

Seon Ki Park · Liang Xu *Editors*

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. III)

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. III)

Seon Ki Park · Liang Xu
Editors

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications (Vol. III)



Springer

Editors

Seon Ki Park
Environmental Science and Engineering
Ewha Womans University
Seoul
Republic of Korea

Liang Xu
Marine Meteorology Division
Naval Research Laboratory
Monterey, CA
USA

ISBN 978-3-319-43414-8
DOI 10.1007/978-3-319-43415-5

ISBN 978-3-319-43415-5 (eBook)

Library of Congress Control Number: 2016950903

© Springer International Publishing Switzerland 2017

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.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To Yoshi K. Sasaki

Preface

Since the 2007 Annual Meeting in Bangkok, the Asia Oceania Geosciences Society (AOGS) has hosted a series of sessions on data assimilation (DA), named “Yoshi K. Sasaki Symposium on Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications.” We started this symposium after the first DA session in the 2005 AOGS meeting in Singapore to honor Dr. Yoshi Kazu Sasaki, the former George Lynn Cross Professor of the School of Meteorology at University of Oklahoma, for his lifelong contributions to DA. Yoshi was a pioneer in broaching the use of variational method to produce optimal initial conditions for numerical prediction. His approach is further developed and nowadays widely adopted in various prediction systems in geosciences. Unfortunately and sadly we lost this great scientist on March 12, 2015, and we decided to make this volume be a memorial to Dr. Sasaki. Yoshi had contributed superior chapters to the previous two volumes—both frontline works on tornado DA.

The first volume of this book was published in March 2009 with 27 chapters—a collection of notable invited papers along with those selected from the previous symposiums. Among many important chapters in the volume, John M. Lewis, one of Yoshi’s Ph.D. students, contributed a chapter titled “Sasaki’s Pathway to Deterministic Data Assimilation.” Milija Županski, the last Ph.D. student of Yoshi, discussed theoretical and practical issues of ensemble DA. François-Xavier Le Dimet, who was a postdoctoral scientist under Yoshi’s supervision, described application of the variational approach to hydrologic DA. Yoshi himself proposed a new theory based on the entropic balance, titled “Real Challenge of Data Assimilation for Tornadogenesis.” I. Michael Navon provided a thorough review of DA for numerical weather prediction, especially on 4D-Var, which became the most cited chapter in this series (66 times as of April 2016 from Google Scholar).

The second volume was published in May 2013, again with 27 chapters, by collecting both invited papers and selected papers from the previous symposiums held in 2009 (Singapore), 2010 (Hyderabad) and 2011 (Taipei). Volume II included excellent overviews of estimation theory, nudging and variational methods, and Markov chain Monte Carlo methods. Most prominently, Yoshi extended his entropy balance theory for tornado DA from the previous volume, and contributed a

chapter titled “Entropic Balance Theory and Radar Observation for Prospective Tornado Data Assimilation.”

We have felt a need to publish another volume because some new research results had been presented at the symposiums since publication of the last volume. When we heard about Yoshi’s passing, we immediately decided to publish a memorial volume for him. Here we include a special dedication section, titled “In Memory of Yoshi,” by collecting memories on and photos of Yoshi from some authors. This volume includes excellent overviews of variational DA (François-Xavier Le Dimet et al.), coupled system DA (Milija Županski), representer-based variational DA (Boon Chua and Liang Xu), and soil moisture DA (Viviana Maggioni and Paul Houser).

In this volume, theoretical and methodological aspects encompass inverse theory, variational methods with/without adjoint model, representer-based variational method, quantification of information and forecast uncertainty, sensitivity tools, error representation modeling, the maximum likelihood ensemble filter, ensemble forecast, conditional nonlinear optimal perturbation approach, etc., with applications to oceanic, atmospheric, and land surface DA; coupled atmosphere-chemistry DA; stratospheric and mesospheric DA; terrestrial ecosystem; bottom topography mapping; radar/lidar/satellite assimilation; adjoint sensitivity; and targeting observations. Operational 4D-Var applications are also included for the JMA Nonhydrostatic Model (NHM) and the US Navy Coastal Ocean Model (NCOM).

This book will be useful to individual researchers as well as graduate students as a reference to the most recent progresses in the field of data assimilation. The publication is partly supported by the Korea Environmental Industry & Technology Institute through the Eco Innovation Program (ARQ201204015). We appreciate Boon Chua at Naval Research Laboratory, Takeshi Enomoto at Kyoto University, and François-Xavier Le Dimet at University of Joseph Fourier, who have served as the co-conveners of the Sasaki Symposium. We are very honored to dedicate this book to the late Yoshi Sasaki—our friend and mentor, for his monumental contributions to the advance of data assimilation.

Seoul, Republic of Korea
Monterey, USA
April 2016

Seon Ki Park
Liang Xu

In Memory of Yoshi

John M. Lewis

Yoshi, My Teacher

I came to University of Oklahoma (OU) in August 1963 to work on one of Professor Yoshikazu Sasaki's National Science Foundation (NSF) research projects. I had no intention to work for a degree beyond M. Sc. in geophysics, I just received from University of Chicago. A phone call from Professor George Platzman to Professor Walter Saucier sealed the deal.

I taught a refresher course in math for the AFIT [Air Force Institute of Technology] students in August and then started the research. Instead of grinding out some data analysis, Professor Sasaki gave me a problem based on current work by two of his colleagues, Professor Platzman at U of C and Professor Akio Arakawa at UCLA. It was a problem in numerical weather prediction (NWP). He took the spectral form of Burgers' nonlinear advection equation (Burgers 1939), a surrogate for the barotropic vorticity equation, and introduced me to the yet unknown form of Arakawa's innovative physically based finite differencing scheme. Both works yet unpublished (Platzman 1964; Arakawa 1966). He told me to go to Bizzell Memorial Library and check out *Methods of Mathematical Physics, Volume 2* (Courant and Hilbert 1962). He said: "with the following long wave initial condition, find the analytic solution to this problem by the method of characteristics." All new to me—spectral form of solution, truncation of the basis functions, a novel form of finite differencing applied to the grid-point version of the model, and analytic solution via the method of characteristics. Wow! I was excited. Then he laid the problem before me: "You'll find that the barotropic wave breaks and it will be your job to determine the limitations of Arakawa's scheme near the point of breaking. When you get some result come back and we'll talk." What a joyful feeling filled me: I am free, I understand the problem, I have no idea of the result, but it looks interesting and challenging, and if I can solve it, I'll contribute to Professor Sasaki's NSF project and likely learn something myself.

It took me about three and a half months to solve the problem—from late September to early January. When I returned from my Christmas holiday in California, I visited “Doc” (as all of us called him based on the revered term used by his first doctoral student—Rex Inman). I laid out the analytic solution with the beautiful breaking wave and the difficulty encountered by Arakawa’s innovative scheme near the point of breaking (Lewis 1964). He was ecstatic. Neither before nor since have I ever felt so filled with success.

I needed a break from academia and took a job with Shell Oil Company—oil exploration in the Gulf of Mexico via seismic prospecting. I got married to Sherry McDowell, the Mayo Clinic nurse who took part of her training at Lying-In Hospital in Chicago. Although I was making money by the barrowful, I kept contact with Professor Sasaki. I missed him so much that I asked Sherry if she would consider leaving her job in Houston and moving to Norman where I could work with him again. She encouraged me and back we went to Oklahoma. I learned variational analysis from Doc and got my degree in 1969. Fleet Numerical Weather Central (FNWC) in Monterey wanted to use Yoshi’s variational method to develop an operational upper air analysis over the data-sparse Pacific Ocean to help direct U.S. Navy ships and planes from the West Coast to the South China Sea during the Vietnam War. I got that job and Yoshi became a hero in absentia after Tom Grayson, USAF Colonel Bob Long, and I got the variational method to work over the entire globe. It remained operational for nearly 25 years—unheard of in operations. It ran every 6 hours for all those years—so robust, hardly ever broke down, and so efficient. This product, labeled the Global Band Analysis, was a favorite of the Navy forecasters in places like Guam and Rota-Spain. When this operational analysis was about to be “dethroned” in the early 1990s, Ed Barker, a Sasaki protégé and a meteorologist at Naval Research Laboratory (NRL), told me the forecasters in the field begged to have the Global Band retained. What a tribute to Yoshi.¹

As I look back at my apprenticeship under Professor Sasaki, I view him as a philosopher as well as a scientist, a Socratic scholar-teacher who knew how to bring out the best in every student. Questioning, no preaching, not telling, simply presenting, and then let the student go. When I left OU, I truly felt there was no problem I could not solve—maybe a little overconfident, but not full of myself. Yoshi was self-effacing and I like to think I inherited some of that from him.

One of my great thrills was to get a Christmas card from Professor Sasaki a year or two after I left OU. The writing inside the card is attached (Fig. 1). The spirit of Yoshi’s words makes it clear how much we enjoyed learning together. It was mostly a “one-way street” despite Yoshi’s compliments. But I never hesitated to doubt his idea and his derivation and he accepted that. But his acceptance came

¹Sasaki took sabbatical leave to NEPRF (Navy Environmental Prediction Research Facility) in 1973. NEPRF was collocated with FNWC and the Naval Postgraduate School (NPS). At the end of his sabbatical year, he was offered the position of Director of Research at NEPRF. He declined and returned to OU. Nevertheless, his influence of the U.S. Navy’s work in NWP has been long lasting.

with the admonition, “OK, so you don’t accept my idea and derivation; fine, work on it tonight and present your idea to me tomorrow.” That was Socratic and that is how I learned in part from the master. But most of my learning “came at his feet,” in his office, and at the blackboard. I appreciated University of Chicago and University of Oklahoma (Figs. 2–4) and their wonderful graduate programs, but I always felt that I had some special training not associated with either institution; I was trained at University of Tokyo, second generation, so I thought of myself as a scientific Nisei. Forever, Thank You Doc.

John M. Lewis
Visiting Research Professor, Desert Research Institute
Ph.D. in Meteorology, University of Oklahoma, 1969
Protégé of Yoshikazu Sasaki

References

- Arakawa A (1966) Computational design for long-term numerical integration of the equations of fluid motion: Two-dimensional incompressible flow. Part I *J Comp Phys* 1:119–143
- Burgers JM (1939) Mathematical examples illustrating relations occurring in the theory of turbulent fluid motion. *Verhandelingen der Koninklijke Nederlandse Akademie van Wetenschappen (Eerste Sectie)*, 17(2):53 p
- Courant R, Hilbert D (1962) Methods of mathematical physics, Vol 2. John Wiley and Sons, 830 pp
- Lewis JM (1964) Computational instability arising from numerical solutions to the non-linear advection equation. Unpublished manuscript, 24 pp
- Platzman GW (1964) An exact integral of complete spectral equations for unsteady one-dimensional flow. *Tellus* 16:422–431

John, Sherry, Debra:

Have a wonderful and nice Christmas! In
~~our~~ ^{our} (or these) short stay in this wonderful world,
Koko and I always remember every nice things
about you. Our children are enjoying toys
soon left for them. Koko and I are very often
talking about you. Especially, to John, you
really made unreplaceable assistance to me
for the work which is significant. I have
been trying hard to establish "Weather and
Environment Analysis Techniques Development Program"
on which I received considerable interest
Yoshi

Have very nice
holiday season!

Merry Christmas!

May all your hopes and dreams
come true
in the New Year *

Yoshi, Koko

Otto, Jimmy, Larry x Anna
Sasaki

Fig. 1 A Christmas card from Yoshi (ca 1972)



Fig. 2 Yoshi at his office in the Engineering Lab in 1970, taken by Jim Heimbach, one of Yoshi's doctoral students



Fig. 3 Yoshi outside Felgar Hall, OU Campus (ca 1965)

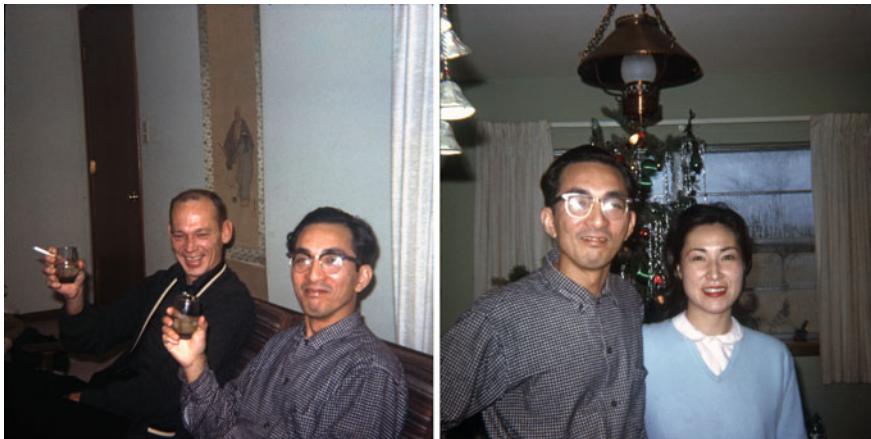


Fig. 4 Rex Inman and Yoshi (*left*) and Yoshi and Koko (*right*) (ca. 1965)

François-Xavier Le Dimet

In the 1980s, the numerical models for predicting the evolution of geophysical flows were known to have spectacular improvements mainly due to the increase of computational power. But a model, whatever be its quality, is not sufficient to carry out a prediction: it is mandatory to provide an initial condition deduced from dynamical observation of the flow. “Analysis” was the term used to build up the initial condition and at that time it was the weak point of the prediction process.

I had read some papers written by Yoshi Sasaki and especially those published in *Monthly Weather Review* in 1970, which were based on the Calculus of Variations to mix the information provided by models and the information contained in observations.

I was a former student of Jacques-Louis Lions and I had attended his courses on Optimal Control for Partial Differential Equations at Université Pierre et Marie Curie in Paris. Optimal Control is based on Calculus of Variations, if data assimilation is considered from this viewpoint then the dynamics is easily introduced by considering the initial condition as the control variable. The price to be paid is to have to deal with a more complex Optimality System, because it will take into account the evolution of the adjoint variables (Lagrange Multipliers). In 1981, I wrote a report on how to use Optimal Control for Data Assimilation and sent it to Yoshi. Several weeks later (no e-mail at that time!) I received an invitation from Yoshi Sasaki. I arrived in Norman in May 1982 and I was received by Yoshi and Koko Sasaki, their welcome was so warm and friendly! On OU campus I had an office at the Cooperative Institute for Mesoscale Meteorological Studies (CIMMS), headed by Yoshi. It was a small building close to the railway track. I stayed for

4 months at OU, working on some developments in applications of Variational Methods in meteorology. I introduced Luther White to Yoshi. Luther was at the Math Department and he has been working with Lions during a stay at INRIA near Paris. He was working on inverse problems a very close topic to DA. Cooperation with Luther lasted many years and several papers have been published mainly on applications of DA in hydrology. During my stay, Yoshi invited Roland Glowinski from Université Pierre et Marie Curie and from the University of Houston and also a former Lions student to give a talk at OU on inverse problem. I visited OU for many years either at CIMMS or at the Math department or at NSSL with John Lewis. I also had a strong cooperation with Baxter Vieux in hydrology; in this domain there is a strong demand for DA and inverse modeling. Thanks to Yoshi a solid link was created between OU and the universities of Clermont-Ferrand where I had my position. Millie Audas in charge of the Office of International Affairs at OU pushed for student exchanges and for more than a decade she informed me on the number of marriages celebrated between students of our universities, then came the first divorce!

With Yoshi and Luther we discussed about organizing a Symposium on Variational Methods in Geosciences: it was held at Norman in May 1985, with Jacques Louis Lions as an invited speaker. It was the first meeting between Sasaki and Lions and it has been followed by several other meetings in USA, France, and Japan when Lions became the chairman of the French Space Agency (CNES). A few years later Lions received the Japan Prize, a very high distinction in Japan. During the stay at OU I wrote the first draft of a paper published in 1986 in Tellus, based on several CIMMS scientific reports. The meeting in Norman was quite successful and 30 years after it remains in the air. I am still remembering Koko playing some Japanese music instrument after the banquet.

Probably because of this event I was asked by World Meteorological Organization to co-organize the first WMO International Symposium on the Assimilation of Observations in Meteorology and Oceanography; it was held in Clermont-Ferrand in July 1990, and more than 250 scientists attended this meeting, with Lions and Sasaki as invited speakers. We took advantage of the meeting to have some sight seeing tour in Auvergne with Yoshi, he was very glad to taste the typical cooking of Auvergne cheese and “tripoux.” A dozen of people from OU came to the WMO meeting, we were helped by Emily Glasgow from OIP for receiving our hosts (Fig. 5). Yoshi and Koko visited France several times and I have been happy and proud to host them home (Fig. 6).

After moving to Université Joseph Fourier in Grenoble, I visited OU less frequently nevertheless I have always been in touch with Yoshi, Luther, and Baxter.

For the last time I met Yoshi and Koko at the AMS annual meeting held in New Orleans in 2012, Yoshi was accompanied by Koko and their son, Yoshi received the title of Honorary Member of the American Meteorological Society and delivered a short talk on his scientific achievements.

For me Yoshi was a great scientist, a mentor, and a friend. I will always remember his kindness and warmness. Thanks to him I was opened to a new world. All the people who had met Yoshi will not forget him.

Writing these lines I found that Lions and Sasaki were both born in 1928, both had to suffer from WWII, Sasaki in a destroyed Japan and Lions joined the “Résistance” when he was 15 years old; nevertheless both used their scientific aura to promote peace and friendship in the world. May we follow their paths.

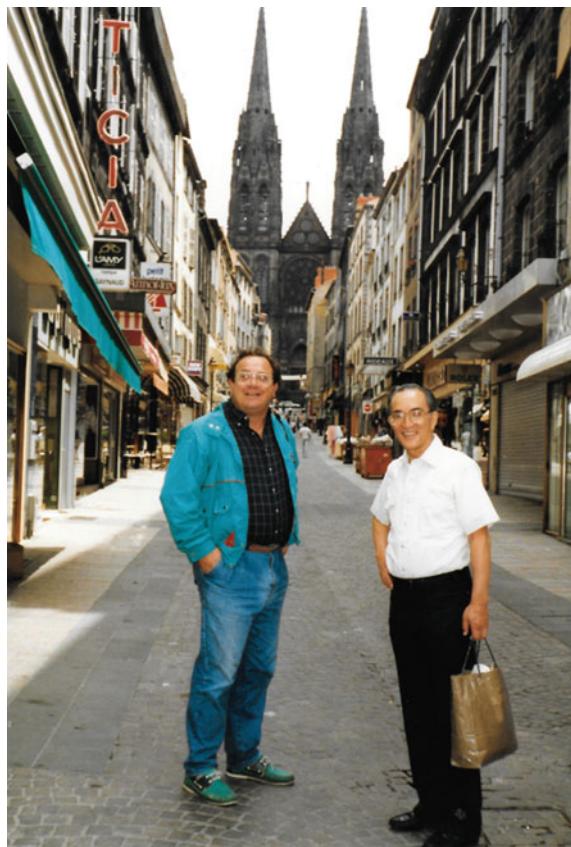
François-Xavier Le Dimet

Professeur Emérite, Lab. Jean-Kuntzman, Université Joseph Fourier
Postdoctoral Scientist, CIMMS/University of Oklahoma, 1982



Fig. 5 Yoshi at the 1st WMO Symposium on Data Assimilation in Clermont-Ferrand, France in July 1990. *Left*: Yoshi on the *first row*, Akira Kasahara at the *left end* on the *second row*, and Regis Juvanon du Vachat, Luc Fillion and Richard Ménard, from *left to right*, on the *third row*. *Right*: From *left to right* Christian Pariset, Yoshi, and Werner Wergen

Fig. 6 François-Xavier Le Dimet and Yoshi in Clermont-Ferrand, France (ca. 1988). The cathedral of Clermont-Ferrand is in the background, which is black because it was built with volcanic stones



Milija Županski

I have fond memories of Prof. Yoshi Sasaki. He was my principal advisor for both M.S. and Ph.D. degrees at University of Oklahoma. Yoshi was very generous in allowing me to find my own way in research, and helping me when I needed help. I greatly appreciated this freedom. Yoshi was also very supportive of my desire to go beyond standard variational applications in meteorology, eventually leading to several one-to-one courses.

There was a time, however, which I remember in most vivid detail. I was at the time a Ph.D. student with Yoshi and worked on Alpine lee cyclogenesis using variational methods to solve a set of nonlinear equations. A conference related to this subject was held in Sestola, Italy (near Parma) that we attended. Also, Koko joined Yoshi on this trip. We rented a car from the airport in Rome, and he let me drive a really nice Italian car. On our way to Sestola, we visited Florence and Pisa,

and talked about the Renaissance art, artistic details of old buildings, and many other things. I enjoyed this trip very much, and still cherish the time I have got to spend with Yoshi and Koko in a relaxed atmosphere outside the work.

