 2\tilde{\mathbf{x}}'^T \Sigma \tilde{\mathbf{x}}')}} \right) \quad (77)$$

Neural network covariance function is derived from single hidden layer neural network. Users can also define their own covariance function from their prior knowledge about the input-output relationship.

Given the type of covariance function, the behavior of Gaussian processes regression is controlled by the value of hyper-parameters. An appropriate estimation of hyper-parameters can be obtained by maximizing the marginal log-likelihood function:

$$\log[p(\mathbf{y}|X)] = -\frac{1}{2}\mathbf{y}^T(K + \sigma_n^2 I)^{-1}\mathbf{y} - \frac{1}{2}\log|K + \sigma_n^2 I| - \frac{n}{2}\log 2\pi \quad (78)$$

where $K = K(\mathbf{X}, \mathbf{X})$. Any effective global optimization methods, such as the SCE-UA algorithm (Duan et al. 1993), can be used to find appropriate hyper-parameters. For more technical and theoretical detail of various types of Gaussian processes models and how to select the best model and hyper-parameters, please refer to the book written by Rasmussen and Williams (2006).

5 Summary

This chapter provides a brief introduction to various types of regression models that has been, or can be potentially used in hydrometeorological ensemble forecasting, including not only the classical linear and nonlinear regressions but also Gaussian processes regression (kriging), regression tree, MARS, SVM, ANN, etc. These regression methods account for the relationships between the predictand and predictors with various types of assumptions about the parametric models and solve the model with appropriate numerical methods.

There are many interesting connections between these regression methods. For instance, the MARS method is a combination of regression tree and linear regression. Similarly, the classical linear regression problem can also be interpreted under the framework of GPR, SVM, and even ANN, as the most simplified cases having been shown in previous introductions. Both GPR and SVM solve nonlinear problems with the help of the “kernel trick,” which uses a nonlinear mapping function $\phi(\mathbf{x})$ to transform the lower dimensional and nonlinear predictand \mathbf{x} to higher dimensional linear space and solve the transformed linear problem instead of the original nonlinear problem. The MARS model can be used as a sensitivity analysis method with the help of ANOVA. The GPR can simulate the behavior of infinite ensemble of multilayer perceptrons with a specific covariance function. The reason for the over-fitting problem of various regression models can be explained with the statistical learning theory and can be dealt with cross-validation, information criteria, or other specific methods. The interesting connections indicate that although the regression methods are proposed under different assumptions, they are sharing some common statistical fundamentals. Consequently, a technology suitable for one model can also be applicable to many other regression models.

References

- L. Breiman, Random forests. *Mach. Learn.* **45**(1), 5–32 (2001)
- L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, *Classification and Regression Trees* (Chapman and Hall/CRC, Boca Raton, 1984)
- G. Cybenko, Approximation by superpositions of a sigmoidal function. *Math. Control Signals Syst.* **2**(4), 303–314 (1989). <https://doi.org/10.1007/BF02551274>

- Q.Y. Duan, V.K. Gupta, S. Sorooshian, Shuffled complex evolution approach for effective and efficient global minimization. *J. Optim. Theory Appl.* **76**(3), 501–521 (1993). <https://doi.org/10.1007/bf00939380>
- J.H. Friedman, Multivariate adaptive regression splines. *Ann. Stat.* **19**(1), 1–14 (1991)
- Y. Guo, Y. Liu, A. Oerlemans, S. Lao, S. Wu, M.S. Lew, Deep learning for visual understanding: A review. *Neurocomputing* **187**, 27–48 (2016). <https://doi.org/10.1016/j.neucom.2015.09.116>
- T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, 2nd edn. (Springer, New York, 2009)
- A.E. Hoerl, R.W. Kennard, Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**(1), 55–67 (1970). <https://doi.org/10.1080/00401706.1970.10488634>
- A.K. Jain, M. Jianchang, K.M. Mohiuddin, Artificial neural networks: A tutorial. *Computer* **29**(3), 31–44 (1996). <https://doi.org/10.1109/2.485891>
- R. Koenker, G. Bassett, Regression quantiles. *Econometrica* **46**(1), 33–50 (1978). <https://doi.org/10.2307/1913643>
- Y. LeCun, Y. Bengio, G. Hinton, Deep learning. *Nature* **521**(7553), 436 (2015). <https://doi.org/10.1038/nature14539>
- P. López López, J.S. Verkade, A.H. Weerts, D.P. Solomatine, Alternative configurations of quantile regression for estimating predictive uncertainty in water level forecasts for the upper Severn River: A comparison. *Hydrol. Earth Syst. Sci.* **18**(9), 3411–3428 (2014). <https://doi.org/10.5194/hess-18-3411-2014>
- D.W. Marquardt, An algorithm for least-squares estimation of nonlinear parameters. *J. Soc. Ind. Appl. Math.* **11**(2), 431–441 (1963). <https://doi.org/10.2307/2098941>
- M. Minsky, S.A. Papert, *Perceptrons: An Introduction to Computational Geometry* (MIT Press, Cambridge, MA, 1969)
- C.E. Rasmussen, C.K.I. Williams, *Gaussian Processes for Machine Learning* (MIT Press, Cambridge, MA, 2006)
- J. Schmidhuber, Deep learning in neural networks: An overview. *Neural Netw.* **61**, 85–117 (2015). <https://doi.org/10.1016/j.neunet.2014.09.003>
- V.N. Vapnik, *Estimation of Dependencies Based on Empirical Data* (Springer, New York, 1982)
- V.N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd edn. (Springer, New York, 2002)

Index

A

- Acoustic Doppler current profiling (ADCP), 750
Acoustic Doppler velocimetry (ADV), 750
Adaptive metropolis, 596–598
Adaptive neuro-fuzzy inference system (ANFIS), 372
Adjoint model, 686
Advanced Microwave Sounding Radiometer 2 (AMSR2), 706
Advanced SCATterometer (ASCAT), 707, 734, 736, 739
African Flood Forecasting System (AFFS), 1199
Aggregated forecast performance, 1169–1170, 1174
Aggregational approaches, 500–504
Akaike information criteria (AIC), 1481
Aleatoric uncertainties, 241
Analog(s)
 description, 799
 ensemble, 137
 general method, 799–801
 streamflow data, 801
Analysis, 704, 705, 713, 714, 716, 725, 729, 732, 734
Analysis of variance (ANOVA), 646–648, 1488–1490
ANMAX model, 360
Ant colony optimization (ACO), 538–540
Antecedent precipitation index (API) models
 applications, 346
 calculation, 344–345
 graphical, 345–346
Applications of hydrologic ensemble forecasts,
 see Ensemble forecasts, NYC water supply system
A priori estimation, 485
AquaLog/Hydrog LAEF, 1206
Arakawa-C staggered grid method, 77
Artificial neural network (ANN), 1455, 1501
 applications, 364
 network training, 363–364
 structure, 361–362
ASCAT, *see* Advanced SCATterometer (ASCAT)
Ashokan Release Channel (ARC), 1355
Ashokan Reservoir, 1331, 1334, 1335, 1355, 1360
Asynchronous EnKF (AEnKF)
 asynchronous approaches, 754
 characteristics of, 754
 vs. EnKF, 768–772
 vs. PF, 768–772
Atmospheric boundary layer (ABL), 471
Atmospheric model, 16
Atmospheric motion, 45, 48
Atmospheric predictability, 16–18
Australian Bureau of Meteorology (BMRC), 158
Autocorrelated error model, 507
Autocorrelation function, 580
Automatic calibration, 491
Automatic parameter optimization, 424–425
Autoregressive conditional heteroscedasticity (ARCH) model, 360
Autoregressive integrated moving average (ARIMA) model, 357, 1453
Autoregressive model, 355, 1453
Autoregressive moving average (ARMA) models, 356–357