Milija Županski
Senior Research Scientist, CIRA/Colorado State University
Ph.D. in Meteorology, University of Oklahoma, 1990

Jidong Gao

The paper in this volume is dedicated to the late Dr. Yoshi Sasaki who is the founding father of the variational data assimilation in numerical weather prediction, and its applications to radar meteorology. He also had a genuine personality and treated any individual he met with respect. He proposed the initial idea for the application of variational calculus in numerical weather prediction in 1955 in his Ph.D. study. He published three important papers in 1970 when he was a professor at the School of Meteorology, University of Oklahoma (OU). One of them, titled “Some basic formalisms in numerical variational analysis,” laid the foundation for the development of the four-dimensional variational data assimilation method in late 1980s (so-called 4DVAR) which continues to be used operationally in the world’s major meteorological centers to this day. In 1988, he was funded by NASA for a project titled “Variational High-resolution Data Assimilation and Short-Range Weather Forecasting” which focused on application of the variational analysis to radar data—still a very hot research topic in the meteorological community. Personally, I learned a great deal from Dr. Sasaki and his many important publications. When I joined the Cooperative Institute of Mesoscale Meteorological Studies/OU as a visiting scholar in 1995, I could barely speak English. But I visited his office many times from 1995 to 2004 in OU Sarkeys Energy Center and asked questions about his research and discussed my research with him. He was always very kind and friendly. He listened to my questions very patiently and carefully and tried his best to make sure that I understood his answers. I gained a lot of knowledge about variational data assimilation directly from him during that period of time. I will always remember him as one of the greatest scientists I ever met in my life.

Jidong Gao
Research Meteorologist, National Severe Storm Laboratory/NOAA
Research Scientist, CIMMS/CAPS/University of Oklahoma, 1995–2010

S. Lakshmivarahan

Reminiscences on Dr. Yoshi Sasaki

My name is S. Lakshmivarahan and I joined the then School of Electrical Engineering and Computer Science (EECS), University of Oklahoma (OU) in the fall of 1978. Since day one, I had heard a lot about Professor Yoshi Sasaki and his groundbreaking work in variational methods for data assimilation. While we shared offices in different floors of the same building—the old Engineering Lab, it was only in late 1980s I was introduced to him by one of his former student, John M. Lewis, National Severe Storms Laboratory (NSSL) when John and I started working in this area. John, Kelvin Drogemeier and I collaborated in designing and offering a two level system of courses in Dynamic Data Assimilation for the first time here at OU. It was offered as a graduate-level course at the School of Meteorology for the first two years and but these courses were moved over to the newly formed School of Computer Science within the College of Engineering at OU where it has been offered as Scientific Computing I and II. This effort culminated in the publication of our book, Lewis et al. (2006). Dr. Sasaki has been our guiding light and he had given total and unconditional support for our efforts and we shall remain eternally grateful to him.

While all of Dr. Sasaki's work were deeply entrenched in the classical deterministic approach, he was curious to find out the underpinnings of stochastic modeling. I had the pleasure of having him attend one of my courses on “Stochastic Differential Equations and Its Applications to Finance.” He attended the first half dealing with the exposition of the Stochastic Calculus as developed by K. Ito. He later shared his immense joy of meeting Professor K. Ito at the Kyoto University during one of his trips to Japan.

Professor Sasaki single handedly convinced the mighty Hitachi Corporation to endow a Professorship—known as the Hitachi Chair, only the second of its kind (the first of such chair was established at the Stanford University) at the School of Computer Science, OU. Again thanks to generous support from the local Hitachi corporation office in Norman, Oklahoma, the Hitachi Distinguished Lecture series was established at our School. Dr. Sasaki never missed a lecture in this series and had always interacted with the visiting scholars with his views and questions. Continuing our spirit of great regard and admiration, we have dedicated our recent monograph (Lakshmivarahan et al. 2016) to the memory of Professor Yoshi Sasaki.

S. Lakshmivarahan
George Lynn Cross Research Professor, Computer Science,
University of Oklahoma

References

- Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic Data Assimilation: A Least Squares Approach. Cambridge University Press
- Lakshmivarahan S, Lewis JM, Jabrzemski R (2016) Forecast Error Correction using Dynamic Data Assimilation, Springer (To appear)

Kazuo Saito

Photos with Prof. Sasaki and JMA's Condolences to His Wife

I visited the Center for Analysis and Prediction of Storm of Oklahoma University and the National Weather Center of NOAA in March 2012 subsequent to the 11th National Severe Weather Workshop held in Oklahoma City (Fig. 7). On the evening of March 6, after my seminar, I visited the laboratory of Prof. Sasaki. He was very fine and he introduced me his post-doctoral student and his latest works. He regretted that he would not be able to meet MRI's recent invitation to visit Tsukuba for health concern over the physical condition of his wife, Koko, and expressed his hope to visit Japan at the next opportunity. I was impressed by his activity and enjoyed discussion with Prof. Sasaki for about one hour.

In April 2016, at the sad news of loss of Prof. Sasaki, the then Director-General of the Japan Meteorological Agency (JMA), Noritake Nishide, sent official condolences to his wife regarding his tremendous contribution to the improvement of Japanese meteorological service (Fig. 8). Prof. Sasaki mentored several visiting scientists from Japan at Oklahoma University, and his leading came to fruition at JMA as the operational mesoscale model and its variational data assimilation systems. His supervision to JMA visitors about Doppler radar observations and tornado nowcastings also significantly promoted their practical implementation.



Fig. 7 Photos taken by Kazuo Saito during his visit to OU in March 2012. Kazuo and Yoshi at Yoshi's office at OU (right)

In Memory of Yoshi

きりあることとなっております。
このほかにも、防災気象情報の調査のために議員が米国に滞
在した際、米国の官民連携の状況や方法について、実視聴をもど
に現地の方々を交えてご教説いただいたことがございます。この
ことは、気象衛星のP/F-1事業や緊急地震速報の導入について

例えは、乳業業者の人々の一つである数値化専門分野では、
さとう木暮先生の突然の引退に接した際は、思わず
泣き落とす先生でした。お電話になつたことは、枚挙に暇があり
ません。

Fig. 8 Official condolences by JMA to Yoshi's wife

Kazuo Saito
Senior Director for Research Affairs
Meteorological Research Institute, Japan

Liang Xu

Yoshi's NRL Monterey Connection

One of the important predecessors of the Marine Meteorology Division (MMD) at the Naval Research Laboratory (NRL) in Monterey, California, USA is the Naval Environmental Prediction Research Facility (NEPRF). There is a deep connection between the data assimilation systems used at MMD in NRL and the late Professor Yoshi Sasaki. It is not very widely known that Yoshi was instrumental in helping laying the foundation of the numerical modeling and data assimilation for the US Navy in Monterey, CA, USA. I would like to take the opportunity to provide an excerpt from “The Genesis of Numerical Modeling and Data Assimilation” by Rosmond and Barker (http://www.nrlmry.navy.mil/MMD_History/text/index.htm).

In 1975-1976, Professor Yoshi Sasaki from the University of Oklahoma spent a year as acting director of research of the new Naval Environmental Prediction Research Facility (NEPRF). NEPRF was now co-located with Fleet Numerical Weather Central (FNWC) on the airport annex, and had access to their computer systems and data bases. During this time FNWC was running the Kessel-Winninghoff hemispheric PE model as their primary NWP system. It was run in both hemispheres, but left a major void in the tropics. They had an in-house effort to develop a global model, but were struggling to support it. The FNWC commanding officer at the time, Capt. Ron Hughes, realized that they could not maintain the long-term continuity of research effort needed to develop and maintain a large global NWP system. He came to NEPRF and requested that the lab assume responsibility for NWP operational forecast system development in support of FNWC. In response, Prof. Sasaki and the NEPRF Commanding Officer, Capt. Cody Sherar, formed the Numerical Modeling Department (NUMOD), with Tom Rosmond as the department head.

To jump start the NEPRF global modeling effort, in 1976 Tom Rosmond spent 2 weeks at UCLA getting a crash course on the UCLA general circulation model, at the time considered one of the best global models in the world, under the guidance of Professor Akio Arakawa. This model was the basis for the first generation NEPRF global forecast system. The other components of the system were a Barnes successive corrections objective analysis and a variational balance equation initialization scheme, both developed by Ed Barker. Professor Sasaki's expertise with variational methods was a contributing factor in the development of these systems.

Figure 9 is an excerpt from “A History of the U.S. Navy’s Numerical Objective Analysis and Data Assimilation Systems” presented by Dr. Edward Barker at the Symposium on the 50th Anniversary of Operational Numerical Weather Prediction.

I first met Yoshi at the SIAM Conference on Mathematical and Computational Issues in the Geosciences (GS03) on March 20, 2003, in Austin, Texas. Yoshi and I were both in the session entitled “Data Assimilation in the Ocean and Atmosphere.” After our session concluded, Yoshi, his wife Koko, Professor Andrew Bennett, and I strolled the streets of downtown Austin. I enjoyed their company and had a great time.

The last time I met Yoshi and Koko was in June 19, 2008 at the 2nd Sasaki Symposium in Data Assimilation for Atmospheric, Oceanographic, and Hydrologic Applications as a participant and a co-convener. I enjoyed Yoshi’s presentations and scientific discussions during the symposium, and I equally enjoyed the time that Yoshi and his wife Koko, Professor Seon K. Park and his Ph.D. student H. Kim, Dr. H.-S. Lee, and I spent together at a beautiful Korean restaurant not too far from the convention center. Figure 10 includes two pictures that I took at the restaurant. I still vividly remember that Yoshi was very good at encouraging Ms. Kim, the young Ph.D. student, to keep working in the data assimilation field. We also had the opportunity to discuss and appreciate the Chinese, Japanese, and Korean cultures among other interesting topics at the restaurant.

Liang Xu
Meteorologist, Naval Research Laboratory

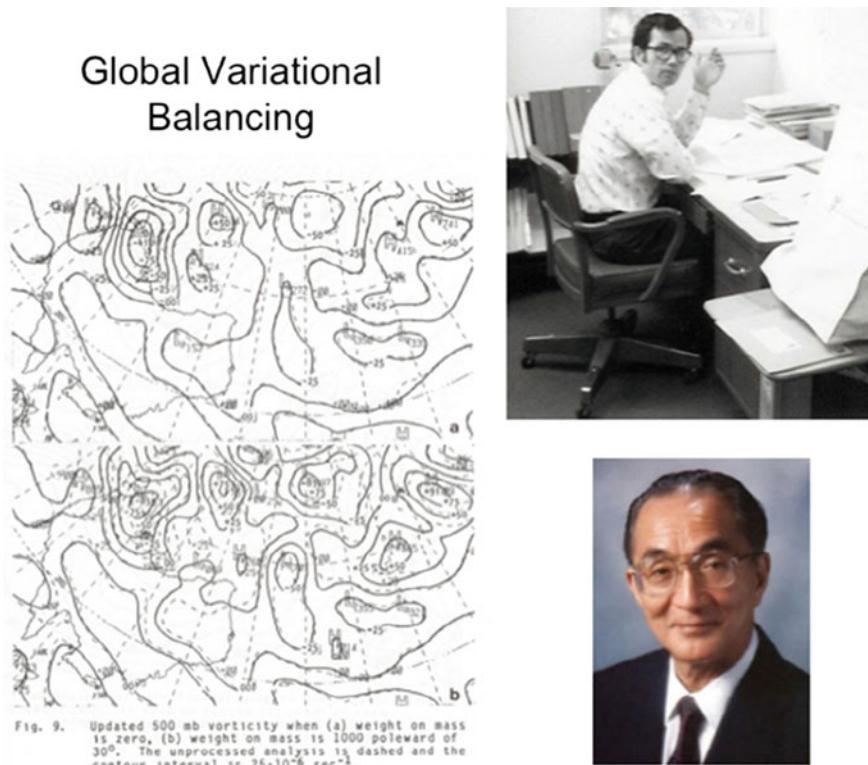


Fig. 9 The 500 hPa vorticity analysis based on the global variational nonlinear balancing algorithm suggested by Yoshi (Sasaki 1958). The person on the right top of the picture is Dr. Edward Barker who was responsible to implement the variational algorithm. This figure was provided through the courtesy of Dr. Edward Barker



Fig. 10 Koko and Yoshi (left) and, from left to right, Dr. H.-S. Lee, Ms. H. Kim, and Prof. Seon K. Park (right) in a restaurant in June 2008 at Busan, Korea

Reference

- Sasaki Y (1958) An objective analysis based on the variational method. *J Meteor Soc Japan*, 36:738–742

Seon Ki Park

Yoshi, My Mentor

My memory with Yoshi dates back to the Fall Semester 1990, when I first met him as a new Ph.D. student at OU School of Meteorology. I came to know him through his course, *Mesoscale Dynamics*, during the first semester at OU. One day Yoshi took me to Kelvin Droegemeier and asked if he could serve as my academic advisor. I owe a debt of gratitude to Yoshi that allowed me to have such a wonderful advisor as Kelvin.

I always enjoyed taking his courses, especially because his classes were full of discussions between him and students. He always encouraged me to ask many questions and initiate discussions in the classes. He invited me many times to the teatime discussions in his office after classes. Through those teatime discussions, I could grow up in many aspects. Yoshi not only was a good instructor but also became my lifelong mentor. He also helped me a lot in completing my dissertation as a member of the advisory committee.

After I came back to Korea with a faculty job at the Ewha Womans University, I've never forgotten Yoshi's pioneer spirit of establishing the Meteorology program at OU. In 2007, my university's first meteorology-related research institution, called the Severe Storm Research Center, was established. In 2009, another institution, named Center for Climate/Environment Change Prediction Research, was established based on a grant from the Korean government. Finally the Department of Atmospheric Science and Engineering was founded in 2012 and has accepted students from meteorological centers from developing countries through the Ewha-WMO Fellowship Program. All these achievements were inspired by Yoshi's pioneer spirit.

I started convening a session in the Asia Oceania Geosciences Society, titled *Yoshi K. Sasaki Symposium on Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications* in honor of Yoshi for his lifelong contribution to data assimilation in geosciences. Yoshi and Koko attended the second symposium at Busan, Korea in 2008, and gave a special lecture. I have also hosted an event called "Dinner with Yoshi" under the sponsorship of KMA, and we had joyful time with about 50 invited scientists and students (see the photos Figs. 11–14).

Liang Xu and I have published a series of editorial books under the same title as the symposium with the Springer. The first volume was published in 2009 and was dedicated to Yoshi and Koko. The second volume was published in 2013 with dedication to Yoshi and Roger Daley. In these two volumes, Yoshi himself contributed chapters on new directions in tornado data assimilation. Now at the sad news of Yoshi's passing, we hereby publish the third volume as the memorial volume for Yoshi. The symposium and the book series will continue in memory of Yoshi. I do appreciate everything that I have received from you, Yoshi, and I miss you a lot.



Fig. 11 The AOGS poster announcing Yoshi's Special Lecture on Data Assimilation (left), and Yoshi, Koko and Seon K. Park in front of the poster (right) at Busan, Korea in June 2008



Fig. 12 At the Dinner with Yoshi co-hosted by the Korean Meteorological Administration and Ewha Womans University at BEXCO, Busan, Korea on June 18, 2008



Fig. 13 Seon K. Park, Koko and Yoshi (*left*), and Shigeo Yoden, Tieh-Yong Ko, Yoshi, Koko, Seon K. Park and Hadi Tri Wahyu at the Dinner with Yoshi



Fig. 14 Koko and Yoshi talking with François-Xavier Le Dimet and Oliver Talagrand (*left*), and Yoshi giving a speech (*right*) at the Dinner with Yoshi

Seon Ki Park
Professor, Environmental Science and Engineering
Professor, Atmospheric Science and Engineering
Ewha Womans University
Ph.D. in Meteorology, University of Oklahoma, 1996

Contents

Variational Data Assimilation: Optimization and Optimal Control	1
François-Xavier Le Dimet, Ionel M. Navon and Răzvan Ștefănescu	
Data Assimilation for Coupled Modeling Systems	55
Milija Županski	
Representer-Based Variational Data Assimilation Systems: A Review	71
Boon S. Chua and Liang Xu	
Adjoint-Free 4D Variational Data Assimilation into Regional Models	83
M. Yaremchuk, P. Martin, G. Panteleev, C. Beattie and A. Koch	
Convergence of a Class of Weak Solutions to the Strong Solution of a Linear Constrained Quadratic Minimization Problem: A Direct Proof Using Matrix Identities	115
S. Lakshmivarahan	
Information Quantification for Data Assimilation	121
Sarah King, Wei Kang, Liang Xu and Nancy L. Baker	
Quantification of Forecast Uncertainty and Data Assimilation Using Wiener's Polynomial Chaos Expansion	141
Junjun Hu, S. Lakshmivarahan and John M. Lewis	
The Treatment, Estimation, and Issues with Representation Error Modelling	177
Daniel Hodyss and Elizabeth Satterfield	
Soil Moisture Data Assimilation	195
Viviana Maggioni and Paul R. Houser	

Surface Data Assimilation and Near-Surface Weather Prediction over Complex Terrain	219
Zhaoxia Pu	
Recent Developments in Bottom Topography Mapping Using Inverse Methods	241
Edward D. Zaron	
The Impact of Doppler Wind Lidar Measurements on High-Impact Weather Forecasting: Regional OSSE and Data Assimilation Studies	259
Zhaoxia Pu, Lei Zhang, Shixuan Zhang, Bruce Gentry, David Emmitt, Belay Demoz and Robert Atlas	
A Three-Dimensional Variational Radar Data Assimilation Scheme Developed for Convective Scale NWP	285
Jidong Gao	
Data Assimilation Experiments of Refractivity Observed by JMA Operational Radar	327
Hiromu Seko, Ei-ichi Sato, Hiroshi Yamauchi and Toshitaka Tsuda	
Assessment of Radiative Effect of Hydrometeors in Rapid Radiative Transfer Model in Support of Satellite Cloud and Precipitation Microwave Data Assimilation	337
Peiming Dong, Wei Han, Wei Li and Shuanglong Jin	
Toward New Applications of the Adjoint Sensitivity Tools in Data Assimilation	361
Dacian N. Daescu and Rolf H. Langland	
GPS PWV Assimilation with the JMA Nonhydrostatic 4DVAR and Cloud Resolving Ensemble Forecast for the 2008 August Tokyo Metropolitan Area Local Heavy Rainfalls	383
Kazuo Saito, Yoshinori Shoji, Seiji Origuchi and Le Duc	
Validation and Operational Implementation of the Navy Coastal Ocean Model Four Dimensional Variational Data Assimilation System (NCOM 4DVAR) in the Okinawa Trough	405
Scott Smith, Hans Ngodock, Matthew Carrier, Jay Shriver, Philip Muscarella and Innocent Souopgui	
Stratospheric and Mesospheric Data Assimilation: The Role of Middle Atmospheric Dynamics	429
Saroja Polavarapu and Manuel Pulido	
A Coupled Atmosphere-Chemistry Data Assimilation: Impact of Ozone Observation on Structure of a Tropical Cyclone	455
Seon Ki Park, Sujeong Lim and Milija Županski	

Contents	xxx
Adjoint Sensitivity with a Nested Limited Area Atmospheric Model	467
Clark Amerault	
On the Impact of the Diabatic Component in the Forecast Sensitivity	
Observation Impact Diagnostics.	483
Marta Janisková and Carla Cardinali	
Application of Conditional Nonlinear Optimal Perturbation to Target	
Observations for High-Impact Ocean-Atmospheric Environmental	
Events	513
Qiang Wang and Mu Mu	
Responses of Terrestrial Ecosystem to Climate Change: Results	
from Approach of Conditional Nonlinear Optimal	
Perturbation of Parameters	527
Guodong Sun and Mu Mu	
Index	549

Contributors

Clark Amerault Naval Research Laboratory, Monterey, CA, USA

Robert Atlas NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, FL, USA

Nancy L. Baker Naval Research Laboratory, Monterey, CA, USA

C. Beattie Department of Mathematics, Virginia Tech, Blacksburg, VA, USA

Carla Cardinali European Centre for Medium-Range Weather Forecasts, Reading, UK

Matthew Carrier Stennis Space Center, Naval Research Laboratory, Bay St. Louis, MS, USA

Boon S. Chua Science Applications International Corporation, Monterey, CA, USA

Dacian N. Daescu Portland State University, Portland, OR, USA

Belay Demoz University of Maryland, Baltimore County, Baltimore, MD, USA

François-Xavier Le Dimet Lab. Jean-Kuntzman, Université Grenoble-Alpes, GRENOBLE Cedex 9, France

Peiming Dong Beijing Piesat Information Technology Co., Ltd., Beijing, China; Numerical Prediction Center, Chinese Meteorological Administration, Beijing, China

Le Duc Japan Agency for Marine-Earth Science and Technology, Yokohama, Japan; Meteorological Research Institute, Tsukuba, Japan

David Emmitt Simpson Weather Associates, Charlottesville, VA, USA

Jidong Gao NOAA/National Severe Storms Laboratory, National Weather Center, Norman, OK, USA

Bruce Gentry NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

Wei Han Numerical Prediction Center, Chinese Meteorological Administration, Beijing, China

Daniel Hodyss Naval Research Laboratory, Marine Meteorology Division, Monterey, CA, USA

Paul R. Houser George Mason University, Fairfax, VA, USA

Junjun Hu School of Computer Science, University of Oklahoma, Norman, OK, USA

Marta Janisková European Centre for Medium-Range Weather Forecasts, Reading, UK

Shuanglong Jin Numerical Prediction Center, Chinese Meteorological Administration, Beijing, China

Wei Kang Applied Mathematics Department, Naval Postgraduate School, Monterey, CA, USA

Sarah King Naval Research Laboratory, Monterey, CA, USA

A. Koch Department of Marine Science, University of Southern Mississippi, Hattiesburg, MS, USA

S. Lakshmivarahan School of Computer Science, University of Oklahoma, Norman, OK, USA

Rolf H. Langland Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

John M. Lewis Desert Research Institute (DRI), Reno, NV, USA; National Severe Storm Laboratory (NSSL), Norman, OK, USA

Wei Li Numerical Prediction Center, Chinese Meteorological Administration, Beijing, China

Sujeong Lim Department of Atmospheric Science and Engineering, Ewha Womans University, Seoul, Republic of Korea; Korea Institute of Atmospheric Prediction System, Seoul, Republic of Korea

Viviana Maggioni George Mason University, Fairfax, VA, USA

P. Martin Naval Research Laboratory, Stennis Space Center, Bay St. Louis, MS, USA

Mu Mu Function Laboratory for Ocean Dynamics and Climate, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China; Key Laboratory of Ocean Circulation and Wave, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China; State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China; Institute of Atmospheric Sciences, Fudan University, Shanghai, China

Philip Muscarella Stennis Space Center, Naval Research Laboratory, Bay St. Louis, MS, USA

Ionel M. Navon Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

Hans Ngodock Stennis Space Center, Naval Research Laboratory, Bay St. Louis, MS, USA

Seiji Origuchi Meteorological Research Institute, Tsukuba, Japan

G. Panteleev International Arctic Research Center, University of Alaska, Fairbanks, AK, USA

Seon Ki Park Department of Environmental Science and Engineering, Ewha Womans University, Seoul, Republic of Korea

Saroja Polavarapu Environment and Climate Change Canada, Toronto, Ontario, Canada

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

Manuel Pulido Department of Physics, FACENA, Universidad Nacional del Nordeste and CONICET, Corrientes, Argentina

Kazuo Saito Meteorological Research Institute, Tsukuba, Japan

Ei-ichi Sato Meteorological Research Institute, Tsukuba, Japan

Elizabeth Satterfield Naval Research Laboratory, Marine Meteorology Division, Monterey, CA, USA

Hiromu Seko Meteorological Research Institute, Tsukuba, Japan

Yoshinori Shoji Meteorological Research Institute, Tsukuba, Japan

Jay Shriver Stennis Space Center, Naval Research Laboratory, Bay St. Louis, MS, USA

Scott Smith Stennis Space Center, Naval Research Laboratory, Bay St. Louis, MS, USA

Innocent Souopgui Stennis Space Center, University of Southern Mississippi, Hattiesburg, MS, USA

Guodong Sun State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing, China

Răzvan Ștefănescu Department of Mathematics, North Carolina State University, Raleigh, NC, USA

Toshitaka Tsuda RISH/Kyoto University, Uji, Japan

Qiang Wang Key Laboratory of Ocean Circulation and Waves, Institute of Oceanology, Chinese Academy of Sciences, Qingdao, China; Function Laboratory for Ocean Dynamics and Climate, Qingdao National Laboratory for Marine Science and Technology, Qingdao, China

Liang Xu Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

Hiroshi Yamauchi Japan Meteorological Agency, Tokyo, Japan

M. Yaremchuk Naval Research Laboratory, Stennis Space Center, Bay St. Louis, MS, USA

Edward D. Zaron Department of Civil and Environmental Engineering, Portland State University, Portland, OR, USA

Lei Zhang Department of Atmospheric Sciences, University of Utah, Salt Lake City, USA

Shixuan Zhang Department of Atmospheric Sciences, University of Utah, Salt Lake City, USA

Milija Županski Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO, USA

Variational Data Assimilation: Optimization and Optimal Control

François-Xavier Le Dimet, Ionel M. Navon and Răzvan řtefanescu

1 Introduction

In the last few years due to a constant increase in the need for more precise forecasting and nowcasting, several important developments have taken place in meteorology directed mainly in two different directions: modeling at either large scale or at smaller scales to include an ever increasing number of physical processes and parametrization of subgrid phenomena and adding new sources of data such as satellite data, radar, profilers, and other remote sensing devices. While this led to an abundance of widely distributed data it also created difficulties since most of the information is heterogeneous in space or time and present different levels of accuracy.

Therefore, a cardinal problem is how to link together the model and the data. This problem induces several questions: (i) How to retrieve meteorological fields from sparse and/or noisy data in such a way that the retrieved fields are in agreement with the general behavior of the atmosphere? (Data Analysis); (ii) How to insert pointwise data in a numerical forecasting model? This information is continuous in time, but localized in space (satellite data for instance)? (Data assimilation problem) (iii) How to validate or calibrate a model (or to invalidate it) from observational data? The dual question in this case being how to validate (invalidate) observed data when the behavior of the atmosphere is predicted by a numerical weather prediction model.

F.-X. Le Dimet (✉)
Lab.Jean-Kuntzman, Université Grenoble-Alpes, BP 53, 38041
GRENOBLE Cedex 9, France
e-mail: fxld@yahoo.com

I.M. Navon
Department of Scientific Computing, Florida State University,
483 Dirac Science Library, Tallahassee, FL, USA

R. řtefanescu
Department of Mathematics, North Carolina State University,
Box 8205, Raleigh, NC, USA

For these questions a global approach can be defined by using a variational formalism.

1.1 *Historical Perspective*

Numerical weather prediction has started in the '50 with Charney et al. (1950) when an atmospheric forecast was obtained after an integration of a mathematical model starting from an initial condition. In the early years of numerical meteorology this initial condition was determined by optimal interpolation i.e. an interpolation weighted by statistics on the atmospheric fields. When Yoshi Sasaki arrived in Oklahoma, a state frequently devastated by tornadoes, he worked on mesoscale meteorology; the methods of large scale meteorology can not be directly applied at this scale due to the lack of observations and of statistics on rare events. Interpolation methods tend to regularize the fields at mesoscale level where the identification of discontinuities such as squall lines is of great importance. To retrieve these fields, Sasaki, in his pioneering basic papers has proposed to use the mathematical model itself as a constraint in order to retrieve the atmospheric fields: variational methods applied to meteorology were born. At the same period that the Optimal Control methods for Partial Differential Equation were developed, Lions (1971) pioneered the theoretical basic support of these methods, consisting of the Calculus of Variations in adequate functional spaces.

Optimal Control makes the difference between a “State Variable” and a “Control Variable” and permits to alleviate the difficulties linked to the determination of boundary and/or initial conditions in numerical models. The '80 and '90 have witnessed important improvements in the computational tools with advent of high performance and parallel supercomputers, the development of more precise numerical weather prediction models along with a better understanding of the underlying atmospheric physics and the coverage of networks of observations especially with launching of dedicated satellites. A consequence was that of rendering obsolete the optimal interpolation methods, mainly because the retrieved fields were not in agreement neither with the physics of the models nor with their dynamics. At the end of this period variational methods were successfully introduced in many national operational centers.

1.2 *Variational Methods in Meteorology: A Perspective*

There are two main approaches employed when modeling a system described by a state variable, X . The first approach consists of finding a set of equations F such that X is the unique solution of the state equation

$$F(X) = 0. \quad (1.1)$$

In most cases system F must have as many equations as X has components in order to possess a unique solution—this is the problem of closure. In meteorology this problem has often been solved by using various artifacts such as adding supplementary equations. The second approach to the problem of closure is the variational one consisting in finding X as the solution of a problem of optimization i.e. by finding the extremum of some known functional J .

Such an approach was proposed in theoretical mechanics more than 250 years ago by Euler (1952, 1766) and by Lagrange (1761, 1762). In the domain of numerical analysis Sobolev or Galerkin type methods are also based upon variational principles (Ritz 1908; Galerkin 1915).

In meteorology, using the most general terms, we assume the state of the atmosphere to be described by the set of equations (1.1).

As mentioned, if this system possesses fewer equations than unknowns, the system is said to be non-closed. However, one can still close it by introducing a variational approach.

If X_{obs} denotes an observation of a meteorological field, we will choose from among all the solutions of the system $F(W) = 0$, the solution closest to the observation X_{obs} . The resulting solution will be the optimal solution. In this manner a connection is established between the data and the observations.

In meteorology, the first application of variational methods has been pioneered by Sasaki et al. (1955) and Sasaki (1958). Later on, Washington and Duquet (1963), Stephens (1966, 1968) and Sasaki (1969, 1970a, b) have given a great impetus towards the development of variational methods in meteorology.

In a series of basic papers Sasaki (1969, 1970a, b) generalized the application of variational methods in meteorology to include time variations and dynamical equations in order to filter high-frequency noise and to obtain dynamically acceptable initial values in data void areas. In all these approaches, the Euler-Lagrange equations were used to calculate the optimal X .

Numerous other manuscripts applying these ideas appeared in the meteorological literature during the 70's using the variational formulation. In parallel with the introduction of variational methods in meteorology, starting in the 60's and 70's, mathematicians in coordination with other scientific disciplines have achieved significant advances in optimization theory and optimal control, both from the theoretical viewpoint as well as from the computational one. In particular significant advances have been achieved in the development of optimization algorithms (Gill et al. 1981; Fletcher 2013; Powell 1982; Bertsekas 1982; Lungenberger 1984 to cite but a few).

Optimal control methods have been introduced by Pontryagin et al. (1962), and they have been generalized for systems governed by partial differential equations (Lions 1971).

The application of an optimal control theory to meteorological problems has for the first time supplied the correct framework for a unified approach to analysis, data assimilation and initialization for meteorological problems.

Other techniques strongly related to variational and optimization theory, such as optimum interpolation, Kalman-Bucy filtering (Ghil et al. 1981), smoothing splines (Wahba 1975, 1981a,b), Kriging, generalized cross-validation (GCV) (Wahba and Wendelberger 1980) have also emerged. For a unified approach Lorenc (1986) manuscript could be consulted.

1.2.1 Variational Methods: For Which Purposes?

The first applications of variational methods targeted the objective analysis of meteorological fields, i.e. to retrieve fields from pointwise distributed data in space. In most of the important meteorological situations the temporal evolution of the fields is crucial, therefore, some attempts were carried out towards extending variational analysis to dynamic analysis. Introducing sparsity of data in time using variational tools has led to 4-D data assimilation for numerical weather prediction models. To perform a forecast a meteorological model requires an initial condition. This initial condition must be as close as possible to the observations while remaining compatible with the model. The problem of initialization may be stated as a variational problem and solved in this way.

A general formalism of variational problems has to deal with observations but these observations may not necessarily be physical ones. For instance they may result out of a numerical model (output of a numerical model). Furthermore, the constraints imposed upon the analysis may have no physical origin and could only have been introduced for numerical purposes.

Many applications were carried out in similar situations as mentioned above resulting in a global approach of variational methods, such as for instance enforcing conservation of integral invariants in numerical models (Navon 1981; Navon and De Villiers 1983), or design of discretization schemes (Sasaki 1976). A major difficulty for the classical approach to variational methods for meteorologically significant problems, in particular for those where dynamics play a prominent part, is the fact that the size of the discrete problem to be solved is prohibitive.

A way to circumvent this difficulty is to introduce optimal control methods permitting a significant reduction of the problem size. These techniques, upon which we will expand in a later section, introduce the adjoint of the numerical model. Knowledge of the adjoint of the model turns out to be particularly useful, because it can be applied towards a sensitivity analysis (Hall et al. 1982; Cacuci and Hall 1984) or for environmental studies such as the estimation of the impact of industrial pollution upon the environment (see Marchuk (1982)).

In this review paper we will present the most important contributions concerning applications of variational methods using the general formalism of mathematical programming.

1.3 *Variational Methods in Meteorology: The Optimization Theory View Point*

Numerical weather prediction is based on the integration of a dynamic system of partial differential equations modeling the behavior of the atmosphere.

From a mathematical view point this approach is equivalent to the classical Cauchy problem. Therefore discrete initial conditions describing the state of the atmosphere have to be provided prior to the integration.

In order to retrieve a complete description of the atmosphere one can add information to the raw data using one of the following families of several methods: (i) Perform a simple interpolation, no information is added to the data. This procedure is purely algorithmic; (ii) Add as information the statistical structure of the fields and use an optimal interpolation type method. Unfortunately this information is not always available or may be inadequate for instance as is the case with a paroxysmal event; (iii) variational method.

Variational methods are based on the fact that a given meteorological observation has not an intrinsic credibility. The same measurement of wind, to give just an example, may be used to study the flow around a hill, or may be inserted in a mesoscale model, or may be used in a global model of atmospheric circulation. According to the particular framework where the data will be used, variable trust will be attributed to the same data.

Variational methods try to achieve a best fit, with respect to some ‘a priori’ criterion, of data to a model by placing the data into the most adequate framework where it should be used, and permits us to link the data and the model.

In the first part of the paper we will show how variational methods can be defined and which are the ingredients necessary to build a variational method, all this in the perspective of the surveyed accumulated work. Then we will show how to solve related variational problems in the framework of a systematic classification of the reviewed work. This classification will permit us to review different variational methods as well as the context in which they were performed.

The last section will be devoted to future developments and potential applications of variational methods in meteorology.

2 Ingredients of a Variational Method

2.1 *Definition of a Variational Method*

In the most condensed way a variational method may be defined as a search, amongst all the possible solutions of a model, of the solution closest to a given observation. Therefore a variational method will be defined by the following ingredients:

- An atmospheric variable X , describing the state of the atmosphere.
- A model which may be mathematically written as:

$$\frac{dX}{dt} + A(X) = 0 \quad (2.1)$$

where A is a linear or non-linear operator.

We suppose that system (2.1) is not closed by which we mean that in order to obtain an unique solution to (2.1) some additional information has to be provided.

- A control variable U that may comprise the initial conditions, boundary conditions, or both, the vector X itself or a part of it. Once U is defined—a unique solution $X(U)$ of (2.1) will be associated with it. The vector control variable U must belong to some set of admissible control U_{ad} . The definition of U_{ad} may include physical information which can be stated in the form of inequalities.
- An observation X_{obs} of the meteorological fields.
- J , a cost function measuring the difference between a solution of (2.1) associated with U and the observations X_{obs} .

The variational problem is determined in terms of these last five items and it can be stated as following problem:

Determine U^* which belongs to U_{ad} and minimizes the cost function J . (2.2)

The second stage of the solution of the variational problem will be to determine, or at least to approximate U^* (and therefore the optimal associated state of the atmosphere $X(U^*)$).

In order to achieve this, we first have to set up an optimality condition and then to perform an algorithm for solving problem (2.2).

2.1.1 The Optimality Condition

A general optimality condition is given by the variational inequality (see Lions (1968))

$$(\nabla J(U^*), V - U^*) \geq 0 \text{ for all } V \text{ belonging to } U_{ad}, \quad (2.3)$$

where ∇J is the gradient of the functional J with respect to the variable U .

In the case where U_{ad} has the structure of a linear space, variational inequality (2.3) is reduced to the equality

$$\nabla J(U^*) = 0. \quad (2.4)$$

2.1.2 The Algorithm of Solution

As stated above—variational problems are problems of optimization with or without constraints. There exist standard procedures (Le Dimet and Talagrand 1986; Navon and Legler 1987) to solve them.

A common requirement of these procedures is the need to explicitly supply the gradient of J with respect to U to the code.

Moreover, the basic problem to be solved is always a problem of unconstrained minimization for which the method of conjugate gradient may be used (see Navon and Legler (1987)).

3 Variational Analysis

Basically, the problem of retrieving meteorological fields X from observations X_{obs} , in such a way that X verify some model:

$$F(X) = 0 \quad (3.1)$$

and are as close as possible, in the sense of a given functional J , to the observations X_{obs} , is a problem of optimization with constraints.

Sasaki (1970a,b) in historical paper has introduced two formalisms. The *weak constraint formalism* consists in minimizing without constraint the functional J defined by

$$J_1(X) = J(X) + K\|F(X)\|^2. \quad (3.2)$$

It is easily seen that for large values of K , $F(X)$ has to be small for minimizing J_1 , therefore, for a specified value of K , constraint (3.1) is only approximately verified. In what follows K is a generic constant used as a coefficient of a weak constraint. This is justified by the fact that Eq. (3.1) is not a perfect representation for the atmosphere and therefore should not be satisfied with a greater precision than its own accuracy.

The optimal condition, which in the Euler-Lagrange equation gives the optimal analyzed field X^* , is the solution of the equation

$$\nabla J_1(X^*) = \nabla J(X^*) + 2K \cdot F'(X^*) \cdot F(X^*) = 0. \quad (3.3)$$

In this equation ∇J_1 (respectively ∇J) is the gradient of J_1 (respectively J) with respect to X , while F' is the Jacobian matrix of F . No standard method exists for solving (3.3). As such a method of solution has to be chosen in agreement with the particular expressions for J and F . In the majority of cases, and even always when F is non-linear, an iterative algorithm has to be carried out.

The second formalism is called strong constrained where the model has to be exactly verified. In consequence we have to deal with a problem of optimization under constraint and the approach by optimal control permits to alleviate to some extent the difficulties linked to this formalism.

4 Optimal Control Techniques

4.1 General Results

Optimal control methods for distributed systems have been extensively studied and applied in many areas such as mechanics, economics, engineering, oceanography, etc.

Due to the fact that the formalism of optimal control problems includes the minimization of a functional, the cost function, they are variational methods and as such their numerical solution requires the computation of the gradient of the cost functional with respect to the state variable.

In many cases, the cost function is only an implicit function of the state variable which may be an initial condition or a boundary condition. Therefore, more sophisticated mathematical techniques must be used for estimating the gradient. One such particular method, the adjoint model technique, was specially developed for this purpose. A difficulty of this approach is the necessity to write well-posed problems and to carefully specify the functional framework of the variational problem.

We assume that the state of the atmosphere is described by a variable X belonging to some Hilbert space \mathcal{H} (of finite or infinite dimension) and by a model written as

$$F(X) = 0 \quad (4.1)$$

We suppose that X may be split into two parts, Y and U , each part belonging to the Hilbert spaces \mathcal{Y} and \mathcal{U} , respectively.

Therefore, (4.1) may be written as

$$F(Y, U) = 0 \quad (4.2)$$

where U is the control variable, chosen in such a way that for each given U , equation (4.2) has a unique solution $Y(U)$.

In this way we may define G by

$$G : \mathcal{Y} \rightarrow \mathcal{U} \quad (4.3)$$

for each U belonging to \mathcal{U} . Then

$$G(Y) = U \quad (4.4)$$

has a unique solution in \mathcal{Y} .

Furthermore, we will assume that for each Y belonging to \mathcal{Y} , $\frac{\partial F}{\partial Y}(Y)$ is an isomorphism from \mathcal{Y} to \mathcal{U} .

Therefore, it is possible to define an inverse function Φ such that:

$$\begin{aligned}\Phi : \mathcal{U} &\rightarrow \mathcal{Y} \\ U &\rightarrow \Phi(U) = Y\end{aligned}$$

verifying:

$$\begin{aligned}\Phi(G(Y)) &= Y \\ \Phi'(U) &= \left[\frac{\partial F}{\partial Y}(\Phi(U)) \right]^{-1}\end{aligned}$$

Another Hilbert space has to be defined: the space of observations Θ in which an observation Z_{obs} is given. As pointed out, the observation is not necessarily a physical one, and it is not supposed to verify the equations of the model.

Let C be an operator from the space of the state variable to the space of observations; for each value of the control U we associate a state of the atmosphere $Y(U)$ and a model observation

$$Z(U) = C(Y(U)). \quad (4.5)$$

The cost function $J(U)$ is a measure of the distance between the model observation associated to the control U and the observation. It is defined by:

$$J(U) = \frac{1}{2} \|C(Y(U)) - Z_{obs}\|_{\Theta}^2 \quad (4.6)$$

Therefore, the problem is to determine the optimal control variable U^* defined by

$$U^* = \arg(\min J(U) \mid u \in \mathcal{U}). \quad (4.7)$$

From a theoretical viewpoint, the system of optimality giving U^* is dependent upon the gradient of J with respect to U .

From a numerical viewpoint, U^* may be estimated by an iterative method starting from a first given U_0 . In the same way, the numerical implementation of the iterative method requires the computation of the gradient of J with respect to U .

For deriving the gradient, a systematic method is the following:

- Let V be some variable belonging to \mathcal{U} ; then the directional derivative of J in direction V will verify:

$$\begin{aligned}J'(U, V) &= \nabla J(U) \cdot V = (C'(Y) \cdot V, C(Y) - Z_{obs})_{\Theta} \\ &= \langle C'(Y)V, \Lambda_{\Theta}(C(Y) - Z_{obs}) \rangle_{\Theta', \Theta}\end{aligned} \quad (4.8)$$

where Λ_{Θ} is the canonical isomorphism between Θ and its dual space Θ' , and $\langle \cdot, \cdot \rangle$ denotes the duality between Hilbert spaces.