B

- Backpropagation method, 1501
Bayes, 1280, 1285
Bayesian analysis, 605
Bayesian forecasting system (BFS), 299–301

- Bayesian hierarchical modeling (BHM), 837–838
- Bayesian inference, 495, 497, 571–573, 1451–1452
- generalized likelihood function, 583
 - likelihood function, 575–583
 - posterior distribution, 588–590
 - prior distribution, 573–575
- Bayesian joint probability model (BJP), 834–835, 1163–1165
- Bayesian model averaging (BMA), 290–293, 812, 835, 1077, 1480
- Bayesian total error analysis (BATEA), 504–505
- Bayes' law, 1467, 1472, 1478, 1481, 1483
- Bayes' theorem, 627–628, 1434
- Behavioral simulation space, 620
- Bias correction approaches, 827
- Black-box hydrological models
- antecedent precipitation index models, 343–347
 - artificial neural network models, 361–365
 - frequency analysis models, 374–381
 - fuzzy logic models, 365–374
 - regression models, 347–353
 - time series models, 353–361
- Black box model, 327, 328, 334
- BMA, *see* Bayesian model averaging (BMA)
- Boundary perturbation, 116
- Boussinesq equation, 420
- Box-Cox transformation, 292
- Box-Jenkins models, 356
- Brazil, 1309, 1325, 1326
- Bred growing mode (BGM), 105–106
- Brier score (BS), 302, 1479
- Brier skill score (BSS), 1396
- Butterfly effect, 16
- C**
- Calibrate meteorological field, 247
- Calibration, 327, 328, 333–334, 708, 722
- Canal, 1374, 1375
- Canonical correlation analysis (CCA), 213–214
- Canopy
- evaporation, 418, 452, 470
 - interception, 418
 - transpiration, 418
 - water content, 449
- Cartesian coordinate system, 71
- Catchment, 270, 315, 318, 320, 321, 323, 324, 326, 327, 329, 332, 334
- model, 1051–1052
- Catskill System, 1332–1336
- Centro de Previsao de Tempo e Estudos Climaticos (CPTEC), 160, 172
- CFS precipitation forecast skill, variation of, 23, 24
- CFSv2, *see* Climate forecast system version 2 (CFSv2)
- Chaotic system
- and atmospheric predictability, 48
 - described, 41
 - implications of, 41
 - perturbations, 46
 - phase space, 43
- Chinese Meteorological Administration (CMA), 159, 171
- Civil protection, 1141, 1143, 1145, 1149, 1152
- Classification and regression tree (CART), 1504
- Climate change, 321, 323, 324, 333, 336
- Climate forecast system version 2 (CFSv2), 1281, 1282, 1284, 1285
- Climate model, 1280, 1283, 1285
- Climate variability
- canonical correlation analysis, 213
 - constructed analogue method, 214–217
 - coupled general circulation models, 219–220
 - dynamical models, 217–218
 - ENSO (*see* El Niño-Southern Oscillation (ENSO))
 - linear inverse modeling, 217
 - local forcing and feedbacks, 207
 - low-hierarchy climate models, 218–219
 - patterns and indices, 201
 - and prediction, 233
 - sources of predictability, 210–211
 - teleconnections patterns, 208–209
 - two-way interaction models, 219
- Climatological ESP method, 15, 18
- Cognitive biases, 1138, 1155
- Cold season processes
- fractional snow cover, 459–460
 - frozen soil thermodynamics and hydraulics, 462–463
 - snow albedo, 460–461
 - snow melt, 461–462
 - snowpack evolution, 459
 - snow sublimation, 460–461
 - soil heat flux, 462
- Color coding, 595
- Community Hydrologic Prediction System (CHPS), 1192, 1350

- Conceptual hydrological models, 390
 evapotranspiration, 398
 infiltration, 393
 precipitation, 391
 river routing, 402–405
 runoff generation, 399
 Sacramento model, 409
 soil moisture storage, 396–398
 tank model, 406–407
 Xinanjiang model, 407–409
- Conceptual model, 318, 327, 331, 334
- Conceptual rainfall-runoff (CRR) models, 541
- Conditional forecast, 785
- Conditional nonlinear optimal perturbation (CNOP), 110–111
- Conditional probability density function, 1436
- Conditional probability method, 1422
- Conditional Seasonal Storage Objective (CSSO), 1335, 1355
- Conjugate families, 1472
- Conjugate priors, 574
- Connecting ensemble members, 268–269
- Conservation releases, 1346, 1362
- Consolidation methods, 228–230
- Consortium for Small-Scale Modeling (COSMO) ensemble, 290
- Constructed analogue method, 214
- Continental flood, 1269, 1274
- Continuous ranked probability score (CRPS), 176, 181, 902–904
- Continuous ranked probability skill score (CRPSS), 182, 186
- Convective acceleration, 1388
- Convergence monitoring, 605–607
- Copula
 advantage of, 1448
 applications of, 1448
 empirical copula, 1449
 vine copula, 1450
- Cost-loss decision model, 913–915
- Cost structure model, 1379, 1380
- Coupled General Circulation Models, 16
- Coupled ocean–atmosphere–land systems, 60
- Courant-Friedrichs-Lowy (CFL) criterion, 69
- Covariance, 786, 789
- Croton System, 1332
- Crowdsourced measurements, 750
- Cumulative distribution function (CDF), 1432, 1475
- Curse of dimensionality, 1475
- D**
- Dams, 1290, 1293, 1294, 1296, 1299, 1303
- Danjiangkou reservoir, 1420
- Data, 316, 322, 324, 326, 328, 330, 332, 335
 assimilation, 626, 1297, 1298, 1301–1302, 1481–1483
 visualization, 1343, 1347
- Data assimilation (DA)
 definition, 676
 deterministic *vs.* ensemble methods, 679
 linear *vs.* non-linear methods, 679
 model structures, 694–695
 smoothers *vs.* filters, 678
 See also Streamflow data assimilation
- Data assimilation, NWP, 88, 90–91
 3D-VAR, 91, 93
 4D-VAR, 91–93
 ensemble Kalman filter, 93, 94
 least squares theory, 88–90
- Decision-making, 1025, 1026, 1133, 1135, 1141, 1154, 1155, 1289–1302
- Decision rule, 1467
- Decision support, 1347
- Decompositional approaches, 504–505
- Delaware System, 1336
- Delta test (DT), 659–660
- Department of Environmental Protection (DEP)
 conservation releases, 1362
 probabilistic analysis and decision making, 1353
 spill mitigation, 1355
- Design of experiment, 640, 643, 646
- Deterministic forecasts, 5, 851, 857, 860
- Deterministic method, 679
- Deterministic multi-objective search methods, 544–547
- Development of the European Multi-model Ensemble system for seasonal-to-interannual prediction (DEMETER), 289
- D8 flow direction method, 417
- DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm, 293, 602
- Differential evolution Markov chain (DE-MC) algorithm, 599
- Digital elevation model (DEM), 1061, 1390, 1393, 1395, 1396
 database, 416
 gridded basin division, 416
 terrain property data, 416
- 3-Dimensional variational (3D-VAR) data assimilation method, 91, 93
- Direct research methods, 529–531