- Let R be a linear operator from \mathcal{Y} to \mathcal{U} , we define its dual operator to be the operator R^* from \mathcal{U}' to \mathcal{Y}' defined by:

$$\langle R \cdot \mathcal{Y}, U' \rangle_{\mathcal{U}} = \langle Y, R^* \cdot U' \rangle_{\mathcal{Y}}$$

Using the dual operator of C' in (4.8) gives:

$$\nabla J(U) \cdot V = \langle V, C'(Y)^* \Lambda_{\Theta} (C(Y) - Z_{obs}) \rangle_{\mathcal{U}, \mathcal{U}'}$$

- Let us now define the adjoint system by:

$$\left(\frac{\partial F}{\partial Y} \right)^* P = -C'(Y)^* \Lambda_{\Theta} (C(Y(U)) - Z_{obs}) \quad (4.9)$$

Then:

$$\begin{aligned} \nabla J(U) \cdot V &= \langle V, \left(\frac{\partial F}{\partial Y} \right)^* \cdot P \rangle_{\mathcal{U}, \mathcal{U}'} \\ \nabla J(Y) \cdot V &= \langle \frac{\partial F}{\partial Y} \cdot V, P \rangle_{\mathcal{Y}, \mathcal{Y}'} \end{aligned} \quad (4.10)$$

J is a functional defined on the space \mathcal{U} , so its gradient belongs to the dual space \mathcal{U}' . Theoretically, it is always possible to identify a Hilbert space to its dual. However, in practical problems there exist inclusion relations between the spaces used here, and when a space has been identified to its dual, it is no longer possible to identify subspaces with their duals.

In the practical phase of optimal control methods we were always operating in finite-dimensional spaces where no such problems exist.

Therefore Eq. (4.10) permits us to compute the gradient of J , applied to the direction V by determining P , the adjoint variable, as the solution of the adjoint system (4.9).

From this abstract situation let us extract two more practical examples enabling us to see how the gradient is computed. For an initial condition problem we will consider the case where the control variable is the initial condition, while for a boundary value problem we will see how to compute the gradient when the control variable is the value on the boundary.

4.2 Control of the Initial Condition

After a spatial discretization, we will assume that the state of the atmosphere, modeled by a vector Θ is verifying for the time interval $[0, T]$ the equation:

$$\frac{d\Theta(t)}{dt} = H(\Theta(t)) \quad (4.11)$$

where $\Theta(t)$ belongs to a finite dimensional space.

With an initial condition $\Theta(0) = \mu$, Eq. (4.11) has a unique solution $\Theta(\mu, t)$.

For the sake of simplicity, we will assume that a continuous observation $\tilde{\Theta}$, in time, is given on the time interval $[0, T]$. The distance between a solution of (4.11) and the observation is defined by

$$J(\mu) = \frac{1}{2} \int_0^T \left\| \Theta(\mu, t) - \tilde{\Theta}(t) \right\|^2 dt \quad (4.12)$$

where $\| \cdot \|$ is the Euclidean norm in finite dimensional space. With respect to the general theory developed above the space of the state variable is the same as the space of the observations. In practice, the observations are pointwise in both space and in time, therefore, Dirac's measures have to be introduced in the definition of J .

The derivation of the gradient of J with respect to μ is obtained as follows:

Let v be some element belonging to the space of the initial conditions. The directional derivative of Θ in direction v is defined by:

$$\hat{\Theta}(\mu, v) = \lim_{\alpha \rightarrow 0} \frac{\Theta[(\mu + \alpha), t] - \Theta(\mu, t)}{\alpha} \quad (4.13)$$

where $\hat{\Theta}(\mu, v)$ is the solution of the differential system:

$$\begin{aligned} \frac{d\hat{\Theta}(\mu, v)}{dt} &= \frac{\partial H}{\partial \Theta}[\Theta(\mu, t)] \cdot \hat{\Theta}(\mu, v) \\ \hat{\Theta}(0) &= v \end{aligned} \quad (4.14)$$

obtained by writing (4.11) with initial condition μ , then with initial condition $\mu_\alpha v$ and by letting the scalar α tend to zero. In (4.14) the expression $\frac{\partial H}{\partial \Theta}$ denotes the Jacobian of H .

The directional derivative of J in direction v is obtained by taking the derivative of (4.12) leading to:

$$J'(\mu, v) = \int_0^T \left(\hat{\Theta}(\mu, v, t), \Theta(\mu, t) - \tilde{\Theta}(t) \right) dt \quad (4.15)$$

Let ψ be the dual variable to Θ , ψ is defined as the solution of the adjoint system to (4.11) given by:

$$\begin{aligned} \frac{d\psi}{dt}(\mu, t) + \left[\frac{\partial H}{\partial \Theta} \Theta(\mu, t) \right]^T \cdot \psi(\mu, t) &= \left(\Theta(\mu, t) - \tilde{\Theta}(t) \right) \\ \psi(T) &= 0 \end{aligned} \quad (4.16)$$

Let us write the scalar product of (4.15) with $\hat{\Theta}$, then by integrating from 0 to T , we obtain:

$$J'(\mu, v) = \int_0^T \left(\frac{d\psi}{dt} + \left[\frac{\partial H}{\partial \Theta} \Theta(\mu, t) \right]^T \cdot \psi(\mu, t), \hat{\Theta}(\mu, v, t) \right) dt$$

The time derivative in (4.16) is integrated by parts and then by using (4.14) we obtain:

$$J'(\mu, v) = \nabla J(\mu) \cdot v = \psi(\mu, 0) \cdot v \quad (4.17)$$

Therefore, the gradient of J is obtained as the value at time zero of the dual variable. The backward integration of the adjoint system from T to 0 permits us to estimate the gradient of the cost functional and to perform a descent-type method.

An important remark for potential applications of control methods is the fact that with a different cost function only the right hand side of (4.16) has to be changed. The main difficulty encountered for programming optimal control methods is to write the left hand side of (4.16). This one is independent of the cost function and is intrinsic for a given model. Once it has been written and derived it can be used for other purposes such as data assimilation, initialization, sensitivity analysis, etc.

4.3 Control of the Boundary

For the sake of simplicity, we will suppose that on a domain Ω , of boundary Γ , some field is verifying the Laplace equation

$$\Delta U = f. \quad (4.18)$$

Together with a boundary condition $U/\Gamma = V$, Laplace equation (4.18) has a unique solution, $U(V)$.

Let \mathcal{T} be a set of points belonging to Ω , where some observations \tilde{U} of U are performed.

$$\mathcal{T} = \{Z_1, Z_2, \dots, Z_N\}$$

The cost function is defined by

$$J(V) = \frac{1}{2} \sum_{i=1}^N \left(U(V, Z_i) - \tilde{U}(Z_i) \right)^2, \quad (4.19)$$

while the directional derivative \overline{U} of U in a direction H is the solution of

$$\begin{aligned} \Delta \overline{U}(H) &= 0 \\ \overline{U}(H)/\Gamma &= H. \end{aligned} \quad (4.20)$$

The directional derivative of J verifies

$$J'(V, H) = \sum_{i=1}^N \left(\bar{U}(Z_i), U(V, Z_i) - \tilde{U}(Z_i) \right). \quad (4.21)$$

The adjoint system to (4.19) is introduced with P the dual variable to U .

$$\begin{aligned} \Delta P &= \sum_{i=1}^N U(V, Z_i) - \tilde{U}(Z_i) \\ P/\Gamma &= 0 \end{aligned} \quad (4.22)$$

As above, (4.22) is multiplied by $\bar{U}(H, Z_i)$ integrated on Ω , and after an integration by parts we find

$$\nabla J(V) = \frac{\partial P}{\partial n}/\Gamma, \quad (4.23)$$

$\frac{\partial P}{\partial n}$ being the normal derivative of P on the boundary Γ . The estimation of the gradient for carrying out a descent-method requires the estimation of the gradient of J , which is obtained by solving the adjoint system (4.22).

Let us point out that this case is especially simple due to the fact that the Laplacian operator is self-adjoint. Therefore, a Laplace's equation solver may be used to solve both the direct and the adjoint problem.

This problem could have been solved using a classical variational formalism, for instance with a weak constraint formalism we would have to minimize the functional

$$J(U) = \frac{1}{2} \sum \left(U(Z_i) - \tilde{U}(Z_i) \right)^2 + \frac{1}{C} \int_{\Omega} (\Delta U - f)^2 dy. \quad (4.24)$$

The Euler-Lagrange equation for (4.24) is a fourth order partial differential equation with complicated boundary conditions. From a numerical viewpoint the size of the discrete problem associated with (4.24) is equal to the number of grid points in the discrete point of view domain Ω .

By comparison, for the optimal control approach the dimension of the problem to be solved is only equal to the number of points on the discrete boundary. In this way we have obtained a significant reduction of the size of the problem.

5 Weak Constraints in Variational Data Assimilation

The canonical approach for variational data assimilation (VDA), based on Optimal Control, implicitly assumes that the model is without error. This is not true because of the physical errors due to approximation in the physics of the problem, for instance in the parametrization of non linear interactions and also in the physical processes and the mathematical error due to discretization of the equations and also to iterative processes carried out to solve non linear problems or subproblems.

To alleviate this problem Sasaki has introduced the concept of *weak constraint* permitting to have a model that is only approximately verified.

5.1 Three Basic Methods in Constrained Optimization

Let us consider the constrained optimization problem:

Minimize $J(X)$ subject to the constraint $G(X) = 0$, where X belongs to some space \mathcal{X} , J is a mapping from \mathcal{X} to \mathbb{R} and G is a mapping from \mathcal{X} to some linear space \mathcal{Y}

Here we assume the differentiability of J and G . There are three basic algorithms to obtain a numerical solution to this problem:

5.1.1 Duality Methods

In this method we introduce a Lagrange multiplier Λ in the dual space \mathcal{Y} of and the Lagrangian \mathcal{L} defined by:

$$\mathcal{L}(X, \Lambda) = J(X) + (\Lambda, G(X)) \quad (5.1)$$

Then optimal solution of the constrained optimization problem is a saddle point of the Lagrangian and is characterized by:

$$\frac{\partial \mathcal{L}}{\partial X} = \nabla J + \left[\frac{\partial G}{\partial X} \right]^t \cdot \Lambda = 0 \quad (5.2)$$

$$\frac{\partial \mathcal{L}}{\partial \Lambda} = G(X) = 0 \quad (5.3)$$

In Sasaki's terminology this is the strong constraint formalism, it is worthwhile to point out that X is the state variable of the problem, Sasaki doesn't make the difference between state variable and control variables that could be the initial condition and/or boundary conditions. The equations above are the Optimality System (O.S.) it can be solved by an iterative algorithm of the form:

$$X_{n+1} = X_n - \rho_n \cdot D_n \quad (5.4)$$

$$\Lambda_{n+1} = \Lambda_n + \eta_n \cdot W_n \quad (5.5)$$

where D_n is a direction of descent, estimated from the gradient of J , X_{n+1} realizes the minimum of \mathcal{L} along this direction. On the same token W_n is a direction of ascent and Λ_{n+1} realizes the maximum of \mathcal{L} along this direction. ρ_n and η_n are scalars. In practice

some stopping criterion for the iterative algorithm has to be defined and therefore, at the end of the process, the constraint is not exactly satisfied and, by this way, an error is introduced. This error cannot be controlled. The Lagrange multiplier introduced in this method is nothing else than the adjoint variable used in the terminology of VDA problems stated as problems of Optimal Control. In practice the convergence of this type of algorithms is slow.

5.1.2 Penalty Methods

In this approach we define a *penalized functional* J_ϵ by:

$$J_\epsilon(X) = J(X) + \frac{1}{\epsilon} \|G(X)\|^2. \quad (5.6)$$

X_ϵ is the minimizer of J_ϵ , when $\epsilon -> 0$ then $X_\epsilon -> X^*$ solution of the original constrained optimization problem.

X_ϵ is the solution of the equation:

$$\nabla J_\epsilon(X) = \nabla J(X) + \frac{2}{\epsilon} \left[\frac{\partial G}{\partial X} \right]^t \cdot G(X) = 0. \quad (5.7)$$

As above, the minimization of the penalized functional is solved by an iterative algorithms of descent type. A major inconvenient of this method is to become quickly ill conditioned when ϵ is small. This is the basic *weak constraint formalism* of Sasaki with the difference that ϵ is fixed and doesn't change with the iterations. A consequence is that the constraint is not exactly verified, the choice of ϵ could permit some control on the amplitude of the error on the constraint.

5.1.3 Augmented Lagrangian Methods

This algorithm is a combination of duality and penalization, it is defined by an Augmented Lagrangian:

$$\mathcal{L}_\epsilon(X, \Lambda) = J(X) + (\Lambda, G(X)) + \frac{1}{\epsilon} \|G(X)\|^2. \quad (5.8)$$

$(X_\epsilon, \Lambda_\epsilon)$, saddle point of the Augmented Lagrangian is a solution of the constrained optimization problem. It is evaluated by a descent-ascent iterative method. The penalty term added to the Lagrangian can be considered as a regularization term in the sense of Tykhonov and make the problem well conditioned. This is exactly the sense of the background term in the usual terminology of VDA.

5.1.4 Remarks

For dynamical models if there is no difference between state variable and control variable then all the evolution of the model as to be considered as the state variable and we have to deal with huge numerical problems. At the present time, for operational models, the size of the variable, at a given time is of the order of 1 billion, therefore if we want to carry out an analysis on 1000 time step, the dimension of the variable will be of the order of 10^{12} , this is out of the scope of numerical optimization.

The *weak constraint formalism* permits to consider some error on the model but the algorithms doesn't permit neither to evaluate this error nor to identify its source. In the next sections we will see how to alleviate this inconvenient.

5.2 Direct Control of the Error in VDA

5.2.1 General Formalism

Let's go back to the general formalism of VDA with a dynamical model. In this approach we introduce some state error Y in some space \mathcal{Y} as state variable, and the model is now written as:

$$\frac{dX}{dt} = F(X) + \Pi.Y \quad (5.9)$$

$$X_0 = U. \quad (5.10)$$

Π is a linear operator from \mathcal{X} to \mathcal{Y} , a priori Y depends on time but it can be steady state. The cost function, we want to minimize with respect to U and Y is defined by:

$$J(U, Y) = \frac{1}{2} \int_0^T \|H[X(U, Y, t)] - X_{obs}(t)\|^2 dt + \frac{1}{2} \|U - U_0\|^2 + \frac{1}{2} \|Y\|^2, \quad (5.11)$$

The last term in the definition of the cost function is to have the error Y as small as possible, while H is the linear observation operator. In order to simplify notations the covariances errors are the identity. Using more complex covariances is straightforward.

5.2.2 Optimality System

As usual, we introduce two directions to compute the directional derivatives. The gradient of J has two components:

$$\nabla J = \begin{pmatrix} \nabla_U J \\ \nabla_Y J \end{pmatrix}.$$

We introduce the adjoint variable P , the solution of the following system:

$$\begin{aligned} \frac{dP}{dt} &= \left[\frac{\partial P}{\partial X} \right]^T \cdot P + H^t(HX - X_{obs}) \\ P(T) &= 0. \end{aligned} \quad (5.12)$$

Then we obtain:

$$\nabla J = \begin{pmatrix} -P(0) + U - U_0 \\ \Pi^t P + Y \end{pmatrix}. \quad (5.13)$$

For practical purposes Y has to be located in a space with a dimension comparable to the dimension of the initial condition U , if the dimension of \mathcal{Y} were too large then the problem would become numerically intractable. A way to alleviate this difficulty is to discretize the error spaces of test functions, we will write Y under the form:

$$Y = \sum_{i=1}^n \sum_{j=1}^m k_{ij} \phi_i \psi_j, \quad (5.14)$$

where k_{ij} are the elements of matrix K , Φ is a time dependent vector with elements ϕ_i and Ψ represents a vector of steady state elements ψ_i . Therefore we have $Y = \Phi^t K \Psi$ and the model becomes:

$$\frac{dX}{dt} = F(X) + \Pi \cdot \Phi^t K \Psi \quad (5.15)$$

$$X_0 = U. \quad (5.16)$$

With a cost function:

$$J(U, K) = \frac{1}{2} \int_0^T \|H[X(U, K, t)] - X_{obs}(t)\|^2 dt + \frac{1}{2} \|U - U_0\|^2 + \frac{1}{2} \|K\|^2, \quad (5.17)$$

The adjoint model is the same than above but the second component of the gradient i.e. the gradient with respect to K becomes:

$$\nabla_U J = \int_0^T \Phi \Pi^t P \Psi dt + K. \quad (5.18)$$

5.2.3 Control of the Error of Observation

By the same token we can consider an error of observation and try to identify it, therefore we introduce a control variable Z belonging to space of observation. The model remains the same only the cost function is going to change and becomes:

$$J(U, Y) = \frac{1}{2} \int_0^T \|H[X(U, Y, t)] - X_{obs}(t) - Z(t)\|^2 dt + \frac{1}{2} \|U - U_0\|^2 + \frac{1}{2} \|Z\|^2, \quad (5.19)$$

It's easy to see that the adjoint model becomes:

$$\begin{aligned} \frac{dP}{dt} &= \left[\frac{\partial P}{\partial X} \right]^t \cdot P + H^t(HX - X_{obs} - Z) \\ P(T) &= 0. \end{aligned} \quad (5.20)$$

and the gradient has a component with respect to U which is unchanged and we also have a component of the gradient with respect to Z which is:

$$\nabla_Z J = (2Z + X_{obs} - HX) \quad (5.21)$$

As we did above the error of observation can be discretized in an adequate base in order to reduce the dimension of the system. In theory both observation errors and model errors could be jointly controlled.

5.3 Weak Constraint: Control of Systematic Error

The error of the model could be a random error and/or a systematic error. To identify a systematic error an additional term can be included in the model written as:

$$\frac{dX}{dt} = F(X) + E(K, t) \quad (5.22)$$

$$X_0 = U. \quad (5.23)$$

where K is a low order parameter we want to identify. We assume that the error is governed by the system:

$$\frac{dE}{dt} = G(E, K) \quad (5.24)$$

$$E_0 = V. \quad (5.25)$$

In this case the assimilation will be the determination of (U^*, V^*, K^*) minimizing the cost function defined by:

$$J(U, V, K) = \frac{1}{2} \int_0^T \|H[X(U, V, t)] - X_{obs}(t)\|^2 dt \quad (5.26)$$

$$+ \frac{1}{2} \int_0^T \|E\|^2 dt + \frac{1}{2} \|U - U_0\|^2 + \frac{1}{2} \|K\|^2, \quad (5.27)$$

To get the Optimality System we need to introduce two adjoints variables P and Q as the solution of:

$$\frac{dP}{dt} = \left[\frac{\partial P}{\partial X} \right]^t + H^t(HX - X_{obs} - Z) \quad (5.28)$$

$$P(T) = 0 \quad (5.29)$$

$$\frac{dQ}{dt} = \left[\frac{\partial Q}{\partial E} \right]^t \cdot Q + E \quad (5.30)$$

$$Q(T) = 0 \quad (5.31)$$

Thanks to a backward integration of the adjoint model we get the three components of the gradient $\nabla J(U, V, K)$:

$$\nabla_U J = (U - U_0) - P(0) \quad (5.32)$$

$$\nabla_V J = -Q(0) + E \quad (5.33)$$

$$\nabla_K J = \left[\frac{\partial G}{\partial K} \right]^t \cdot Q + \frac{\partial G}{\partial K} \quad (5.34)$$

The control of both errors of model and of data can be carried out simultaneously if a model of observation error were added.

5.4 Example: Saint-Venant's Equations

Saint-Venant's equations, also known as shallow water equations, are used for an incompressible fluid for which the depth is small with respect to the horizontal dimensions. General equations of geophysical fluid dynamics are vertically integrated using the hydrostatic hypothesis, therefore vertical acceleration is neglected. In Cartesian coordinates they are:

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} + v \frac{\partial u}{\partial y} - fv + \frac{\partial \phi}{\partial x} = 0 \quad (5.35)$$

$$\frac{\partial v}{\partial t} + u \frac{\partial v}{\partial x} + v \frac{\partial v}{\partial y} + fu + \frac{\partial \phi}{\partial y} = 0 \quad (5.36)$$

$$\frac{\partial \phi}{\partial t} + \frac{\partial u \phi}{\partial x} + \frac{\partial v \phi}{\partial y} = 0 \quad (5.37)$$

In this system $X = (u, v, \phi)^T$ is the state variable, u and v are the components of the horizontal velocity; ϕ is the geopotential (proportional to the height of the free surface) and f the Coriolis' parameter. For sake of simplicity, the following hypothesis are used:

(a) The error of the model is neglected. Only the initial condition will be considered as a control variable.

(b) Lateral boundary conditions are periodic. This is verified in global models.