- Direct soil evaporation, 453
 Direct streamflow measurements, 750
 Disaggregation model, 360
 Disaster response, 1119, 1124
 Displacement methods, 935
 - feature-based methods, 938
 - field deformation approaches, 937
 Distributed hydrologic model (DHM), 414–415
 - AEnKF, 754–756
 - 4DVAR, 757–760
 - particle filter, 759–763*See also* Physically based distributed hydrological models (PBDHMs)
 Distributed models, 328–330, 334
 Distributions-oriented approach, 853, 854, 888
 Downhill simplex method, 529–531
 Downscaling, 1280, 1285
 Drought, 1280, 1283, 1285
 Dual-resolution hybrid ensemble, 135
 Dynamical climate prediction models, 217
 Dynamical hydrology models, 823–825
 Dynamical systems
 - chaotic, 45
 - complex, 42
 - concepts of, 38
 - definition, 39
 - discrete, 39
 - historical perspective, 37–40
 - theory, 46
- E**
 Early flood warning, 1264
 Early warning, 1115–1116
 Earth system models, 226
 Economic value, 1300, 1302
 - of forecast, 979
 EFAS, *see* European Flood Alert System (EFAS)
 EKF, *see* Extended Kalman filter (EKF)
 Elbe, floods, *see* Probabilistic forecasting, flood inundation
 El Niño-Southern Oscillation (ENSO), 198
 - data assimilation schemes, 223
 - interannual variability, 208
 - phenomenon, 201–207
 - plume forecasts, 228
 - tropical ocean-global atmosphere program, 211
 Empirical copula, 1449
 Empirical cumulative distribution function, 1443
 Empirical orthogonal functions (EOFs), 201
 Empirical orthogonal teleconnection, 208
 Energy, 1290, 1292, 1294, 1297, 1300, 1303
 Energy score, 989, 990, 997, 1001, 1002, 1010
 Ensemble-based graphics, 1095
 Ensemble dressing method, 809, 1078
 Ensemble forecast(s), 101–103, 130,
 - 1119–1122, 1227, 1231, 1256, 1456–1457
 - description, 851–853
 - forecast quality, 864–885
 Ensemble forecasts, NYC water supply system conservation releases, 1362
 EPP, 1351
 physically based methods, 1348
 probabilistic analysis and decision making, 1355
 spill mitigation, 1360
 statistical forecast methods, 1349
 Ensemble hydrometeorological forecasting system, 18–28
 Ensemble inflow forecasts, 1310, 1312, 1314
 Ensemble Kalman filter (EnKF), 8, 114–115,
 - 299, 686–687, 754, 1482
 - and MLEF, 770–773
 - vs. PF, 772
 Ensemble method, 679
 Ensemble model output statistics (EMOS) method, 290, 293–294
 Ensemble of data assimilations (EDA), 187
 Ensemble post processor (EPP), 1349, 1351
 Ensemble prediction, 1264, 1268, 1270, 1272, 1281
 Ensemble prediction systems (EPSs), 286, 1041, 1043
 - convective-scale, 49
 - forecast, 287
 - global, 49, 51
 - operational, 49, 51, 55
 - regional, 49, 55
 - role, 36
 - storm-scale, 49
 - types of, 49
 Ensemble size, 135, 139–141
 Ensemble square root filter, 688
 Ensemble streamflow prediction (ESP), 1296, 1300, 1301, 1456
 - bias correction approaches, 827
 - forecast calibration, 827–829
 - historical precipitation series, 1421
 - rainfall runoff ESP, 1419, 1420
 - seasonal runoff forecasts, 830
 - short range flood forecasting, 1415, 1417
 - trace weighting, 829–830

- Ensemble transform Kalman filter (ETKF), 111–114, 168
Ensemble transform with rescaling (ETR), 106
Environment Agency, 1145, 1146, 1148
EPIC, *see* European Precipitation Index based on simulated Climatology (EPIC)
Epistemic uncertainty, 241, 749
Error model equations, 497
Estimation
 of score, 990–993
 uncertainty, 784, 788
European Centre for Medium-Range Weather Forecast (ECMWF), 51, 68, 70, 79, 82, 83, 86, 87, 94, 160, 173, 180, 186, 188, 190, 288
European Flood Alert System (EFAS), 290, 1142, 1148–1152, 1154
European Flood Awareness System (EFAS), 1192, 1386, 1388, 1392, 1394, 1395
European Precipitation Index based on simulated Climatology (EPIC), 1069, 1226, 1228, 1230, 1234, 1252
European Runoff Index based on Climatology, 1069
Evapotranspiration (ET), 314, 316, 321, 323, 324, 328, 331, 336, 398–399
Evolutionary PF-MCMC (EPFM), 692
Existing post-processing methodologies, 242
Extended Fourier amplitude sensitivity test, 650
Extended Kalman filter (EKF), 683–684
Extended Streamflow Prediction (ESP), 14, 15
Extreme forecast index (EFI), 227, 943–944
Extreme precipitation events, 1231
Extreme rainfall alert (ERA), 1067
Extreme-value verification, 942
- F**
- Filtering, 677, 678
Filter methods
 neighborhood methods, 935, 936
 scale-separation methods, 935, 936
Finite difference equations (FDEs), 73
First-order sensitivity index, 649
Fixed forecast model, 227
Flash flood early warning, 1233, 1234, 1238, 1239, 1245, 1254
Flexible Flow Management Program (FFMP), 1343, 1363, 1368
Flood, 1038, 1040, 1042, 1374, 1376, 1378, 1379
 control, 1415
 forecasting, 245
 model, 486
Flood Early Warning System for the Po River (FEWSPo), 1193
Flow accumulation, 417
Forecast accuracy, 1166–1169
 CRPS, RMSE and RMSEP skill scores, 1172–1173
 forecast precision, 1173
 tercile hit rates, for low and high flows, 1173
Forecast centres, 1192
Forecast communication, 1039, 1041
Forecasted available water (FAW), 1364, 1367
Forecast evaluation, 895
Forecast horizons, 1210
Forecasting, 265–266
Forecast precision, 1173
Forecast products and services, 27–28
Forecast quality, 1292, 1300–1302
Forecast reliability, 1166, 1170
Forecast users, 6
Forecast verification, 850, 1014, 1016, 1029, 1030
 accuracy, 856
 aspects, 925–928
 association, 856
 bias, 855–856
 calibration-refinement decomposition, 858–860
 correlation, 944–946
 discrimination, 857
 double-penalty and small-scale errors, 934–935
 ensemble forecasts, 851–853
 ensemble probability distribution forecast, 877–885
 extreme events, 941–944
 high-resolution gridded forecasts, 935–938
 (*see also* Hydrological forecast)
 likelihood-base rate decomposition, 860–862
 mathematical formulation, 853–855
 point forecasts, 928–931
 practical considerations, 885–889
 pre-processing gridded observations, 933–934
 probability forecast for discrete event, 869–877
 reliability, 856
 resolution, 856
 sharpness, 857
 single-valued forecast, 865–869
 skill score, 856, 862–864
 sources of gridded observations, 931–933