(c) Observations are supposed to be continuous with respect to time. Of course this is not the case in practice. The observation operators are taken identity matrices. If $X_0 = (u_0, v_0, \phi_0)^T$ is the initial condition and the cost function is given by:

$$J(X_0) = \frac{1}{2} \int_0^T [\|u - u_{obs}\|^2 + \|v - v_{obs}\|^2 + \gamma \|\phi - \phi_{obs}\|^2] dt, \quad (5.38)$$

where γ is a weight function, then the directional derivatives $\bar{X} = (\bar{u}, \bar{v}, \bar{\phi})^T$ in the direction $h = (h_u, h_v, h_\phi)^T$ (in the control space) will be solutions of the linear tangent model:

$$\frac{\partial \bar{u}}{\partial t} + u \frac{\partial \bar{u}}{\partial x} + \bar{u} \frac{\partial u}{\partial x} + v \frac{\partial \bar{u}}{\partial y} + \bar{v} \frac{\partial u}{\partial y} - f \bar{v} + \frac{\partial \bar{\phi}}{\partial x} = 0 \quad (5.39)$$

$$\frac{\partial \bar{v}}{\partial t} + u \frac{\partial \bar{v}}{\partial x} + \bar{u} \frac{\partial v}{\partial x} + v \frac{\partial \bar{v}}{\partial y} + \bar{v} \frac{\partial v}{\partial y} + f \bar{u} + \frac{\partial \bar{\phi}}{\partial y} = 0 \quad (5.40)$$

$$\frac{\partial \bar{\phi}}{\partial t} + \frac{\partial \bar{u} \phi}{\partial x} + \frac{\partial u \bar{\phi}}{\partial x} + \frac{\partial \bar{v} \phi}{\partial y} + \frac{\partial v \bar{\phi}}{\partial y} = 0 \quad (5.41)$$

Introducing three adjoint variables p, q, φ we can compute the adjoint system, it writes:

$$\frac{\partial p}{\partial t} + \frac{\partial p u}{\partial x} + v \frac{\partial p}{\partial y} - q \frac{\partial v}{\partial x} - f q + \frac{\partial \varphi \phi}{\partial x} = u_{obs} - u, \quad (5.42)$$

$$\frac{\partial q}{\partial t} - p \frac{\partial u}{\partial y} + u \frac{\partial q}{\partial x} + \frac{\partial q v}{\partial y} + f p + \frac{\partial \varphi \phi}{\partial y} = v_{obs} - v, \quad (5.43)$$

$$\frac{\partial \varphi}{\partial t} + \frac{\partial u \varphi}{\partial x} + \frac{\partial v \varphi}{\partial y} + \frac{\partial p}{\partial x} + \frac{\partial q}{\partial y} = \gamma(\phi_{obs} - \phi). \quad (5.44)$$

5.4.1 Control of Errors with Dynamics: An Example

Now we consider the Burgers equation with homogeneous boundary conditions and the state variable is u defined on $[0, 1] \times [0, T]$ with an error $E(x, t)$ governed by a diffusion equation with an unknown parameter γ . The unknown initial conditions on u and E to be found are α and β . The model and the error are characterized by:

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} - \frac{\partial}{\partial x} \left(\mu \frac{\partial u}{\partial x} \right) + E = F, \quad (5.45)$$

$$u(0) = \alpha. \quad (5.46)$$

$$\frac{\partial E}{\partial t} - \frac{\partial}{\partial x} \left(\gamma \frac{\partial E}{\partial x} \right) = 0, \quad (5.47)$$

$$E(0) = \beta. \quad (5.48)$$

The problem is to determine $\alpha^*, \beta^*, \gamma^*$ minimizing the cost function J defined by:

$$J(\alpha, \beta, \gamma) = \frac{1}{2} \int_0^T \int_0^1 (u - u_{obs})^2 dx dt + \frac{1}{2} \int_0^T \int_0^1 E^2 dx dt + \frac{1}{2} \int_0^1 \gamma^2 dx + \frac{1}{2} \alpha^2 + \frac{1}{2} \beta^2. \quad (5.49)$$

For sake of simplicity we have chosen the simplest form with a complete observation, identity observation operators and ignoring the statistical information. The forcing term is denoted by F and parameter μ is known. As usual the next step is to derive the Gâteaux derivatives u and E in some directions in the spaces of control variables then introducing p and q two adjoint variables as the solution of the equations:

$$\frac{\partial p}{\partial t} + u \frac{\partial p}{\partial t} + \frac{\partial}{\partial x} \left(\mu \frac{\partial p}{\partial x} \right) = u_{obs} - u \quad (5.50)$$

$$p(T) = 0 \quad (5.51)$$

$$\frac{\partial q}{\partial t} + \frac{\partial}{\partial x} \left(\gamma \frac{\partial q}{\partial x} \right) + E + p = 0 \quad (5.52)$$

$$q(T) = 0. \quad (5.53)$$

Then the components of the gradient ∇J verify:

$$\nabla J_\alpha = -p(0) + \int_0^1 (u - u_{obs}) dx \quad (5.54)$$

$$\nabla J_\beta = -q(0) + \beta \quad (5.55)$$

$$\nabla J_\gamma = \int_0^T \left(\frac{\partial E}{\partial x} \frac{\partial q}{\partial x} + q \frac{\partial^2 E}{\partial x^2} \right) dx \quad (5.56)$$

6 Second Order Methods

The optimality system, the Euler-Lagrange equation, provides only a necessary condition for optimality. In the linear case, the solution is unique if the Hessian is positive definite. From a general point of view the information given by the Hessian is important for theoretical, numerical and practical issues. For operational models it is impossible to compute the Hessian itself, as it is a square matrix with around 10^{18} terms, nevertheless the most important information can be extracted from the spectrum of the Hessian which can be estimated without an explicit determination of this matrix. This information is of importance for estimating the condition number of the Hessian for preparing an efficient preconditioning.

A general method to get this information is to apply the techniques described above to the couple made by the direct and adjoint models (Le Dimet et al. 2002), leading to a so called second order adjoint. The following steps are carried out:

- Linearization of the direct and adjoint models with respect to the state variable.
- Introducing second order adjoint variables.
- Transposition to put in light the linear dependence with respect to the directions.

The system obtained, the second order adjoint, is used to compute the product of the Hessian by any vector. Of course if we consider all the vectors of the canonical base then it will be possible to get the complete Hessian.

The determination of this product permits to access some information.

- By using Lanczos type methods and deflation, it is possible to compute the eigenvectors and eigenvalues of the Hessian.
- To carry out second order optimization methods of Newton-type are used for equations of the form:

$$\nabla J(X) = 0$$

The iterations are:

$$X_{n+1} = X_n - H^{-1}(X_n) \cdot \nabla J(X_n)$$

where H is the Hessian of J . At each iteration a linear system should be solved. This is done by carrying out some iterations of a conjugate gradient methods which requires computing the Hessian-vector product.

For the Saint-Venant equations the second order adjoint system is given by

$$\begin{aligned} \frac{\partial \hat{u}}{\partial t} + u \frac{\partial \hat{u}}{\partial x} + v \frac{\partial \hat{v}}{\partial y} + \hat{u} \frac{\partial v}{\partial y} - \hat{v} \frac{\partial v}{\partial y} - f \hat{v} + \phi \frac{\partial \hat{\phi}}{\partial x} \\ = \tilde{v} \frac{\partial \bar{v}}{\partial x} - \bar{u} \frac{\partial \tilde{u}}{\partial x} - \bar{v} \frac{\partial \bar{u}}{\partial y} + \tilde{u} \frac{\partial \bar{v}}{\partial y} - \bar{\phi} \frac{\partial \tilde{\phi}}{\partial x} - \bar{u} \end{aligned} \quad (6.1)$$

$$\begin{aligned} \frac{\partial \hat{v}}{\partial t} + \hat{u} \frac{\partial u}{\partial y} - u \frac{\partial \hat{v}}{\partial x} + \hat{v} \frac{\partial u}{\partial x} + v \frac{\partial \hat{v}}{\partial y} + f \hat{u} + \phi \frac{\partial \hat{\phi}}{\partial y} \\ = \tilde{u} \frac{\partial \bar{u}}{\partial x} - \bar{u} \frac{\partial \tilde{v}}{\partial x} - \tilde{v} \frac{\partial \bar{u}}{\partial y} + \bar{u} \frac{\partial \tilde{v}}{\partial y} - \bar{\phi} \frac{\partial \tilde{\phi}}{\partial y} - \bar{v} \end{aligned} \quad (6.2)$$

$$\frac{\partial \hat{\phi}}{\partial t} + \frac{\partial \hat{u}}{\partial x} + \frac{\partial \hat{v}}{\partial y} + u \frac{\partial \hat{\phi}}{\partial x} + v \frac{\partial \hat{\phi}}{\partial y} = -\bar{u} \frac{\partial \tilde{\phi}}{\partial x} - \bar{v} \frac{\partial \tilde{\phi}}{\partial x} - \gamma \bar{\phi}, \quad (6.3)$$

where $Q = (\hat{u}, \hat{v}, \hat{\phi})^T$ and $R = (\bar{u}, \bar{v}, \bar{\phi})^T$ are the second and first order adjoint variables.

From the formal point of view we see that first and second order differ by second order terms which do not take into account the adjoint variable. The computation of second derivatives require storing both the trajectories of the direct and adjoint models. For very large models it could be more economical to recompute these trajectories.

6.1 Sensitivity analysis

In the environmental sciences the mathematical models contain parameters which cannot be estimated very precisely either because they are difficult to measure or because they represent some subgrid phenomena. Therefore it is important to be able to estimate the impact of uncertainties on the outputs of the model. Sensitivity analysis is defined by:

- X is the state vector of the model, K a vectorial parameter of the model $F(X, K) = 0$.
- $G(X, K)$ is the response function: a real value function
- By definition the sensitivity of the model is the gradient of G with respect to K . The difficulty encountered comes from the implicit dependence of G on K through X , solution of the model.

Several methods can be used to estimate the sensitivity:

- By finite differences. We get:

$$\frac{\partial G}{\partial K_i} \simeq \frac{G(X(K + \alpha e_i), K + \alpha e_i) - G(X(K), K)}{\alpha},$$

where K_i is the i th component of vector K .

The main inconvenience of this method is its computational cost: it requires solving the model as many times as the dimension of the model. Furthermore the determination of the parameter α may be tricky. If it too large, the variation of G could be nonlinear, since small value round off errors may dominate the variation of G . The main advantage of this method is that it is very easy to implement.

- Sensitivity via an adjoint model. Let $F(X, K) = 0$ be the direct model. We introduce its adjoint:

$$\left[\frac{\partial F}{\partial X} \right]^* \cdot P = \frac{\partial G}{\partial X},$$

where $\left[\frac{\partial F}{\partial X} \right]^*$ is the adjoint operator of $\frac{\partial F}{\partial X}$. Then the gradient is given by:

$$\nabla G = \frac{\partial G}{\partial K} - \left[\frac{\partial F}{\partial K} \right]^* \cdot P$$

The advantage of this method is that the sensitivity is obtained by only one run of the adjoint model. The price to be paid is the derivation of the adjoint code.

6.2 Sensitivity in the Presence of Data

In geophysics a usual request is the estimation of the sensitivity with respect to observations. What will be the impact of an uncertainty on the prediction? It is clear that observations are not directly used in the direct model, since they can be found only among the forcing terms in the adjoint model. Therefore to apply the general formalism of sensitivity analysis we should apply it not to the model itself but to the optimality system, i.e. the model plus the adjoint model. A very simple example with a scalar ordinary differential equation is given in Le Dimet et al. (1997) showing that the direct model is not sufficient to carry out sensitivity analysis in the presence of data. Deriving the optimality system will introduce second order derivatives.

Consider a model of the form $F(X, I) = 0$ where I stands for the input of the model. Assuming F is a steady state or time dependent operator, $J(X, I)$ is the cost function in the variational data assimilation and P is the adjoint variable, then the optimality system is:

$$\begin{cases} F(X, I) = 0 \\ \left[\frac{\partial F}{\partial X} \right]^* \cdot P - \frac{\partial J}{\partial X} = 0 \\ \left[\frac{\partial F}{\partial I} \right]^* \cdot P - \frac{\partial J}{\partial I} = 0 \end{cases} \quad (6.4)$$

The optimality system can be considered as a generalized model: \mathcal{F} with the state variable $Z = \begin{pmatrix} X \\ P \end{pmatrix}$.

If i is a perturbation on I , we get:

$$\frac{\partial F}{\partial X} \cdot \hat{X} + \frac{\partial F}{\partial I} \cdot i = 0 \quad (6.5)$$

$$\left[\frac{\partial^2 F}{\partial X^2} \cdot \hat{X} + \frac{\partial^2 F}{\partial X \partial I} \cdot i \right]^* \cdot P + \left[\frac{\partial F}{\partial X} \right]^* \cdot \hat{P} - \frac{\partial^2 J}{\partial X^2} \cdot \hat{X} - \frac{\partial^2 J}{\partial X \partial I} \cdot i = 0 \quad (6.6)$$

$$\left[\frac{\partial^2 F}{\partial I^2} \cdot i + \frac{\partial^2 F}{\partial X \partial I} \cdot \hat{X} \right]^* \cdot P + \left[\frac{\partial F}{\partial I} \right]^* \cdot \hat{P} - \frac{\partial^2 J}{\partial I^2} \cdot i - \frac{\partial^2 J}{\partial I \partial X} \cdot \hat{X} = 0 \quad (6.7)$$

$$\hat{G}(X, I, i) = \frac{\partial G}{\partial X} \cdot \hat{X} + \frac{\partial G}{\partial I} \cdot i, \quad (6.8)$$

where \hat{X} and \hat{P} are the Gâteaux derivative of X and P in the direction i . Let us introduce the second order adjoint variables Q and R . Here G is the response function introduced in Sect. 6.1 and we are looking to estimate the gradient of G with respect to I . By taking the inner product of (6.5) and (6.6) by Q , and of (6.7) by R , and adding it, we obtain:

$$\begin{aligned}
& < \hat{X}, \left[\frac{\partial F}{\partial X} \right]^* \cdot Q + \left[\frac{\partial^2 F}{\partial X^2} \cdot Q \right]^* \cdot P - \left[\frac{\partial^2 J}{\partial X^2} \right]^* \cdot Q - \left[\frac{\partial^2 J}{\partial I \partial X} \right]^* \cdot R + \left[\frac{\partial^2 F}{\partial X \partial I} \cdot P \right]^* \cdot R > \\
& + < \hat{P}, \left[\frac{\partial F}{\partial X} \right] \cdot Q + \left[\frac{\partial F}{\partial I} \right] \cdot R > + \\
& < i, \left[\frac{\partial F}{\partial I} \right]^* \cdot Q + \left[\frac{\partial^2 F}{\partial X \partial I} \cdot Q \right]^* \cdot P - \left[\frac{\partial^2 J}{\partial X \partial I} \right]^* \cdot Q - \left[\frac{\partial^2 J}{\partial I^2} \right]^* \cdot R + \left[\frac{\partial^2 F}{\partial I^2} \cdot R \right]^* \cdot P > = 0
\end{aligned}$$

Identifying in (6.8) it is seen that if Q and R are defined as solutions of:

$$\begin{cases} \left[\frac{\partial F}{\partial X} \right]^* \cdot Q + \left[\frac{\partial^2 F}{\partial X^2} \cdot Q \right]^* \cdot P - \left[\frac{\partial^2 J}{\partial X^2} \right]^* \cdot Q - \left[\frac{\partial^2 J}{\partial I \partial X} \right]^* \cdot R + \left[\frac{\partial^2 F}{\partial X \partial I} \cdot P \right]^* \cdot R = \frac{\partial \mathbf{G}}{\partial X} \\ \left[\frac{\partial F}{\partial X} \right] \cdot Q + \left[\frac{\partial F}{\partial I} \right] \cdot R = 0 \end{cases} \quad (6.9)$$

Then we get the gradient of G with respect to I (the sensitivity) by:

$$S = \frac{d\mathbf{G}}{dI} = - \left[\frac{\partial F}{\partial I} \right]^* \cdot Q - \left[\frac{\partial^2 F}{\partial X \partial I} \cdot Q \right]^* \cdot P + \left[\frac{\partial^2 J}{\partial X \partial I} \right]^* \cdot Q + \left[\frac{\partial^2 J}{\partial I^2} \right]^* \cdot R - \left[\frac{\partial^2 F}{\partial I^2} \cdot R \right]^* \cdot P + \frac{\partial \mathbf{G}}{\partial I} \quad (6.10)$$

To summarize the algorithm is the following:

- i. Solve the optimality system (6.4) to get X and P
- ii. Solve the coupled system (6.9) to compute Q and R
- iii. Then the sensitivity is given by (6.10).

7 Sensitivity with Respect to Sources

7.1 Identification of the Fields

Let us assume that the flow described by the state variable X satisfies the following time dependent differential system between time 0 and final time T :

$$\begin{cases} \frac{dX}{dt} = F(X), \\ X(0) = U. \end{cases} \quad (7.1)$$

The pollutant, considered as a passive tracer, is described by its concentration whose evolution is given by the following equations:

$$\begin{cases} \frac{dC}{dt} = G(X, C, S), \\ C(0) = V, \end{cases} \quad (7.2)$$

where C is the pollutant's concentration and S is a function of space and time and represents the production of pollutant.

The first task is to retrieve the fields from observations $X_{obs} \in \mathcal{X}_{obs}$ corresponding to the state variable X and $C_{obs} \in \mathcal{C}_{obs}$ associated with the concentration C of the pollutant. We introduce a cost function J defined by

$$J(U, V) = \frac{1}{2} \int_0^T \|EX - X_{obs}\|^2 dt + \frac{1}{2} \int_0^T \|DC - C_{obs}\|^2 dt \quad (7.3)$$

where E is an operator from the space of the state variable toward the space of observations and D from the space of concentration toward the space of observations of concentration. For sake of simplicity, we do not introduce regularization terms in the cost function. In practice they are of crucial importance. For retrieving the state variable and the concentration, we have to determine U^* and V^* which minimize J . They are solutions of the Optimality System. If P and Q are defined as the solutions of the following adjoint models:

$$\begin{cases} \frac{dP}{dt} + \left[\frac{\partial F}{\partial X} \right]^* \cdot P + \left[\frac{\partial G}{\partial X} \right]^* \cdot Q = E^*(EX - X_{obs}); \\ P(T) = 0; \end{cases} \quad (7.4)$$

$$\begin{cases} \frac{dQ}{dt} + \left[\frac{\partial G}{\partial C} \right]^* \cdot Q = D^*(DC - C_{obs}); \\ Q(T) = 0, \end{cases} \quad (7.5)$$

then from the backward integration of the system we get the gradient:

$$\nabla J_U = -P(0) \quad (7.6)$$

$$\nabla J_V = -Q(0) \quad (7.7)$$

7.2 Formulation of the Sensitivity Problem

Let Ω_A , a sub-domain of the physical space be the region of interest (response region) and φ the function giving the measure of the effect of interest. By effect of interest, we mean the “effect of the pollutant” and we want to evaluate the sensitivity of φ with respect to the source. Our quantity of interest φ is a function of the concentration C and of the source S . We define the response function as:

$$\Phi_A(C, S) = \int_0^T \int_{\Omega_A} \varphi(C, S) dx. \quad (7.8)$$

In the simplest case, φ can be defined as $\varphi = C$, in which case Φ_A is the mean over the space and the time of the concentration of the pollutant. By definition, the sensitivity with respect to the source S is the gradient of the response function Φ_A with

respect to S . Following the guidelines of the derivation of the gradient as presented in Sect. 7.1, we obtain the tangent linear and adjoint models. Their solutions are \hat{X} , \hat{C} , \hat{P} and \hat{Q} , the Gâteaux derivatives with respect to S (in the direction s) of the variables of the optimality system given by Eqs. (7.1), (7.2), (7.4) and (7.5). The models are elaborated again below:

$$\begin{cases} \frac{d\hat{X}}{dt} = \frac{\partial F}{\partial X} \cdot \hat{X} \\ \hat{X}(0) = \hat{U} \end{cases} \quad (7.9)$$

$$\begin{cases} \frac{d\hat{C}}{dt} = \frac{\partial G}{\partial X} \cdot \hat{X} + \frac{\partial G}{\partial C} \cdot \hat{C} + \frac{\partial G}{\partial S} \cdot s \\ \hat{C}(0) = \hat{V} \end{cases} \quad (7.10)$$

$$\begin{cases} \frac{d\hat{P}}{dt} + \left[\frac{\partial F}{\partial X} \right]^* \cdot \hat{P} + \left[\frac{\partial^2 F}{\partial X^2} \hat{X} \right]^* \cdot P + \left[\frac{\partial G}{\partial X} \right]^* \cdot \hat{Q} \\ \quad + \left[\frac{\partial^2 G}{\partial X^2} \cdot \hat{X} + \frac{\partial^2 G}{\partial C \partial X} \cdot \hat{C} + \frac{\partial^2 G}{\partial S \partial X} \cdot s \right]^* \cdot Q = E^* E \hat{X} \\ \hat{P}(T) = 0, \end{cases} \quad (7.11)$$

$$\begin{cases} \frac{d\hat{Q}}{dt} + \left[\frac{\partial G}{\partial C} \right]^* \cdot \hat{Q} + \left[\frac{\partial^2 G}{\partial C^2} \cdot \hat{C} + \frac{\partial^2 G}{\partial X \partial C} \hat{X} + \frac{\partial^2 G}{\partial S \partial I} s \right]^* \cdot Q = D^* D \cdot \hat{C} \\ \hat{Q}(T) = 0, \end{cases} \quad (7.12)$$

where \hat{U} in (7.9) and \hat{V} in (7.10) are the Gâteaux derivatives of the optimal initial condition U and V , respectively. These terms are important because the information is propagated from the initial condition. For the response function, the Gâteaux derivatives with respect to S is given by:

$$\hat{\varphi}(S, C, s) = \frac{\partial \varphi}{\partial S} \cdot s + \frac{\partial \varphi}{\partial C} \cdot \hat{C} \quad (7.13)$$

To compute the gradient $\nabla_S \varphi(C, S)$, we introduce four second-order adjoint variables Γ , Λ , Φ and Ψ ; the system of equations (7.9) is multiplied by Γ , (7.10) by Λ , (7.11) by Φ and (7.12) by Ψ , all the terms are added together and integrated by parts and we get:

$$\mathcal{Z} + \int_0^T \langle \hat{X}, \mathcal{A} \rangle dt + \int_0^T \langle \hat{C}, \mathcal{B} \rangle dt + \int_0^T \langle \hat{P}, \mathcal{L} \rangle dt + \int_0^T \langle \hat{Q}, \mathcal{W} \rangle dt + \int_0^T \langle s, \mathcal{K} \rangle dt = 0 \quad (7.14)$$

with:

$$\mathcal{Z} = \langle \Gamma(T), \hat{X}(T) \rangle - \langle \Gamma(0), \hat{X}(0) \rangle + \langle \Lambda(T), \hat{C}(T) \rangle - \langle \Lambda(0), \hat{C}(0) \rangle + \langle \Phi(T), \hat{P}(T) \rangle - \langle \Phi(0), \hat{P}(0) \rangle + \langle \Psi(T), \hat{Q}(T) \rangle - \langle \Psi(0), \hat{Q}(0) \rangle, \quad (7.15)$$

$$\begin{aligned} \mathcal{A} = & -\frac{d\Gamma}{dt} - \left[\frac{\partial F}{\partial X} \right]^* \cdot \Gamma - \left[\frac{\partial G}{\partial X} \right]^* \cdot \Lambda + \left[\frac{\partial^2 F}{\partial X^2} \Phi \right]^* \cdot P \\ & + \left[\frac{\partial^2 G}{\partial X^2} \Phi \right]^* \cdot Q + \left[\frac{\partial^2 G}{\partial C \partial X} \Psi \right]^* \cdot Q - E^* E \Phi, \end{aligned} \quad (7.16)$$

$$\mathcal{B} = -\frac{d\Lambda}{dt} - \left[\frac{\partial G}{\partial C} \right]^* \cdot \Lambda + \left[\frac{\partial^2 G}{\partial C \partial X} \Phi \right]^* \cdot Q + \left[\frac{\partial^2 G}{\partial X^2} \Psi \right]^* \cdot Q - D^* D \Psi, \quad (7.17)$$

$$\mathcal{L} = -\frac{d\Phi}{dt} + \left[\frac{\partial F}{\partial X} \right] \cdot \Phi, \quad (7.18)$$

$$\mathcal{W} = -\frac{d\Psi}{dt} + \left[\frac{\partial G}{\partial C} \right] \cdot \Psi \quad (7.19)$$

and

$$\mathcal{K} = - \left[\frac{\partial G}{\partial S} \right]^* \cdot \Lambda + \left[\frac{\partial^2 G}{\partial X^2} \Phi \right]^* \cdot Q + \left[\frac{\partial^2 G}{\partial C \partial S} \Psi \right]^* \cdot Q. \quad (7.20)$$

If the known values are taken into account in the expression of \mathcal{Z} , it becomes:

$$\begin{aligned} \mathcal{Z} = & \langle \Gamma(T), \hat{X}(T) \rangle - \langle \Gamma(0), \hat{U} \rangle + \langle \Lambda(T), \hat{C}(T) \rangle - \langle \Lambda(0), \hat{V} \rangle \\ & - \langle \Phi(0), \hat{P}(0) \rangle - \langle \Psi(0), \hat{Q}(0) \rangle. \end{aligned} \quad (7.21)$$

If Γ , Λ , Φ and Ψ are the solution of the following second order adjoint systems of equations

$$\begin{cases} -\frac{d\Gamma}{dt} - \left[\frac{\partial F}{\partial X} \right]^* \cdot \Gamma - \left[\frac{\partial G}{\partial X} \right]^* \cdot \Lambda + \left[\frac{\partial^2 F}{\partial X^2} \Phi \right]^* \cdot P \\ \quad + \left[\frac{\partial^2 G}{\partial X^2} \Phi \right]^* \cdot Q + \left[\frac{\partial^2 G}{\partial C \partial X} \Psi \right]^* \cdot Q - E^* E \Phi = 0; \\ \quad \Gamma(0) = 0; \\ \quad \Gamma(T) = 0, \end{cases} \quad (7.22)$$

$$\begin{cases} -\frac{d\Lambda}{dt} - \left[\frac{\partial G}{\partial C} \right]^* \cdot \Lambda + \left[\frac{\partial^2 G}{\partial C \partial X} \Phi \right]^* \cdot Q + \left[\frac{\partial^2 G}{\partial X^2} \Psi \right]^* \cdot Q - D^* D \Psi = \frac{\partial \varphi}{\partial C}; \\ \Lambda(0) = 0; \\ \Lambda(T) = 0, \end{cases} \quad (7.23)$$

$$-\frac{d\Phi}{dt} + \left[\frac{\partial F}{\partial X} \right] \cdot \Phi = 0, \quad (7.24)$$

$$-\frac{d\Psi}{dt} + \left[\frac{\partial G}{\partial C} \right] \cdot \Psi = 0, \quad (7.25)$$

then it comes that:

$$\nabla \varphi = \left[\frac{\partial G}{\partial S} \right]^* \cdot \Lambda - \left[\frac{\partial^2 G}{\partial X^2} \Phi \right]^* \cdot Q - \left[\frac{\partial^2 G}{\partial C \partial S} \Psi \right]^* \cdot Q + \frac{\partial \varphi}{\partial S}. \quad (7.26)$$

The conditions $\Gamma(0) = 0$ and $\Gamma(T) = 0$ are imposed because there is no information on $\hat{X}(T)$ and \hat{U} , respectively. For similar reasons, the conditions $\Lambda(0) = 0$ and $\Lambda(T) = 0$ are also imposed. Because the initial conditions are optimal, $\hat{P}(0)$ defined by

$$\hat{P}(0) = \lim_{\alpha \rightarrow 0} \frac{P(0)|_{S+\alpha S} - P(0)|_S}{\alpha} \quad (7.27)$$

is zero. Similarly, $\hat{Q}(0) = 0$.

We get a coupled system of four differential equations (7.22) to (7.25) of the first order with respect to time. Two equations have an initial and a final condition while the two others have no condition at all: that is a non-standard problem.

A theoretical question remains on the existence and the uniqueness of a solution. Some developments in this direction are underway.

7.2.1 Solving the Non-standard Problem

To make it simple, we consider a system of two differential equations, the extension to four equations being straightforward. The method proposed is based on the theory of optimal control (Lions 1971). We consider the generic system on the time interval $[0, T]$

$$\begin{cases} \frac{dX}{dt} = K(X, Y), & t \in [0, T]; \\ \frac{dY}{dt} = L(X, Y), & t \in [0, T] \end{cases} \quad (7.28)$$

with

$$\begin{cases} X(0) = 0; \\ X(T) = 0 \end{cases} \quad (7.29)$$

and no condition on Y . Let transform it into a problem of optimal control. We consider the problem (7.28) with the initial condition

$$\begin{cases} X(0) = 0; \\ Y(0) = U. \end{cases} \quad (7.30)$$

We assume that under these conditions, (7.28) has a unique solution for $t \in [0, T]$. Let $X(T, U)$ be the value of X at time $t = T$ for the value U of $Y(0)$, we define the cost function

$$J_P(U) = \frac{1}{2} \|X(T, U)\|^2. \quad (7.31)$$

The problem becomes the determination of U^* by minimizing J_P . We can expect that at the optimum $X(T, U^*) = 0$. The Gâteaux derivatives with respect to U in the direction u are given by:

$$\frac{d\hat{X}}{dt} = \frac{\partial K}{\partial X} \cdot \hat{X} + \frac{\partial K}{\partial Y} \cdot \hat{Y} \quad (7.32)$$

$$\frac{d\hat{Y}}{dt} = \frac{\partial L}{\partial X} \cdot \hat{X} + \frac{\partial L}{\partial Y} \cdot \hat{Y} \quad (7.33)$$

$$\hat{X}(0) = 0 \quad (7.34)$$

$$\hat{Y}(0) = u \quad (7.35)$$

$$\hat{J}_p(U) = \langle X(T), \hat{X}(T) \rangle \quad (7.36)$$

Let us introduce the adjoint variables W and Z and proceed to the integration by part. We get:

$$\begin{aligned} & \langle \hat{X}(T), W(T) \rangle + \langle \hat{Y}(T), Z(T) \rangle - \langle \hat{Y}(0), Z(0) \rangle \\ & - \int_0^T \left\langle \hat{X}, \frac{dW}{dt} + \left[\frac{\partial K}{\partial X} \right]^* \cdot W + \left[\frac{\partial L}{\partial X} \right]^* \cdot Z \right\rangle dt \\ & - \int_0^T \left\langle \hat{Y}, \frac{dZ}{dt} + \left[\frac{\partial K}{\partial Y} \right]^* \cdot W + \left[\frac{\partial L}{\partial Y} \right]^* \cdot Z \right\rangle dt = 0 \end{aligned} \quad (7.37)$$

If Z and W are defined as the solution of:

$$\frac{dW}{dt} + \left[\frac{\partial K}{\partial X} \right]^* \cdot W + \left[\frac{\partial L}{\partial X} \right]^* \cdot Z = 0; \quad (7.38)$$

$$\frac{dZ}{dt} + \left[\frac{\partial K}{\partial Y} \right]^* \cdot W + \left[\frac{\partial L}{\partial Y} \right]^* \cdot Z = 0; \quad (7.39)$$

$$Z(T) = 0; \quad W(T) = X(T), \quad (7.40)$$

then we get

$$\nabla J_P(U) = Z(0). \quad (7.41)$$

A theoretical question remains on the existence and the uniqueness of a solution. Some development in this direction are underway.

7.2.2 Extension and Potential Development

Without any theoretical difficulty, the development carried out above can be extended to:

- the case of several sources of pollution; in which case the source S becomes a vector $S = (S_1, S_2, \dots, S_m)$, where S_i is the i th source and m is the total number of sources;
- the case of several pollutant with kinetic chemistry. The same methodology can be applied with a vector of concentration in the place of a single concentration; $C = (C_1, C_2, \dots, C_m)$. Some numerical difficulties may arise if the characteristic time of the kinetic are very heterogeneous.

8 Incremental Methods

During the 1990s, the 4D VAR methodology matured and was adopted at several important international Numerical Weather Prediction centers. However, although 4D VAR cost function and gradient can be evaluated at the cost of one integration of the forecast model followed by one integration of the adjoint model, the computational cost to implement it was still prohibitive since a typical minimization requires between 10 and 100 evaluations of the gradient. The cost of the adjoint model is typically 3 times that of the forward model (Griewank and Walther 2008). The analysis window in a typical operational model such as the ECMWF system is 12-h. Thus, the cost of the analysis is roughly equivalent to between 20 and 200 days of model integration with 10^8 variables, making it computationally prohibitive for NWP centers that have to deliver timely forecasts to the public. Amongst the methods that greatly facilitated the adoption, application and implementation of 4D VAR data assimilation at major operational centers and contributed to advance of the technique, the incremental 4D VAR method is of paramount importance. Courtier et al. (1994) (based on an idea of Derber) introduced the incremental formulation of the 4D VAR. The incremental algorithm reduces the cost of 4D VAR mainly by reducing the resolution of the model, thus allowing the 4D VAR method to become computationally feasible.

The incremental 4D VAR algorithm reduces the resolution of the model and eliminates most of the time-consuming physical packages, thereby enabling the 4D VAR method to become computationally feasible. Furthermore, the incremental 4D VAR algorithm removes the nonlinearities in the cost minimization by using a forward integration of the linear model instead of a nonlinear one. The minimization procedure is identical to the usual 4D VAR algorithm except that the increment trajectory is obtained by integration of the linear model.

The reference trajectory (which is needed for integrating the linear and adjoint models and which starts from the background integration) is not updated at every iteration. This simplified iterative procedure for minimizing the incremental cost function is called the inner loop, and is much cheaper computationally to implement by comparison to the full 4D VAR algorithm. However, when the quadratic cost function is approximated in this way, the incremental 4D VAR algorithm no longer converges to the solution of the original problem. Furthermore, the analysis increments are calculated at reduced resolution and must be interpolated to conform to the high-resolution models grid. Consequently, after performing a user-defined number of inner-loops, one outer-loop is performed to update the high-resolution reference trajectory and the observation departures. After each outer-loop update, it is possible to use progressively higher resolutions for the inner-loop. Other simplifications introduced by the incremental 4D VAR method will be briefly described below. The nonlinearity of the model and/or of the observation operator can produce multiple minima in the cost function, which will impact the convergence of the minimization algorithm. The incremental 4D VAR algorithm removes the nonlinearities in the cost minimization by using a forward integration of the linear model instead of a nonlinear one. It also uses a coarser resolution model and eliminates most of the time-consuming physical packages. In this section we will address several algorithmic aspects of incremental 4D VAR that are used in present day implementations of 4D VAR data assimilation. Some aspects related to the incremental method versus the full 4D VAR were addressed by Li et al. (2000). They conducted a set of four-dimensional variational assimilation (4D VAR) experiments using both a standard method and an incremental method and compared the corresponding performances.

8.1 Description of the Method

Courtier et al. (1994) devised an incremental 4D VAR algorithm which removes nonlinearities in the minimization by using a forward integration of a linear model instead of a nonlinear one. The minimization of the cost functional is carried out at a reduced model resolution which leads to an effective reduction of computational cost and memory requirements. The 4D VAR problem consists in finding the state at time t_0 that minimizes the cost function:

$$J(X_0) = \frac{1}{2}(X_0 - X_b)^T B^{-1}(X_0 - X_b) + \quad (8.1)$$

$$\frac{1}{2} \sum_{i=1}^n (H_i(X_i) - Y_i)^T R_{i-1} (H_i(X_i) - Y_i) \quad (8.2)$$

subject to the states X_i satisfying the NWP model (8.3) as a strong constraint. In optimal control language this is referred to as Partial Differential Equations (PDE) constrained optimization. We consider a discrete nonlinear dynamical system given by the equation:

$$X_{i+1} = M_i(X_i), \quad (8.3)$$

where $X_i \in \mathbb{R}^n$ is the state vector at time t_i and M_i represents the nonlinear model operator that propagates the state vector at time t_i to time t_{i+1} for $i = 0, 1, \dots, n-1$. We assume that we have imperfect observations $Y_i \in \mathbb{R}^{p_i}$ at t_i . Here $H_i : \mathbb{R}^n \rightarrow \mathbb{R}^{p_i}$ is known as being the observation operator and maps the state vector toward the observation space. The matrix B contains the background error covariances and R_i are the observation error covariances matrices. In the incremental formulation the solution to the nonlinear minimization problem is approximated by a sequence of minimizations of linear quadratic cost functions. We define $X_0^{(k)}$ as being the k th estimate to the solution and we linearize the equation around the model trajectory from this estimate. The incremental approach is designed to find the analysis increment $\delta x_0^{(k)} = X_0^{(k)} - X_b$, that minimizes the following cost function

$$\hat{J}(\delta x_0^{(k)}) = \frac{1}{2}(\delta x_0^{(k)})^T B^{-1} \delta x_0^{(k)} + \quad (8.4)$$

$$\frac{1}{2} \sum_{i=1}^n (d_k - \mathbf{H}_i \mathbf{M}_i \delta x_0^{(k)})^T R_{i-1} (d_k - \mathbf{H}_i \mathbf{M}_i \delta x_0^{(k)}), \quad (8.5)$$

where $d_i = Y_i - H_i(M_{0 \rightarrow i}(X_b))$ are the innovation vectors at time step i and \mathbf{M}_i , and \mathbf{H}_i denote the tangent linear versions of the forecast model and observation operators. Now the constraint is given by the tangent linear model

$$\delta x_{i+1}^{(k)} = \mathbf{M}_i \delta x_i^{(k)}.$$

The process of minimization is similar to the usual 4-D VAR algorithm except that the control variable is the increment at time t_0 and \mathbf{M}_i depicts the linear tangent model operator evaluated at the current estimate of the nonlinear trajectory usually referred to as the linearized state. The optimization process is described in Fig. 1.

The trajectory is obtained by integration of the linear model. The reference trajectory required by the linear and adjoint models comes from the background integration and is not updated at every iteration. Correspondingly, the iterative procedure of minimizing the incremental cost function is called the inner-loop which is much cheaper computationally to implement, due to the incremental 4-D VAR simplifications. When the quadratic cost function is approximated in this way, the 4-D VAR algorithm no longer converges to the solution of the original problem. The analysis increments are calculated at a reduced resolution and must be interpolated to the high-resolution model's grid. This drawback is partially overcome by executing after a number of inner-loops, a single outer-loop iteration which is updating the high-resolution reference trajectory and the observation departures. Correspondingly, the iterative procedure of minimizing the incremental cost function is called the outer-loop. After each outer-loop update, it is possible to use a progressively

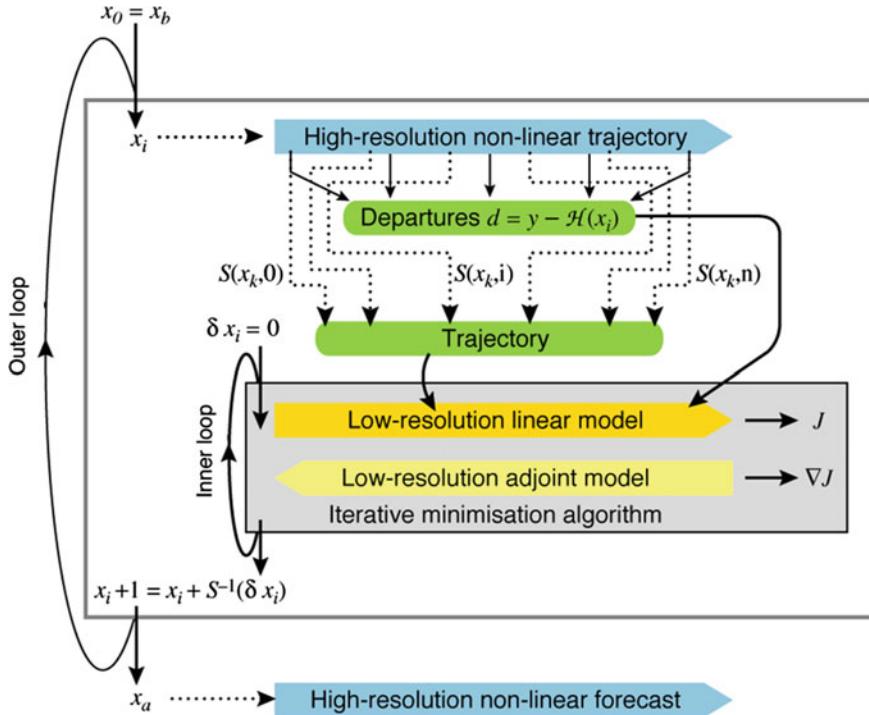


Fig. 1 Incremental 4D-Var

higher resolution for the inner-loops. Such a procedure was carried out in a multi-incremental algorithm proposed by Veersé and Thepaut (1998). The incremental method was shown by Lawless (2010) to be equivalent to an inexact Gauss-Newton method applied to the original nonlinear cost function. The outer-loop iterations can be shown to be locally convergent under certain conditions, provided that the inner-loop minimization is solved with sufficient accuracy (see, e.g., Gratton et al. (2007)). In practice, however, very few outer-loop steps are performed, typically three. The inclusion of full physics in the adjoint model requires the 4-D VAR algorithm to overcome the negative effect of strong nonlinearities present in physics parametrization packages while being able to take advantage of the positive aspects resulting from the consistency between the forecasting nonlinear model and adjoint model. Several approaches have been proposed for mitigating the negative effect of strong nonlinearities in physical processes included in the adjoint model. These approaches involved either direct modifications or simplifications to physical parameterizations. Źupanski and Mesingner (1995) and Tsuyuki (1997) showed beneficial effects when smoothing formulas are used to replace those with discontinuities. The ECMWF system uses simplified physics in the adjoint model, although modifications or simplifications may lead to inconsistencies between the nonlinear forecasting model and the corresponding adjoint model.

In a further, multi-incremental, extension of the incremental 4D VAR method, the inner-loop resolution is increased after each iteration of the outer-loop. In particular, the information about the shape of the cost-function obtained during the early low-resolution iterations provides a very effective preconditioner for subsequent iterations at higher resolution, thus reducing the number of costly iterations. The inner-loops can be efficiently minimized using the conjugate gradient method, provided the cost-function is quadratic, i.e. when the operators involved in the definition of the cost function (the model and the observation operators) are linear. For this reason, the inner-loops have been completely linearized; the non-linear effects are all gathered at the outer-loop level.

9 Developments in Variational Data Assimilation in Last 2 Decades

Starting with the advent of the incremental model a tremendous amount of research efforts focused on implementation of 4-D Var at operational centers. The principle of four-dimensional variational (4D-Var) assimilation usually assumes implicitly that the forecast model is “perfect” within the assimilation window an approach referred to as strong constraint and looks for the model trajectory which best fits the data (background and observations) over the time window. Such a data assimilation method has been implemented in the last few years at various NWP centres with substantial benefit (Rabier 2005). To name a few, Meteo-France, UK Met-Office and Canadian Environment service, led all by pioneering work at ECMWF. However the next step was to consider observation bias, observation error correlation and model error (bias and random) in weak constraint 4-D VAR. See work of Trémolet (2007, 2008) and Akella and Navon (2009).