- Forecast verification (*cont.*)
 spatial techniques, 938–941
 type-II conditional bias, 857
 uncertainty, 857
- Four dimensional variational data assimilation
 (4D-VAR), 91, 94, 686
 with DHM, 757–760
- Fourier amplitude sensitivity test (FAST),
 648–650
- Fractional factorial (FF) screening, 640–643
- Fractional snow cover, 459
- France, 1401, 1402, 1405, 1407, 1409, 1410
- Frequency analysis models
 data sampling methods, 378
 frequency factor method, 375–378
 graphical method, 374–375
 outliers and zeros, 378
 regionalization, 379–380
 uses, 380
- Fuzzy logic models
 adaptive neuro-fuzzy inference system
 (ANFIS), 372
 fuzzy rules, 367
 fuzzy sets, 366
 membership function, 366
 operations with fuzzy sets, 367–369
 types, 370
- G**
- Gamma distribution, 1438–1439
- Gaussian distribution, 1437–1438
- GAussian Mixture importancE Sampling
 (GAME), 625
- Gaussian process (GP) method, 660–663
- Gaussian process regression (GPR), 1510
- General Circulation Models (GCMs), 16, 17,
 246, 256, 1431
 short range precipitation forecast, temporal
 skill in, 20
- Generalized likelihood function, 583–588
- Generalized linear model (GLM), 806, 1348,
 1352, 1495
 components of, 1454, 1455
- Genetic algorithm-MCMC (GA-MCMC)
 technique, 692
- Genetic algorithms (GAs), 532–534
- Geographical information systems (GIS), 331,
 336, 337
- GFS precipitation forecast skill, 21
 variation of, 21, 22
- Gibbs sampler, 1474
- Gilvenko-Cantelli theorem, 1475
- GIS, *see* Geographical information
 systems (GIS)
- Global Ensemble Forecast System (GEFS),
 123, 127, 1350, 1352
- Global Flood Awareness System (GloFAS),
 1152–1155, 1198, 1207
- Global flood partnership, 1123–1126
- Global search methods, 532–544
- GRACE, *see* Gravity Recovery and Climate
 Experiment (GRACE)
- Gradient-based methods, 525, 529
- Gradient-based SA methods
 FF screening, 640
 MOAT method, 644
 OAT method, 640
 PB screening, 643
- Gravity Recovery and Climate Experiment
 (GRACE), 707, 712, 721, 739
- Green and Ampt equation, 419
- Green vegetation fraction (GVF), 443
- Grey-box models, *see* Conceptual hydrological
 models
- Ground water, 313, 317–318, 322, 328
- H**
- Harmonic motion, 38
- Heteroscedasticity, 1487
- Heteroscedastic maximum likelihood estimator
 (HMLE), 569
- Heuristics, 1138
- Hillslope runoff routing, 420, 421
- Hindcast/ing, 227, 1103
- Histogram, 1440–1441
- Holtan equation, 395
- Horton's equation, 394
- HUGO, 1205
- Human expertise, 1292, 1299
- Hybrid prediction technique, 825
- Hydraulic model, 1387–1389
- Hydrological applications examples, 547–553
- Hydrological cycle, 321–322, 390, 391
 appropriate model type, selection of,
 334–335
 base flow, 320
 climate change on, 324
 definition, 313
 evaporation, 316
 ground water, 317
 human activities and land use, 323–324
 infiltration, 316–317
 interception, 315
 mathematical representation of, 322–323
 model calibration, 333

- overland and channel flow, 318–320
precipitation, 315
transpiration, 316
- Hydrological ensemble forecasts, 916, 917
- Hydrological Ensemble Forecast System (HEFS), 8
- Hydrological ensemble prediction, China
annual precipitation, 1414
conditional probability, 1422–1426
rainfall runoff ensemble prediction, 1419–1422
short range ensemble predictions, 1415–1419
Xinanjiang model, 1414, 1419, 1420
- Hydrological Ensemble Prediction Experiment (HEPEX), 8
- Hydrological ensemble prediction system (HEPS), 960–961, 971, 1189
applications of, 1182–1184
Brier skill score, 964–968
correlation, 961
frequency bias, 961–962
public saliency of, 1137, 1146
rank histogram, 968–969
systems objectives, 1208–1210
target variables, 1207–1208
value score, 964
- Hydrological forcings, 1210–1212
- Hydrological forecast, 242, 245, 978, 1015
ensemble streamflow nowcasts, verification of, 955–960
flood peak and flood timing verification, 971–973
long-term, 982
medium-term, verification, 993–998
mid-term, 980–982
NWP driven forecasts, 955
operational HEPS (*see* Hydrological ensemble prediction system (HEPS))
PREVAH, 954
probabilistic assessment, 984–989
- Hydrological models, 13, 524, 583, 787, 788, 1210–1212, 1387, 1388
advantages and limitations, 328–330
black box/empirical models, 327
classification, 483
deterministic models, 326
emerging technology for, 335–336
fully distributed, physically-based models, 328–329
history of, 330–331
lumped conceptual models, 327
parameter calibration, 526–528
- parameter estimation (*see* Parameter estimation)
statistical models, 330
types of, 325–326
uncertainty in, 25–27, 335
uncertainty quantification, 494–507
- Hydrological processes
distributed hydrological models, 414
evaporation, 418
interception, 418
sub-hydrological processes, 417–418
transpiration, 418–419
- Hydrologic Ensemble Forecast Service (HEFS), 1192, 1350, 1352, 1359, 1365
- Hydrologic post-processing methods, 822
- Hydrologic predictability, 13–15
- Hydrologic routing model, VAR, 756–757
- Hydrology, 5, 1280, 1284, 1285
- Hydrometeorological forecasting, 4, 5, 243, 1038
Bayesian inference, 1451–1452
copulas, 1448–1450
history and application of ensemble, 7–9
hypothesis, 1450–1451
mixed variables and distributions, 1443
non-parametric distributions, 1439–1443
parameter estimation, 1445
parametric distributions, 1437–1439
random number generation, 1445–1446
rationale for, 11–13
statistical method (*see* Statistical forecasting method)
- Hydrometeorological model, 5
- Hydrometeorological process, 676
- Hydrometeorological system, 979
- Hydropower systems, 1080
decision-making process, 1302
forecast quality and value, 1302
human expertise, quality and value of forecasts, 1298–1301
meteorological ensemble predictions and weather services, 1301
operational systems, ensemble streamflow forecasts for, 1295–1298
reservoir-based, 1295
run-of-the-river, 1292, 1294
storage-based, 1292
systematic data assimilation methods, operational implementation of, 1301
time frames in forecasting, 1291
- Hydro-Québec (HQ), 979–980
- HYDRUS-1D model, 616–617
- Hypothesis test, 1450–1451

I

- IDF, *see* Intensity-duration-frequency curves (IDF)
 Ignorance score, 904
 Importance sampling, 592–593
 Increment, 713, 715, 717, 725, 734, 738
 Index-flood method, 379–380
 Infiltration processes, 393–396
 excess-runoff mechanism, 419
 saturation-runoff mechanism, 420
 Informative prior, 574
 Initial condition perturbation
 boundary, 115–117
 bred growing mode, 105
 conditional nonlinear optimal perturbation, 110
 ensemble transform Kalman filter, 111
 ensemble transform with rescaling, 106–108
 random, 103
 scaled time-lagged, 104–105
 singular vector, 108
 Initialization, 709
 Inland waterway transport, 1372–1374, 1376, 1379, 1381
 Innovation, 714, 715, 717, 719, 721, 722, 724, 725, 728, 734, 737
 Input uncertainty model, 751–753
 In-situ stage-discharge measurements, 748–749
 Instantaneous unit hydrograph, 608–610
 Integrated Drought Management Programme, 1126–1127
 Integrated hydrological modelling system (IHMS), 1081
 Intensity-duration-frequency curves (IDF), 1236–1238, 1248
 Interpolation, 247