For Gaussian, temporally-uncorrelated model error, the weak-constrained 4D-Var cost function is The cost function assumes the following form for Gaussian, temporally uncorrelated model error.

$$\begin{aligned}
 J(X) = & \frac{1}{2}(\mathbf{x}_0 - \mathbf{x}_b)^T \mathbf{B}^{-1}(\mathbf{x}_0 - \mathbf{x}_b) \\
 & + \frac{1}{2} \sum_{i=0}^n [\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i]^T \mathbf{R}^{-1} [\mathcal{H}_i(\mathbf{x}_i) - \mathbf{y}_i] \\
 & + \frac{1}{2} \sum_{i=0}^n [\mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1})]^T \mathbf{Q}^{-1} [\mathbf{x}_i - \mathcal{M}_i(\mathbf{x}_{i-1})].
 \end{aligned} \tag{9.1}$$

The matrix \mathbf{Q} is taken usually to be proportional to \mathbf{B} . The pioneering work in weak constraint 4-D VAR is considered to be the one of Derber (1989). This continues to be an area of active research.

9.1 *Estimation of Background and Observation Error Covariances*

Modelling and specification of the covariance matrix of background error constitute important components of any data assimilation system. The main attributes of the background error covariance matrix \mathbf{B} are:

- To spread out the information from the observations; correlations in the background covariance matrix will perform spatial spreading of information from observation points to a finite domain surrounding them;
- To provide statistically consistent increments at the neighboring grid points and levels of the model;
- To ensure that observations of one model variable (e.g., temperature) produce dynamically consistent increments in the other model variables (e.g. vorticity and divergence). For operational models, a typical background covariance matrix contains $10^7 \times 10^7$ elements. Therefore, non-essential components of this important covariance matrix may need to be neglected in order to produce a computationally feasible algorithm.

Construction of background error covariances has been addressed in the literature by the so-called “innovation method”, in which the background errors are assumed to be independent of observation errors. The so-called NMC method was introduced by Parrish and Derber (1992) as a surrogate for samples of background error using differences between forecasts of different length that verify at the same time. The ensemble method for constructing background covariances was proposed by Fisher (2003), while Ingleby (2001) proposed using statistical structures of forecast errors. One can attempt to disentangle information about the statistics of background error from the available information (innovation statistics), or one can try to find a surrogate quantity whose error statistics can be argued to be similar to those of the unknown background errors.

9.2 *Observation Error Covariance*

The problem of variational data assimilation for nonlinear evolution model can be formulated as an optimal control problem to find the initial condition, boundary conditions and/or model parameters. The input data contain observation and background errors, hence there is an error in the optimal solution. For mildly nonlinear dynamics, the covariance matrix of the optimal solution error can be approximated by the inverse Hessian of the cost function w.r.t control variables. For problems characterized by strongly nonlinear dynamics, a new statistical method based on the computation of a sample of inverse Hessians was suggested. This method relies on the efficient computation of the inverse Hessian by means of iterative methods (Lanczos and quasi-Newton BFGS) with preconditioning (Shutyaev et al. 2012; Le Dimet et al. 2002).

Adjoint-based methods makes forecast sensitivity to data assimilation system input parameters $[\mathbf{y}, \mathbf{R}, \mathbf{x}_b, \mathbf{B}]$ possible. Forecast sensitivity to observations (FSO)—is used to monitor the impact of observations to reduce short-range forecast errors. In particular, forecast R -sensitivity (Daescu and Todling 2010; Daescu and Langland 2013) may be used to provide guidance to error covariance tuning procedures. The sensitivity of a scalar measure of forecast error is computed with respect to changes to a set of covariance parameters (Lupu et al. 2014). Forecast R - and B -sensitivities can provide guidance toward the real covariance matrices. The method may show if background information is being over (or under) weighted. In this case it appears the Ensemble Data Assimilation (EDA) based background errors are overweighting the background.

10 Hybrid Data Assimilation

Hybrid Data assimilation is a practical feasible way to introduce flow dependence in the background error covariances required for sequential or variational data assimilation. Starting with Lorenc (2003), Whitaker and Hamill (2002), Buehner et al. (2010b) it was shown that combining the time-varying background error covariance derived from an ensemble of forecasts with stationary, climatological background error covariance leads to improvements. The resulting procedure is so-called, hybrid data assimilation system. Several operational numerical weather prediction centers use three- or four-dimensional variational (3D/4DVar) techniques and have implemented hybrid approaches in these contexts. The hybrid data assimilation involves developing hybrid covariance models, i.e. a linear combination of a static B matrix (built from climatology and typically used in 4D-Var applications) with a flow-dependent B matrix (described using an ensemble). This hybrid approach has been operational at ECMWF for some time (Buizza et al. 2008; Isaksen et al. 2010; Bonavita et al. 2012), and is now operational at the Met Office, UK, for their global model (Clayton et al. 2013) and at Environment Canada (Buehner et al. 2010a). A theoretical basis for the construction of the hybrid covariances, in particular how to weigh static and flow-dependent components, is described by Bishop and Satterfield (2013) and Bishop et al. (2013).

11 Numerical Experiments

This section focuses on non-linear strong constraint 4D-Var experiments and it is divided in two parts. The first one centers on numerical simulations using 1D-Burgers model while the second part concentrates on computer simulations of shallow water equations model. As discussed, the adjoint models are required to alleviate the computational complexity of estimating the gradient during the optimization routines. It is known that such models have validity regions depending on the amount

of perturbation considered as input. To asses the quality of adjoint models, tangent linear and adjoint tests are performed. Then the potential of 4D-Var method is discussed, its objective function and associated gradient as well as the analysis errors with respect to the observations being examined.

11.1 Burgers Model

Burgers' equation is an important partial differential equation from fluid mechanics (Burgers 1948). The evolution of the velocity u of a fluid evolves according to

$$\frac{\partial u}{\partial t} + u \frac{\partial u}{\partial x} = \mu \frac{\partial^2 u}{\partial x^2}, \quad x \in [-1, 1], \quad t \in (0, 0.2]. \quad (11.1)$$

Here μ denotes the viscosity coefficient. The model has homogeneous Dirichlet boundary conditions $u(-1, t) = u(1, t) = 0$, and the integration time is $t \in (0, 0.2]$. An Euler explicit scheme is implemented using a spatial mesh of $nx = 41$ equidistant points on $[-1, 1]$, with $\Delta x = 0.05$. A uniform temporal mesh with $nt = 21$ points covers the interval $[0, 0.02]$, with $\Delta t = 0.001$. A set of initial conditions is depicted in Fig. 2 together with the final solution obtained after integrating the discrete Burgers model (11.1) in time.

For our data assimilation experiments, we add uniform random perturbations $\varepsilon \in U(-0.5, 0.5)$ to the above truth state initial conditions and generate twin-experiment observations at every grid space point location and every time step. The background state or the first guess for the 4D-Var simulations is shown in Fig. 3 along with the final time solution. The background and observation error covariance matrices are taken to be identity matrices.

Fig. 2 Initial and final true states

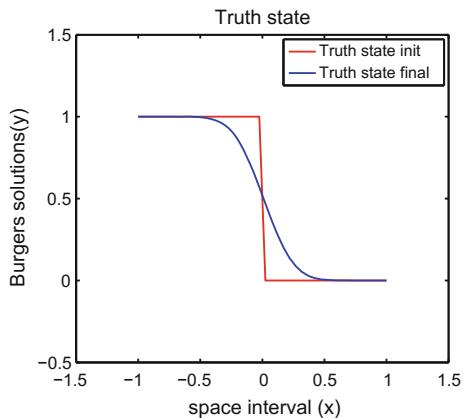
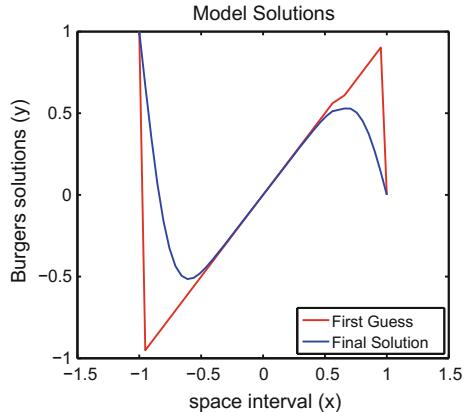


Fig. 3 Initial and final solutions of Burgers model



The objective function of lack of fit between model forecast and observations is minimized using the limited-memory Broyden–Fletcher–Goldfarb–Shanno algorithm (LBFGS) introduced by Liu and Nocedal (1989). This optimization algorithm is a member of quasi-Newton methods and requires the gradient estimation of the 4D-Var objective function. As mentioned earlier, the fastest approach for computing the associated gradient is to employ an adjoint model. Before using it inside the LBFGS algorithm, it is necessary to verify the accuracy of the tangent linear and adjoint models by checking their output agreement with the finite difference approximations. These linearity tests used in this study are derived from the alpha test described by Navon et al. (1992, Eq. (2.20)). The depicted values from Fig. 4 are obtained using

$$\text{adj}_{\text{test}} = \frac{J(\mathbf{x}_0 + \alpha \delta \mathbf{x}_0) - J(\mathbf{x}_0)}{\langle \nabla J(\mathbf{x}_0), \alpha \delta \mathbf{x}_0 \rangle_2}, \quad \text{tl}_{\text{test}} = \frac{\|\mathcal{M}_{0,ff}(\mathbf{x}_0 + \alpha \delta \mathbf{x}_0)(t_f) - \mathcal{M}_{0,ff}(\mathbf{x}_0)(t_f)\|_2}{\|\mathbf{M}_0, t_f(\alpha \delta \tilde{\mathbf{x}}_0)(t_f)\|_2}, \quad (11.2)$$

where α represented on the x axis in Fig. 4 controls the magnitude of the perturbation. The forward and tangent linear models denoted by \mathcal{M} and \mathbf{M} are integrated along the entire time interval. The test formulations (11.2) uses t_f to represent the final time which in our case was set to $t_f = 0.2$ and we employed the Euclidian scalar product and norm.

Figure 5 depicts the minimization performances of the 4D-Var Burgers data assimilation system. The cost function and gradient values are normalized by dividing them with their initial values. After no more than 50 iterations the optimization routine stopped reaching a gradient value smaller than 10^{-4} .

The analysis is obtained and its quality is measured in terms of Euclidian distance away from the observations. In comparison with the first guess, we notice a significant improvement in the error magnitude as shown in Fig. 6. This underlines the performance of the variational 4D-Var algorithm justifying its large scale usage at numerical weather prediction centers around the globe in the form of incremental 4D-Var.

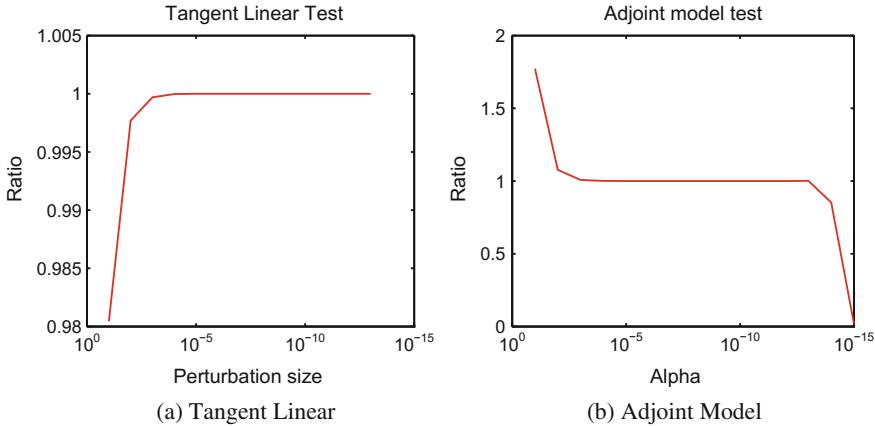


Fig. 4 Linearization tests validating the usage of adjoint and tangent linear models for certain perturbations ranges

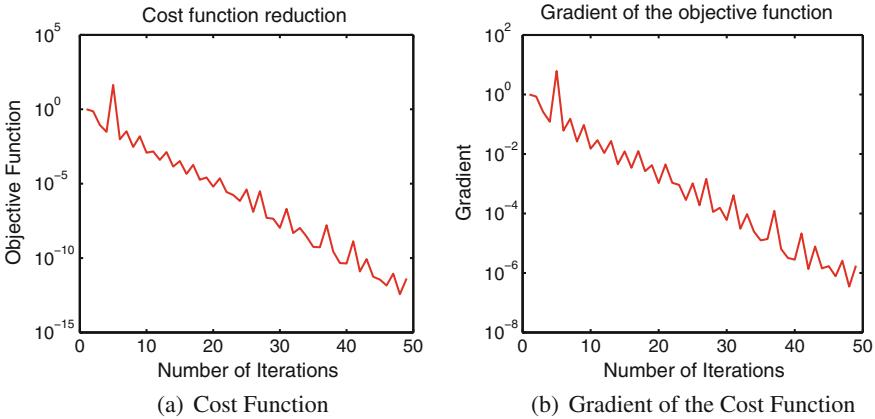


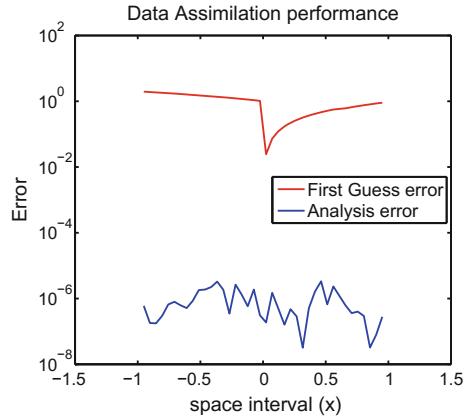
Fig. 5 Reduction in the cost function and its associated gradient

11.2 Shallow Water Equations Model

SWE has proved its capabilities in modeling propagation of Rossby waves in the atmosphere, rivers, lakes and oceans as well as gravity waves in a smaller domain. The alternating direction fully implicit finite difference scheme (Gustafsson 1971) is considered in this chapter and it is stable for large CFL condition numbers (we tested the stability of the scheme for a CFL condition number equal up to 8.9301). We also refer to Fairweather and Navon (1980); Navon and Villiers (1986) for other research work on this topic.

The SWE model using the β -plane approximation on a rectangular domain is introduced (see Gustafsson (1971))

Fig. 6 First guess and optimal initial conditions errors



$$\frac{\partial w}{\partial t} = A(w) \frac{\partial w}{\partial x} + B(w) \frac{\partial w}{\partial y} + C(y)w, \quad (x, y) \in [0, L] \times [0, D], \quad t \in (0, t_f], \quad (11.3)$$

where $w = (u, v, \phi)^T$ is a vector function, u, v are the velocity components in the x and y directions, respectively, h is the depth of the fluid, g is the acceleration due to gravity, and $\phi = 2\sqrt{gh}$.

The matrices A , B and C are assuming the form

$$A = - \begin{pmatrix} u & 0 & \phi/2 \\ 0 & u & 0 \\ \phi/2 & 0 & u \end{pmatrix}, \quad B = - \begin{pmatrix} v & 0 & 0 \\ 0 & v & \phi/2 \\ 0 & \phi/2 & v \end{pmatrix}, \quad C = \begin{pmatrix} 0 & f & 0 \\ -f & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

where f is the Coriolis term.

$$f = \hat{f} + \beta(y - D/2), \quad \beta = \frac{\partial f}{\partial y}, \quad \forall y,$$

with \hat{f} and β constants.

We assume periodic solutions in the x direction for all three state variables while in the y direction

$$v(x, 0, t) = v(x, D, t) = 0, \quad x \in [0, L], \quad t \in (0, t_f]$$

and Neumann boundary conditions are considered for u and ϕ .

Initially $w(x, y, 0) = \psi(x, y)$, $\psi : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$, $(x, y) \in [0, L] \times [0, D]$. Now we introduce a mesh of $n = N_x \cdot N_y$ equidistant grid points on $[0, L] \times [0, D]$, with $\Delta x = L/(N_x - 1)$, $\Delta y = D/(N_y - 1)$. We also discretize the time interval $[0, t_f]$ using N_t equally distributed points and $\Delta t = t_f/(N_t - 1)$. Next we define vectors of the unknown variables of dimension n containing approximate solutions such as

$$\mathbf{w}(t_N) \approx [w(x_i, y_j, t_N)]_{i=1,2,\dots,N_x, \quad j=1,2,\dots,N_y} \in \mathbb{R}^n, \quad N = 1, 2, \dots, N_t.$$

The semi-discrete equations of SWE (11.3) are:

$$\mathbf{u}' = -F_{11}(\mathbf{u}) - F_{12}(\boldsymbol{\phi}) - F_{13}(\mathbf{u}, \mathbf{v}) + \mathbf{F} \odot \mathbf{v}, \quad (11.4)$$

$$\mathbf{v}' = -F_{21}(\mathbf{u}) - F_{22}(\mathbf{v}) - F_{23}(\boldsymbol{\phi}) - \mathbf{F} \odot \mathbf{u}, \quad (11.5)$$

$$\boldsymbol{\phi}' = -F_{31}(\mathbf{u}, \boldsymbol{\phi}) - F_{32}(\mathbf{u}, \boldsymbol{\phi}) - F_{33}(\mathbf{v}, \boldsymbol{\phi}) - F_{34}(\mathbf{v}, \boldsymbol{\phi}), \quad (11.6)$$

where $\mathbf{u}', \mathbf{v}', \boldsymbol{\phi}'$ denote semi-discrete time derivatives, $\mathbf{F} \in \mathbb{R}^n$ stores Coriolis components, \odot is the component-wise multiplication operator, while the nonlinear terms F_{ij} are defined as follows:

$$\begin{aligned} F_{11}, F_{12}, F_{21}, F_{23} : \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad F_{13}, F_{22}, F_{3i} : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}^n, \quad i = 1, 2, 3, 4, \\ F_{11}(\mathbf{u}) = \mathbf{u} \odot A_x \mathbf{u}, \quad F_{12}(\boldsymbol{\phi}) = \frac{1}{2} \boldsymbol{\phi} \odot A_x \boldsymbol{\phi}, \quad F_{13}(\mathbf{u}, \mathbf{v}) = \mathbf{v} \odot A_y \mathbf{u}, \\ F_{21}(\mathbf{u}, \mathbf{v}) = \mathbf{u} \odot A_x \mathbf{v}, \quad F_{22}(\mathbf{v}) = \mathbf{v} \odot A_y \mathbf{v}; \quad F_{23}(\boldsymbol{\phi}) = \frac{1}{2} \boldsymbol{\phi} \odot A_y \boldsymbol{\phi}, \\ F_{31}(\mathbf{u}, \boldsymbol{\phi}) = \frac{1}{2} \boldsymbol{\phi} \odot A_x \mathbf{u}, \quad F_{32}(\mathbf{u}, \boldsymbol{\phi}) = \mathbf{u} \odot A_x \boldsymbol{\phi}, \\ F_{33}(\mathbf{v}, \boldsymbol{\phi}) = \frac{1}{2} \boldsymbol{\phi} \odot A_y \mathbf{v}, \quad F_{34}(\mathbf{v}, \boldsymbol{\phi}) = \mathbf{v} \odot A_y \boldsymbol{\phi}, \end{aligned} \quad (11.7)$$

where $A_x, A_y \in \mathbb{R}^{n \times n}$ are constant coefficient matrices for discrete first-order and second-order differential operators which incorporate the boundary conditions. The numerical scheme was implemented in Fortran and uses a sparse matrix environment. For operations with sparse matrices we employed SPARSEKIT library (Saad 1994) and the sparse linear systems obtained during the quasi-Newton iterations were solved using MGMRRES library (Barrett et al. 1994; Kelley 1995; Saad 2003). Here we did not decouple the model equations as in Stefanescu and Navon (2013) where the Jacobian is either block cyclic tridiagonal or block tridiagonal. By keeping all discrete equations together the corresponding SWE adjoint model can be solved with the same implicit scheme used for forward model.