J

- Japanese Meteorological Administration (JMA), 163, 174
 Join probability density function (PDF), 1435–1436, 1446

K

- Kalman filter(s) (KF), 514, 704, 712, 713, 715, 720, 736, 1052–1053
 applications of, 693
 EKF, 683–684
 linear data assimilation technique, 681–682
 non-linear data assimilation technique, 683–685
 UKF, 684–685

- Kensico Reservoir, 1337
 Kernel density estimation, 1440–1443, 1475
 Kernel dressing, 809, 814
 Korean Meteorological Administration (KMA), 164, 175
 Kostiakov equation, 394
 Kriging, 1506

L

- Land-atmosphere interaction
 land-ABL interaction, 471–474
 NSLAI, 464–471
 Land-surface models (LSMs), 650, 704, 705, 707–710, 714, 717, 728, 736, 737, 1285
 atmospheric forcing data, 438–440
 canopy water content, 449
 cold season processes (*see* Cold season processes)
 land data sets, 440–444
 Noah LSM, 438, 439, 444
 NWP, 84
 soil hydraulics, 456–457
 soil moisture tendency, 449
 soil temperature tendency, 449
 soil thermodynamics, 457–459
 surface energy budget (*see* Surface energy budget)
 surface fluxes, 444
 surface temperature, 455–456
 surface turbulent exchange coefficients, 446–448
 Large-scale hydrologic modeling, 774–775
 Large-scale particle image velocimetry (LSPIV), 750
 LARSIM forecast systems, 1203
 Latent heat flux, 444
 Lateral boundary conditions (LBCs), 78
 Latin hypercube (LH) sampling, 645
 Leaf area index (LAI), 442, 455, 458, 467
 Least squares technique, 497–499
 Least squares theory, 88
 Likelihood function, 575
 generalized, 583
 Linear data assimilation technique, 678–679
 Kalman filter, 681–682
 Linear inverse modeling (LIM), 217
 Linear regression (LR), 654, 1453
 multivariate linear regression, 1491–1492
 quantile regression, 1494
 ridge regression, 1492–1493

- Liuxihe model
automated parameter optimization, 430–432
construction, 426–428
flood forecasting, 431
flow direction, 429
hydrological data, 425
initial parameter values, 427–430
parameter uncertainties, 431
performance validation, 431–433
slopes, 430
soil parameter values, 431
structure, 429
Taiping Watershed, 425, 426
terrain property data, 425–427
- Local feedbacks, 207–208
- Logistic regression, 806, 1454–1455, 1496
- Lorenz attractor, 16
- Lorenz-96 model, 111
- Low-hierarchy climate models, 219
- Low stream flow, 1373, 1374, 1377
- Lumped conceptual model, 406
- M**
- Machine learning methods
ANN, 1501
GPR, 1510
MARS, 1505
random forests, 1505
regression tree, 1505
SVM, 1503
- Madden-Julian Oscillation (MJO), 199, 211, 220
- Marginal likelihood, 622
- Markov Chain Monte Carlo (MCMC)
algorithms, 499, 1472–1474
- Markov Chain Monte Carlo simulation
chain trajectory derived from, 595
differential evolution Markov chain, 598–602
- DREAM algorithm, 602–605
- Gibbs sampler, 1474
- Metropolis-Hastings (MH) algorithm, 594
principle of detailed balance, 593
random walk metropolis algorithm, 596
single chain methods, 596–598
symmetric jumping distribution, 594
- MARS, *see* Multivariate adaptive regression splines (MARS)
- MATLAB code, DREAM, 629–632
- Maximum a posteriori (MAP) estimation, 1471
- Maximum likelihood ensemble filter (MLEF)
vs. EnKF, 770–773
- Maximum likelihood estimation
(MLE), 1445
- McKay correlation ratios, 650
- Mean integrated squared error (MISE), 1476
- Medium-range forecasting, 1084–1085
- Mesoscale surface analyses, 931–932
- Meteorological forcing, 1387
- Meteorological post-processing affects
analogue approach and poor man ensemble, 248
Bayesian post processing methods, 249
model output statistics, 248
spatial and temporal interpolation, 247
- Meteorological Service of Canada (MSC), 165, 176
- Method of moments (MoM), 505, 1444
- Metropolis-Hastings (MH) algorithm, 594
- Microwave brightness temperature, 710, 739
- Millennium drought, 1162
- Mixed variables and distributions, 1443
- MMEFS, 1202
- MOCOM-UA method, 547
- Model calibration, 486, 494, 514
- Model error, 677
- Model evaluation, 938
- Model output statistics, 242, 248
- Model parameter(s)
automatic parameter optimization, 424
classification, 424
determination, 423–424
inference, 261–265
scalar method, 424
- Model physics, 102, 103, 123
- Models, hydrology, *see* Hydrological models
- Moderate resolution imaging spectroradiometer (MODIS), 440, 442
- Modified Puls routing method, 403–404
- Monte Carlo method, 590
acceptance-rejection algorithm, 590–592
importance sampling, 592
Markov chain (*see* Markov chain Monte Carlo simulation)
- Monte Carlo sampling, 645
- Monte Carlo simulation, 6, 7
- Morris one-at-a-time (MOAT) method, 644–646
- Moving average model, 356
- Multi-data hydrologic models, 775–776
- Multi-Ensemble Forecast Processor (MEFP), 1350, 1352
- Multi-model and multi-physics, 118–119
- Multi-model ensemble, 227–228
- Multi-objective optimization, 494

- Multiple model forecast systems
 BHM, 837–838
 BJP, 834–835
 BMA, 835–836
 challenges, 838
 compensatory effects, 838
 weighted resampling, 833–834
- Multiplication rule, 1434
- Multi-purpose, 1303
- Multiscale bias correction (MSBC), 763–767
- Multisensor analysis, 933
- Multivariate adaptive regression splines (MARS), 656–658, 1505
- Multivariate linear regression, 1491–1492
- Multivariate normal (MVN) distribution, 1446–1448
- Multivariate verification, 302, 984
- Muskingum method, 404–405
- N**
- National Centers for Environmental Prediction (NCEP), 166, 177, 180
- National Weather Service (NWS), 8, 1348, 1350, 1352, 1353, 1356
- Navigation, 1372, 1374, 1375, 1379, 1380
- Near-surface land-atmosphere interaction (NSLAI)
 bare soil evaporation, 469–470
 canopy evaporation, 470–471
 potential evaporation, 465–466
 total evapotranspiration, 471
 transpiration, 466–468
- Neighborhood ensemble, 136
- Newton method, 529
- New York City water supply system, 1332
 Catskill System, 1336
 Croton System, 1332
 Delaware System, 1336
 DEP (*see* Department of Environmental Protection (DEP))
 environmental flow requirements, 1343
 OST (*see* Operations Support Tool (OST))
 system interconnections, 1339
 terminal reservoirs, treatment and distribution, 1337
 water quality, 1342
 water supply reliability, 1341
- Non-homogenous Gaussian regression (NGR), 805
- Non-linear data assimilation technique, 678
 Kalman filter, 683–685
 variational methods, 685–686
- Nonlinear interactions, 42, 43
- Nonlinear regression
 logistic regression, 1496
 poisson regression, 1498
- Nonlinear-reservoir bucket model, 492
- Nonparametric methods, 1466, 1474–1477
- North American Multi-model Ensemble (NMME), 1286
- Nowcasting of orographic rainfall (NORA), 1068
- Nowcasting system, 1050–1065
- Nucleus for European Ocean Modelling (NEMO), 161
- Numerical methods and NWP, *see* Numerical weather prediction (NWP)
- Numerical weather prediction (NWP), 68, 438, 442, 706, 714, 715, 733, 734, 739, 740, 1227, 1236, 1256, 1296
 atmospheric model, 48
 atmospheric observations, 69
 basic equations, 70–73
 bottom boundary conditions, 78
 challenges, 94–95
 computer models, 70
 concepts, 36
 coupled numerical models, 87–88
 data assimilation (*see* Data assimilation, NWP)
 development of, 49, 69–70
 dynamic meteorology, 69
 electronic computers, invention of, 69
 FDEs, 73–74
 forecasts, 241
 global and regional models, 79
 grid staggering methods, 76
 initial conditions, 37
 initialization, 46
 land-surface models, 84–86
 LBCs, 78
 mesoscale models, 53
 models, 6, 256, 286, 1098, 1103
 numerical analysis, advances in, 69
 ocean models, 86–87
 physical parameterizations (*see* Physical parameterizations, NWP)
 pillars of, 38
 spectral models, 74–76
 system, 1387
 upper boundary conditions, 76–78
- O**
- OASIS water supply system model, 1343
 demands, 1345
 hydrodynamic water quality model, 1345

- OASIS inflows, 1345
operating rules, 1346
physical data, 1345
- Objective function, 569
- Observational uncertainty, 750–751
- Observation operator, 705, 709–712, 714, 716, 718, 728, 734
- Observing networks, 198, 221
- Occam’s azor, 624
- Ocean-atmospheric coupling, 201, 208, 211, 225
- Ocean mixed-layer (OML) models, 218
- Ocean wave modeling, 86
- One-at-a-time (OAT) method, 640
- Operational flood forecasting services, 1406
- Operational, global medium-range ensembles (OG-ENS), 157, 190
average performance of, 169–179
BMRC Australia, 158–159
characteristics of, 155, 156
CMA China, 159
CPTEC Brazil, 160
ECMWF Europe, 160–163
JMA Japan, 163–164
KMA Korea, 164–165
MSC Canada, 165–166
NCEP United States of America, 166–168
predictability theory to operational ensemble forecasting, 153
UKMO United Kingdom, 168–169
- Operations control language (OCL), 1346
- Operations support tool (OST), 1343
dashboard data visualization, 1348
data acquisition system, 1343
EPP, 1353
OASIS water supply system model
(*see* OASIS water supply system model)
physically based methods, 1351
statistical forecast methods, 1348
- Optimal parameter estimation approaches, 528–547
- Optimization
automatic calibration, 491–493
in Bayesian least squares, 508
goodness-of-fit function, 488–490
multi-objective, 494
- Ordinary differential equations (ODEs), 75
- P**
- Parameter estimation, 638, 639, 1444–1445
a priori estimation, 485
estimation of expensive models, 512
- goodness-of-fit function, 488
initial model states, 512
manual calibration, 487
model calibration, 486
model diagnostics, 507–510
Nash-Sutcliffe efficiency (NSE) metric, 490
non-smooth models, 511
operational improvements, 513
parameter transformations, 510
recursive estimation and data assimilation, 514
root-mean-squared-error (RMSE) metric, 490
sparse-data problems, 513
- Parameter estimation software (PEST), 665
- Parameter screening, 643, 663, 665
- Parametric and structural uncertainty, 749
- Parametric distributions, 1439–1443
- Parametric estimation method, 982
- Parseval’s theorem, 649
- Partial differential equations (PDEs), 68, 69, 73, 76, 80, 677
- Particle filter(s) (PFs), 688
vs. AEnKF, 768–769
applications of, 694
DHM, 759–763
vs. EnKF, 768–769
EPFM, 692
PF-MCMC, 692
SIR, 689–691
SIS, 689
- Particle Filter-Markov Chain Monte Carlo (PF-MCMC), 691
- Particle swarm optimization (PSO) algorithm, 424–425, 536–538
- Pearson correlation coefficient, 900
- Pearson III probability distribution function, 1420
- Pearson product–moment correlation coefficient, 1435
- Penman-Monteith equation, 418
- Periodic motion, 40
- Physically based distributed hydrological models (PBDHMs)
basin divisions, 416
evaporation, 418
excess-runoff mechanism, 419–420
flood forecasting, 419
flow network, 417
higher computational algorithms, 434
hillslope runoff routing, 420–422
interception, 418
Liuxihe model (*see* Liuxihe model)

- Physically based distributed hydrological models (PBDHMs) (*cont.*)
 models proposed, 415
 parameters (*see* Model parameters)
 river runoff routing, 422–423
 saturation-runoff mechanism, 420
SHE model (*see* Système Hydrologique Européen (SHE) model)
 surface runoff (*see* Surface runoff)
 terrain property data, 416–417
 transpiration, 418
 underground runoff and movement, 420–421
- Physical parameterizations, NWP
 basic principles, 80–82
 cloud microphysics and precipitation, 83
 convection, 83
 non-orographic gravity wave drag, 84
 orographic drag, 84
 radiation and chemical processes, 82–83
 soil/surface, 83
 turbulent diffusion and planetary boundary layer scheme, 83–84
- Plackett–Burman (PB) screening, 643–644
- Plotting position, 374
- Poisson regression, 1498
- Poor-man’s ensemble, 134
- Posterior distribution, 588
- Post-processing, 797, 1285
 approaches to hydrological, 787–790
 calibration field consistency, 243–244
 consistency between variables, 246–247
 correlation, 242
 event magnitude, 246
 extreme values, 249–250
 hydrological combination, 250
 hydrological vs. meteorological, 785–787
 interpolation, 242
 meteorological forecasts, 244
 radar precipitation forecasts, 244
 relevant information, 242
 reliability, 242
 requirements and challenges, 790
 seamless forecasting system, 250–251
 sharpness forecast, 242
 shift of tails, 243
 spatial relevance, 245–246
 stationarity, 242
 temporal relevance, 244–245
- Potential evaporation
 evaporative fraction for, 465
 strong land-atmosphere coupling, 466
 weak land-atmosphere coupling, 466
- Precipitation, 241, 313, 315, 321, 322, 324, 336, 391–393
- Precipitation forecast, 20
 skill, seasonal CFS, 22
 temporal scale-dependent uncertainty in, 20
- Predictability, 210
 atmospheric, 48
 chaotic behavior, 41
 complete loss of, 46
 definition, 45
 dependent, 46
 external (or “practical”), 46
 implications for, 45
 intrinsic, 45
 limits of, 40–47
See also Weather and climate systems
- Predictor variables, 270–272
- Pre-operational HEPS systems, 1192
- Pre-processing, 797
- Principal components regression (PCR), 822
- Prior distribution, 573
- Probabilistic forecast/forecasting, 851, 852, 978, 1075–1076
- Probabilistic forecasting, flood inundation
 event description, 1390
 forecast evaluation, 1389
 forecast skill scores, 1394–1396
 hydraulic modelling, 1388–1389
 hydrological modelling, 1388
 modelling framework, 1387
 numerical weather prediction, 1387
 probability maps, 1392–1394
- Probabilistic hydrological forecast, 979–980
- Probabilistic risk, 1359
- Probability
 definition, 1464
 distributions, 1465–1483
 in hydrometeorological forecasting, 1464, 1465
- Probability density function (PDF), 678, 1432
- Probability integral transform diagram, 907–909
- Probability mass function (PMF), 1432
- Probability of precipitation forecasts, 1139
- Production planning model, 998–1008
- Pure fuzzy systems, 370
- Q**
- Quantile mapping (QM), 248
- Quantile regression (QR), 802, 1078, 1493–1494

- Quantitative precipitation estimate (QPE)
mosaics, 933
- Quantitative precipitation forecasting (QPF), 87
- R**
- Radar, 932–933
backscatter, 705, 711
forecasts, 1056
- Radar ensemble generator designed for usage in the Alps using LU decomposition (REAL), 956, 957, 959
- Radiative transfer model, 710, 711, 717, 718, 737
- Rainfall-runoff models, 526
- Rainfall-runoff transformation, 610–612
- Rain gauge analyses, 932
- Random forests, 1503
- Random parameters (RP) scheme, 168
- Random perturbation, 103–104
- Random variables, 1431–1432
distribution function, 1433
expectation of, 1433
probability density and distribution functions, 1432
- Random walk Metropolis (RWM) algorithm, 593, 596
- Realtime experiment, *see* Radar Ensemble generator designed for usage in the Alps using LU decomposition (REAL)
- Real-time modeling, 1343
- Recursive state estimation scheme, 1052–1053
- Reforecasts, 1226, 1227
- Regional regression models, 380
- Regression, 789
coefficients, 1490–1491
generalized linear models, 806
logistic, 806–807
non-homogenous gaussian regression, 805
quantile regression, 802–805
tree, 1503
- Regression-based SA methods
DT, 659
GP method, 661
linear regression, 654–656
MARS, 656
SOT model, 658
- Regression models
applications, 353
multiple linear regression, 351–353
simple linear regression, 347–350
- Reliability, 907, 1103
- Remote sensing, 335, 337, 414, 416, 426
measurements, 749
- Reservoir, 1290, 1297, 1300, 1303
operation(s), 1308, 1316, 1323, 1325, 1342, 1343, 1345
- Residual error models, 504
- Rhine basin, 1072–1075
- Richards equation, 420
- Ridge regression, 1492–1493
- Risk communication, 1133, 1140, 1155
- Risk perception, 1155
- Risk reduction, 1111, 1126
- River ice, 1373, 1374, 1376, 1378
- River Information Service (RIS), 1376
- River Rhine, 1374, 1377, 1379, 1380
- River routing, 402
- River runoff routing, 420, 422
- Root-mean-square-error (RMSE), 175, 176, 180
- Runoff generation, 399–402
- RWsOS Rivers, 1197
- S**
- Sacramento model, 409–410
See also Sacramento Soil Moisture Accounting (SAC-SMA) model
- Sacramento Soil Moisture Accounting (SAC-SMA) model, 547, 549
- Saint-Venant equations
hillslope runoff routing, 421
river runoff routing, 422
- Sampling, 640, 645, 646, 665
- Sampling importance resampling (SIR), 689–691
- Satellite observations, 933
- Scale motion, 46, 48
- Schaake shuffle, 258, 296, 798
- Score estimation, 990–993
- Scoring rules
mathematical definition of, 987
rejection, 988
types, 986
variogram-based, 989
verification, 993
- Screen-level observations, 704, 709–710, 714, 733, 734, 739
- Sea ice modeling, 87
- Seamless prediction methods, hydrologic forecasts, 294–295
- Seasonal CFS precipitation forecast skill, 22–23
- Seasonal forecast/ing, 1081–1084, 1162, 1280, 1281, 1283, 1383
accuracy, 1166–1169
CRPS, RMSE and RMSEP skill scores, 1172–1173
forecast precision, 1173

- Seasonal forecast/ing (*cont.*)
 tercile hit rates, for low and high flows, 1173
 aggregated forecast performance, 1169–1170, 1174
 Bayesian joint probability model (BJP), 1163–1165
 challenges, 1163
 reliability, 1166, 1170
 uncertainty in GCM, 23
 verification, 1165
- Seasonal hydrologic streamflow prediction, 14, 15, 17
- Seasonal prediction, 197, 199, 212, 224, 227
- Seasonal runoff forecasts, 830
- Seasonal streamflow forecasts
 dynamical hydrology models, 823–825
 hybrid and multi-model forecasting, 825
 hydrologic predictions, 821
 statistical/empirical forecasting, 822–823
- Seemingly seamless, 250
- Sensitivity analysis (SA), 638, 663–666
 gradient based methods, 640–646
 qualitative methods, 639
 quantitative methods, 639
 regression-based methods, 654–663
 variance-based methods, 646–653
- Sequential importance sampling (SIS), 689
- Service Central d'Hydrométéorologie et d'Appui à la Prévision des Inondation (SCHAPI), 1142, 1145
- Severity, 1280, 1283
- Shipping, 1372, 1374, 1376, 1379
- Short-range ensemble forecasts
 analogs, 799–802
 Bayesian model averaging, 812–813
 ensemble dressing, 809–811
 regression (*see* Regression)
- Short range GFS precipitation forecast skill, 21–22
- Short range hydrological forecast,
see Hydrological forecast
- Shuffled complex evolution (SCE) algorithm, 424
See also Shuffled complex evolution-UA (SCE-UA)
- Shuffled complex evolution metropolis (SCEM-UA) algorithm, 599
- Shuffled complex evolution-UA (SCE-UA), 540–541
- Shuttle Radar Topography Mission (SRTM), 1390, 1393
 DEM database, 416
- Simplex Downhill method, 525
- Simulated annealing (SA) algorithm, 534–536
- Single chain methods, 596
- Single-component Metropolis Hastings, 1474
- Single variable linear regression, 1488
 ANOVA, 1488–1490
 regression coefficients, 1490–1491
- Singular vectors (SVs), 108–110
- SIR, *see* Sampling importance resampling (SIR)
- Skill score(s), 910–912, 956, 959, 961, 964–968, 979, 984, 995, 996, 998, 999, 1002
- SMAP, *see* Soil Moisture Active Passive (SMAP)
- Smoothing, 678
- SMOS, *see* Soil Moisture Ocean Salinity (SMOS)
- Snow sublimation, 460
- SNOW-17 temperature index snow model, 1350
- Sobol' sensitivity indices, 651
- Soil conservation service (SCS) method, 401–402
- Soil heat flux, 450, 462
- Soil hydraulics, 456
- Soil moisture, 449, 1280, 1284
 retrieval, 705, 707, 709, 715, 716, 733, 734, 736
 storage, 396
- Soil Moisture Active Passive (SMAP), 707, 711, 719, 728, 733, 736, 739
- Soil Moisture Ocean Salinity (SMOS), 706, 707, 719, 723, 728, 729, 732, 736, 737, 739
 satellite mission, 1390
- Soil temperature, 449
- Soil thermodynamics, 457
- Spaghetti plots, 1133, 1134
- Spatial forecast verification, 938–941
- Spatial scale, 11, 23, 24
- Spatial statistics, 939
- Spatio-temporal dependence structures, 297
- Spectral prognosis (S_PROG) model, 1055–1056
- Spread-skill plot, 909–911
- Stage-storage method, 403
- Standardized regression coefficient (SRC), 655
- State Pollutant Discharge Elimination System (SPDES), 1334
- State-space models, 677
- State update, 706, 718, 720, 722, 725
- Statistical dependence
 correlation, 1435
 covariance, 1435
 definition of, 1435

- Statistical/empirical forecasting approach, 822–823
Statistical forecasting method, 1348
 ANN, 1455
 autoregressive, 1453
 ensemble forecast, 1456–1457
 linear regression, 1453
 logistic regression, 1454–1455
 SVM, 1455–1456
Statistical independence, 1435
Steepest descent method, 529
Stochastically perturbed parameterization tendency (SPPT) scheme, 119–120
Stochastic boundary-layer humidity (SHUM), 122
Stochastic convective vorticity (SCV) scheme, 168
Stochastic-Dynamic approach, 7
Stochastic kinetic energy backscatter (SKEB) scheme, 120–122
Stochastic processes, 241
Stochastic total tendency perturbation (STTP) scheme, 123–129, 167
Storage, 1291, 1294
Straight-line model, 487, 489
Streamflow, 746, 1280, 1284, 1285
Streamflow data assimilation
 direct and crowdsourced measurements, 750
 in-situ stage-discharge measurement, 748–749
 large-scale hydrologic modeling, 774–775
 multi-data hydrologic models, 775–776
 observational uncertainty, 750–751
 remote sensing measurement, 749
 timing errors, 776–777
Streamflow forecasts, 788
 hydropower system (*see* Hydropower systems)
 See also Seasonal forecasting
“String of beads” model, 1054–1055
Sub-seasonal forecast range, 189
Sum-of-trees (SOT) model, 658–659
Super-ensemble, 134
Supply reliability, 1339, 1340, 1347, 1359, 1362
Support vector machine (SVM), 1455–1456, 1502
Surface energy budget, 439, 443, 445, 449–450
 canopy evaporation, 452–453
 direct soil evaporation, 453–454
 linearized surface energy balance, 451
 plant transpiration and canopy resistance, 454–455
potential evaporation, 451–452
soil heat flux, 450–451
surface evapotranspiration, 452–453
Surface evapotranspiration, 452
Surface fluxes, 444–445
Surface local feedbacks, 207
Surface runoff
 horizontal and subsurface flow, 420
 routing, 417, 420
 vertical and subsurface flow, 420
Surface temperature, 455
Surface temperature of the sea (SST), 165
Surface turbulent exchange coefficients
 explicit formulations, 447–448
 implicit formulations, 446–447
Surface water, 322
Surface water treatment rule (SWTR), 1342
Surrogate modeling based methods, 541–544
SVM, *see* Support vector machine (SVM)
Swiss forecasting system, 1074
Systematic errors, 788
Système Hydrologique Européen (SHE) model
 automatic parameter optimization, 424
 MIKE 11 modeling package, 422
 Saint-Venant equations, 421
 surface runoff movement, 420
System modeling, OASIS, *see* OASIS water supply system model
- T**
- Takagi-Sugeno-Kang (TSK) fuzzy systems, 370–372
Tangent linear and Adjoint Model Compiler (TAMC), 686
Tangent linear model (TLM), 110
Tank model, 406
Teleconnection(s), 198, 220
 empirical orthogonal, 208
 patterns, 208
 spatial, 201
 statistical techniques for, 209
Teleconnection indices (TCI), 1083
Temperature diurnal cycle, 246
Temporal scale, 11, 23, 25
Temporal scale-dependent uncertainty, in precipitation forecasts, 20
Terrestrial water storage, 709, 712, 739
Thomas-Fiering model, 357–360
Thorpex Interactive Grand Global Ensemble (TIGGE), 1197
operational, global medium-range ensembles (*see* Operational, global medium-range ensembles (OG-ENS))

- THORPEX program, 212
 Time-lagged ensemble, 134
 Time series models
 applications, 361
 nonstationary, 357–358
 stationary, 354–357
 types of, 370–372
 Total probability rule, 1434
 Traffic, 1373, 1374, 1376
 Transient eddy kinetic energy (TEKE), 129
 Turbidity, 1334, 1336, 1338, 1345, 1354, 1360, 1362
- U**
 UK Met Office (UKMO), 168, 178, 1145
 Uncertainty, 335, 787–788, 1212–1214
 in environmental systems modeling, 566
 in GCM seasonal forecasts, 23–25
 hydrologic model, 25
 public understanding of, 1136, 1137
 Uncertainty analysis (UA), 638, 639
 Uncertainty quantification (UQ), 638
 aggregational approaches, 500
 Bayesian inference, 495–496
 boundary conditions, 679
 decompositional approaches, 504
 ensemble simulations, 680–681
 generalized likelihood uncertainty estimation (GLUE), 506
 least squares technique, 498
 method of moments, 505
 model parameters, 680
 model structure, 680
 probabilistic simulations, 680
 tools for Bayesian posterior analysis, 499
 Underground runoff and movement, 420
 Uninformative prior, 574
 Unit hydrograph method, 400–401
 Univariate verification, 301–302
 Unscented Kalman filter (UKF), 684–685
 Users, 1292
 Users' requirements, 1214–1215
 U.S. National Centers for Environmental Prediction (NCEP), 288
- V**
 Vadose zone hydrology, 612–619, 621–622
 Validation, 328, 334
 van Genuchten-Mualem (VGM) model, 616
 Variable infiltration capacity (VIC), 1280, 1281, 1284, 1285
 Variance-based SA methods
 ANOVA, 646
- EFAST, 650
 FAST, 648
 McKay correlation ratios, 650–651
 Sobol' sensitivity indices, 651–652
 Variance decomposition, 646, 650, 652
 Variational assimilation (VAR), hydrologic routing model, 756–757
 Variational data assimilation (DA) methods
 applications of, 693
 4D-Var, 686
 Verification
 cross-validation, 273–274
 metrics, 897, 898, 902
 univariate, 302
 VIC, *see* Variable infiltration capacity (VIC)
 Vigilance, 1400–1401, 1409
 Vine copula, 1450
 Virtual ensembles, 134
 analog ensemble, 137–139
 dual-resolution hybrid ensemble, 135–136
 neighborhood ensemble, 136–137
 poor-man's ensemble, 134
 time-lagged ensemble, 134
 Visualization, 1133–1137, 1142, 1150
 Vorticity confinement, 129–130
- W**
 Water-level forecast, 1214, 1374, 1376, 1377, 1379
 Water management, 1292
 Water partition and balance (WAPABA) model, 257
 Water quality management, 1341
 Water resource management, New York City,
 see New York City water supply system
 Water resources forecasting, 1101
 Watershed, 322, 331, 333, 335, 336, 567
 Water supply reliability, 1340
 Weather and climate systems, from Mesoscale to Global Scales, 47–60
 Weather Research and Forecasting (WRF) Model, 122
 Weighted ensemble dressing, 811
 Weighted resampling, 833
- X**
 Xinanjiang model, 407, 1414, 1419, 1420
- Y**
 Yuba-Feather FCO ensemble-based decision support model, 1098