For the 4D-Var numerical experiments we use the following constants $L = 6000 \text{ km}$, $D = 4400 \text{ km}$, $t_f = 3 \text{ h}$, $\hat{f} = 10^{-4} \text{ s}^{-1}$, $\beta = 1.5 \times 10^{-11} \text{ s}^{-1} \text{ m}^{-1}$, $g = 10 \text{ ms}^{-2}$, $H_0 = 2000 \text{ m}$, $H_1 = 220 \text{ m}$, $H_2 = 133 \text{ m}$. The mesh coordinates are $N_x = 31$, $N_y = 21$ and $N_t = 91$. Next we derived the initial conditions from the initial height condition No. 1 of Grammeltvedt (1969) i.e.

$$h(x, y) = H_0 + H_1 + \tanh\left(9 \frac{D/2 - y}{2D}\right) + H_2 \operatorname{sech}^2\left(9 \frac{D/2 - y}{2D}\right) \sin\left(\frac{2\pi x}{L}\right),$$

$$0 \leq x \leq L, \quad 0 \leq y \leq D.$$

The initial velocity fields were then obtained from the initial height field using the geostrophic relationship

$$u = \left(\frac{-g}{f} \right) \frac{\partial h}{\partial y}, \quad v = \left(\frac{g}{f} \right) \frac{\partial h}{\partial x}.$$

The background and observation error covariance matrices are taken diagonal. The associated background variances are set to 4, 6 and 55 for velocity components u and v and geopotential height ϕ . The observations variances are taken constant for the entire time interval with 3 and 5 associated to velocity components while 50 is used in the case of geopotential height. The corresponding objective function has the following form:

$$J(\mathbf{x}_0) = \frac{1}{2} (\mathbf{w}_0^b - \mathbf{w}_0)^T \mathbf{B}_0^{-1} (\mathbf{w}_0^b - \mathbf{w}_0) + \frac{1}{2} \sum_{i=0}^{N_{obs}} (\mathbf{y}_i - H_i(\mathbf{w}_i))^T \mathbf{R}^{-1} (\mathbf{y}_i - H(\mathbf{w}_i)), \quad (11.8)$$

where \mathbf{w}_0^b is the background state. The observations are not available at every time step but only at 10 locations inside the time interval, i.e. 1, 10, 19, 28, 37, 46, 55, 64, 73, 91. The observation operators H_i are taken as identities meaning that we are observing the state variables only. These observations are obtained by perturbing the trajectory associated with Grammeltvedt conditions. The additive observation noise is set to $0.1 \cdot \varepsilon_o$, where ε_o is sampled from a multivariate distribution $\mathcal{N}(\mathbf{0}, \mathbf{R})$. The background state is obtained by adding normal noise $3 \cdot \varepsilon_b$, where $\varepsilon_b \in \mathcal{N}(\mathbf{0}, \mathbf{B})$. The length of the assimilation window is selected to be 3 h. The implicit scheme allowed us to integrate in time using a larger time step and select $N_t = 91$ time steps.

The Broyden-Fletcher-Goldfarb-Shanno (BFGS) optimization method option contained in the CONMIN software (Shanno and Phua 1980) is employed for the SWE 4-D VAR. BFGS uses a line search method which is globally convergent in the sense that $\lim_{k \rightarrow \infty} \|\nabla J^{(k)}\| = 0$ and utilizes approximate Hessians to include convergence to a local minimum. The discrete tangent linear and adjoint models were derived by hand and their accuracy was verified using (Navon et al. 1992) techniques and the results are depicted in Fig. 7.

The performances of SWE 4D-Var data assimilation system are presented in Fig. 8. The square norm of the gradient is shown in panel (b). The optimization routine stopped after 60 iterations when the local criterium $\nabla J^{(k)} < 10^{-4}$ was satisfied.

Figure 9 depicts the analysis geostrophic wind field in comparison with the first guess field. Clearly the algorithm was able to recover more accurate state variables from the observations as evidenced from the truth wind field shown in panel (a). Similar results are obtained for the geopotential analysis as shown in Fig. 10.

The applications of reduced order modeling techniques have the potential to significantly speed up the solution of variational data assimilation problems with non-linear dynamical models as shown by Ștefănescu et al. (2015). This could represent a very important step for obtaining analyses faster than real time at numerical weather prediction centers. Moreover it has been proved that the solution of reduced

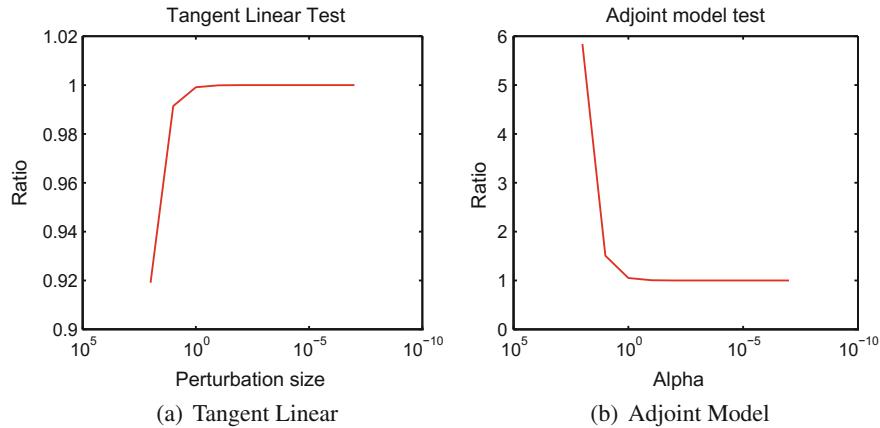


Fig. 7 Swallow water equations tangent linear and adjoint models tests

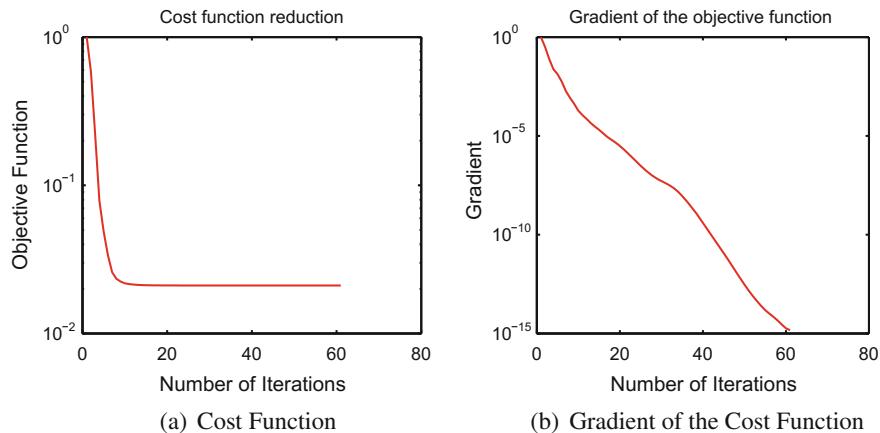


Fig. 8 Reduction in the cost function and its associated gradient

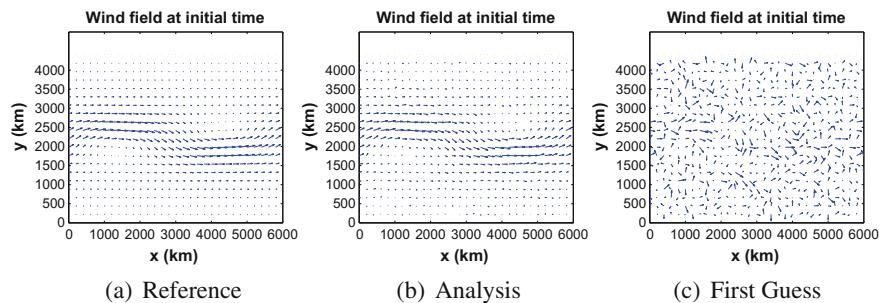


Fig. 9 Initial conditions of wind before and after data assimilation

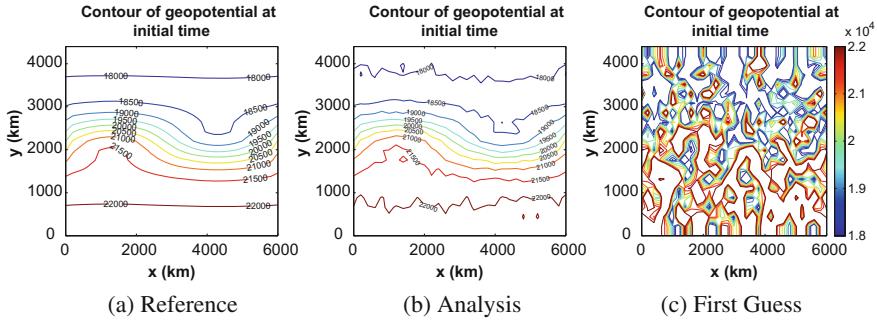


Fig. 10 Geopotential initial conditions before and after data assimilation

optimization system equipped with a trust region algorithm converges to the high-fidelity analyses (Arian et al. 2000) which is not the case with the solution of incremental 4D-Var (Trémoloet 2007).

12 Outlook of Modern Data Assimilation Topics

The 4D-VAR data assimilation is available and implemented today at several operational numerical weather prediction centers starting with European Centre for Medium- Range Weather Forecasts (ECMWF), (Rabier et al. 2007; Klinker et al. 2000) while a similar system was operational at Meteo-France in 2000 (Janisková et al. 1999; Gauthier and Thepaut 2001; Desroziers et al. 2003). More recently 4-D VAR was implemented at UK Met office, Japan and Canada. Park and Županski (2003) survey the status and progress of the four-dimensional variational data assimilation with emphasis on application to prediction of mesoscale/storm-scale atmospheric phenomena. See also Županski et al. (2002). The impact of adopting 4-D VAR was qualified as substantial, resulting in an improvement in NWP quality and accuracy (see Rabier (2005) in special Issue of QJRMS 2005). 4-D VAR combined with improvement in error specifications and with a large increase in a variety of observations has led to improvements in NWP accuracy (Simmons and Hollingsworth 2002). Hollingsworth et al. (2005) show how observing system improvements led to improvements of forecast scores while Bouttier and Kelly (2001) show that the improvement of forecast scores for the southern hemisphere are due to satellite data. Also, error statistics for different sources of observation constitutes an active field of research aimed mainly at obtaining better representation of the specific observation operators. It has become amply evident that in the last 15 years major improvements in NWP are due to large extent to development of sources of observations and that 4-D VAR and sequential data assimilation can take advantage of them due to major research efforts at universities, federal laboratories and operational centers. For new opportunities for research see the article by McLaughlin et al.

(2005) that illuminates and outlines possibilities for enhanced collaboration within the data assimilation community. It is certain that data assimilation concepts have become widely applied in all the geosciences as more geoscience scientific disciplines gain access to larger amounts of data, from satellite remote sensing and from sensor networks, and as Earth system models increase in both accuracy and sophistication.

12.1 *Data Assimilation Applied to Other Fields*

Data assimilation methods are currently also used in other environmental forecasting problems, e.g. in hydrological forecasting. Basically, the same types of data assimilation methods as those described above are in use there. An example of chemical data assimilation using AUTOCHEM can be found at CDA Central. Given the abundance of spacecraft data for other planets in the solar system, data assimilation is now also applied beyond the Earth to obtain re-analyses of the atmospheric state of extra-terrestrial planets. Mars is the only extraterrestrial planet to which data assimilation has been applied so far. Available spacecraft data include, in particular, retrievals of temperature and dust/water ice optical thicknesses from the Thermal Emission Spectrometer on board NASA's Mars Global Surveyor and the Mars Climate Sounder on board NASA's Mars Reconnaissance Orbiter. Two methods of data assimilation have been applied to these datasets: an Analysis Correction scheme and two Ensemble Kalman Filter schemes. Both are using a global circulation model of the Martian atmosphere as forward model. The Mars Analysis Correction Data Assimilation (MACDA) dataset is publicly available from the British Atmospheric Data Centre. Data assimilation is now a part of the challenge for every forecasting problem encompassing multi-physics multi-scale systems. Dealing with biased data is a serious challenge in data assimilation. Further development of methods to deal with biases will be of particular use. If there are several instruments observing the same variable then intercomparing them using probability distribution functions can be useful. Other uses include trajectory estimation for the Apollo program, GPS, and atmospheric chemistry. A particular application is the prediction of future oil production. Data assimilation is extensively used in petroleum reservoir engineering, where it is usually referred to as "history matching". Data assimilation methods are used for uncertainty assessment of performance predictions of wells in oil reservoirs and for generating computational models used for optimizing decision parameters that would improve oil recovery. Recently data assimilation has been extended to blood circulation in hemodynamics to determine and analyze the blood flow patterns in the aortic root—since flow reconstruction by image processing is not accurate enough.

12.2 *Further Applications of Variational Data Assimilation*

At the beginning VDA methods were applied to meteorology, first to mesoscale and then to global models. In a second phase VDA methods were applied to oceanography when optimal interpolation methods became unable to retrieve physically consistent fields.

During this period the observation of the Earth has experienced a major improvement due the advent of satellites.

Neither for the atmosphere nor for the ocean are satellites observing the state variables of the mathematical models, rather what is basically observed and measured are radiances, which are indirectly linked to the state variables, by solving inverse problems and this is the purpose of variational methods, therefore spatial observations were naturally introduced in VDA.

12.2.1 Data Assimilation for Continental Waters

The evolution of rivers is of great importance especially for flood prediction. Going back to the ingredients of VDA, what are they in the framework of continental waters:

- Models: Basically we have to deal with the equations of conservation derived from fluids dynamics. For some rivers such as the Yangtze river, the content of the sediment flow has to be taken into account. There are difficulties to define the geometry of the domain both on the lateral boundaries and also on the bottom of the river which is poorly known in practical applications.
- Sink and source terms, these terms are of various nature: rain, infiltration, sources, they influence the geometry of the domain and most of the time they are poorly known and subject to nonlinearities like for instance the saturation of the soil.
- Data: many rivers are not equipped with sensors, the cost of operational observations is very expensive and furthermore they not accessible to satellite measurements.
- Statistics: Very poor for the case of extreme events.

Therefore the framework of VDA as it is applied in meteorology is very difficult to transpose to hydrology (Vieux et al. 1998), nevertheless VDA is useful for model calibration and sensitivity analysis.

12.2.2 Data Assimilation in Agronomy

In the last few years, encouraging results using radiative transfer model inversion techniques were obtained for land biophysical variables retrieval. However, the inversion of radiative transfer models is a severely ill-posed problem, that may lead to significant uncertainties in the biophysical variables estimates. Improvement of performance of the inversion process requires additional information to be exploited by

including better radiative transfer models, exploitation of proper prior information on the distribution of the canopy and atmosphere variables, knowledge of uncertainties in satellite measurements, as well as possible spatial and temporal constraints.

In their paper, Lauvernet et al. (2008) focus on the use of coupled atmosphere-surface radiative transfer models (SMAC+SAIL+PROSPECT) to estimate some key biophysical variables from top of atmosphere canopy reflectance data. The inversion is achieved over an ensemble of pixels belonging to a spatial window where aerosol properties are assumed to be constant, and over a temporal window of few days where the vegetation state is assumed not to vary. The ensemble inversion scheme accounting for the spatial and temporal constraints is described. Top of the atmosphere reflectance observations are simulated for 13 spectral bands within the visible and near infrared domains. The coupled model is inverted with a variational method implementation aimed at solving very large inverse problems. It is based on the use of the adjoint model and a Quasi-Newton optimisation technique with a BFGS update.

12.2.3 Data Assimilation for Plant Growth

Functional-Structural plant models (FSPM) combine process-based models and architectural models for a better description of plant growth. The process-based models characterize plants mechanics like photosynthesis for agronomic applications.

By contrast, the architectural models were originally developed to analyze botanic patterns and/or topological structures. A typical FSPM can thus simulate not only plant organogenesis but also biomass production and partition at organ level (leaf, fruit, internode). Wu et al. (2012) have used the FSPM GreenLab to model and optimize plant growth thanks to a dynamical system. This growth algorithm is based on a minimal set of physiological knowledge, such as the empirical rules of plant-environment interactions for biomass acquisition and the source-sink relations among organs that compete for assimilates.

Consequently the plant morphological plasticity can be described by a small set of endogenous parameters, thus reducing the complexity of parameters calibration. The model takes the form of a discrete non linear dynamical system, a difficulty arises from the fact that the nonlinearity is partly due to the biological thresholding or saturation effects. An optimal control method has been applied for plant functional-structural growth. Using variational methods based on optimal control two problems have been investigated, the calibration of models by minimizing the discrepancy between observations and computed solutions of the Greenlab model and also by solving an optimal water supply problem applied to growth of sunflower plants. The classical tools of Variational Data Assimilation, such as the adjoint model and optimization algorithms have been employed.

12.2.4 Assimilation of Images

The observation of the earth by satellites is an important source of information if we consider the dynamics of the flows: the evolution of fronts and/or storms provides an intuitive overview of the weather to come. This remark is also true for the ocean with temperature, salinity or the color of the ocean due to biological (algae) activities. From the dynamical point of view the information comes from the evolution of the discontinuities in the images. An important question is what is the nature of these images, basically we have two different phenomena:

- Lagrangian images: this is the case of small cumulus clouds under the tropics. In operational meteorology they are considered as lagrangian tracers, their velocity is estimated then plugged in a classical data assimilation scheme.
- Eulerian images: this is the case of lenticular clouds over a mountain, they looks almost steady state and nevertheless they are the signature of a strong meteorological feature. Therefore an estimation of the wind velocity based on the evolution of this clouds would give a wrong evaluation.

Therefore the problem is how to couple the dynamics of images with numerical models? A functional space has to be defined for the images and also an operator from the space of solutions of the model towards the space of the images. Because the information is in the discontinuities of the images, the metrics should not be too much regularizing; the choice of adequate metrics remains an open problem (Le Dimet et al. 2014).

12.2.5 Data Assimilation in Medicine

For a living person some direct measurement cannot be directly carried out: this is the case of the heart. There exist several mathematical models for the heart counting on several parameters specific of the heart. The methods of data assimilation and the assimilation of images have been used for heart models calibration (Chapelle et al. 2013).

References

- Akella S, Navon IM (2009) Different approaches to model error formulation in 4D-Var: a study with high-resolution advection schemes. *Tellus A* 61(1):112–128
- Arian E, Fahl M, Sachs EW (2000) Trust-region proper orthogonal decomposition for flow control. Technical report, DTIC Document
- Barrett R, Berry M, Chan TF, Demmel J, Donato J, Dongarra J, Eijkhout V, Pozo R, Romine C, Van der Vorst H (1994) Templates for the solution of linear systems: building blocks for iterative methods, 2nd edn. SIAM, Philadelphia, PA
- Bertsekas DP (1982) Constrained optimization and Lagrange multiplier methods. Academic Press, New York

- Bishop CH, Satterfield EA (2013) Hidden error variance theory. Part I: exposition and analytic model. *Monthly Weather Rev* 141(5):1454–1468
- Bishop CH, Satterfield EA, Shanley KT (2013) Hidden error variance theory. Part II: an instrument that reveals hidden error variance distributions from ensemble forecasts and observations. *Monthly Weather Rev* 141(5):1469–1483
- Bonavita M, Isaksen L, Hólm E (2012) On the use of EDA background error variances in the ECMWF 4D-Var. *Q J Royal Meteorol Soc* 138(667):1540–1559
- Bouttier F, Kelly G (2001) Observing-system experiments in the ecmwf 4d-var data assimilation system. *Q J Royal Meteorol Soc* 127(574):1469–1488
- Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B (2010a) Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: description and single-observation experiments. *Monthly Weather Rev* 138(5):1550–1566
- Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B (2010b) Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: one-month experiments with real observations. *Monthly Weather Rev* 138(5):1567–1586
- Buizza R, Leutbecher M, Isaksen L (2008) Potential use of an ensemble of analyses in the ECMWF ensemble prediction system. *Q J Royal Meteorol Soc* 134(637):2051–2066
- Burgers JM (1948) A mathematical model illustrating the theory of turbulence. *Adv Appl Mech* 1:171–199
- Cacuci DG, Hall MCG (1984) Efficient estimation of feedback effects with application to climate models. *J Atmos Sci* 41(13):2063–2068
- Chapelle D, Fragu M, Mallet V, Moireau P (2013) Fundamental principles of data assimilation underlying the Verdandi library: applications to biophysical model personalization within euHeart. *Med Biol Eng Comput* 51(11):1221–1233
- Charney JG, Fjörtoft R, Von Neumann J (1950) Numerical integration of the barotropic vorticity equation. *Tellus A* 2(4):
- Clayton AM, Lorenc AC, Barker DM (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q J Royal Meteorol Soc* 139(675):1445–1461
- Courtier P, Thépaut J-N, Hollingsworth A (1994) A strategy for operational implementation of 4d-var, using an incremental approach. *Q J Royal Meteorol Soc* 120(519):1367–1387
- Daescu DN, Langland RH (2013) Error covariance sensitivity and impact estimation with adjoint 4d-var: theoretical aspects and first applications to navdas-ar. *Q J Royal Meteorol Soc* 139(670):226–241
- Daescu DN, Todling R (2010) Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Q J Royal Meteorol Soc* 136(653):2000–2012
- Derber JC (1989) A variational continuous assimilation technique. *Monthly Weather Rev* 117(11):2437–2446
- Desroziers G, Hello G, Thepaut J-N (2003) A 4d-var re-analysis of fastex. *Q J Royal Meteorol Soc* 129(589):1301–1315
- Euler L (1766) Elementa calculi variationum. Originally published in *Novi Commentarii academiae scientiarum Petropolitanae* 10:51–93
- Euler L (1744) Methodus inveniendi lineas curvas maximi minimive proprietate gaudentes sive solutio problematis isoperimetrii latissimo sensu accepti. Bousquet, Lausannae e Genevae, E65A. OO Ser. I, 24, 1952
- Fairweather G, Navon IM (1980) A linear ADI method for the shallow water equations. *J Comput Phys* 37:1–18
- Fisher M (2003) Background error covariance modelling. In: Seminar on recent development in data assimilation for atmosphere and ocean, pp 45–63
- Fletcher R (2013) Practical methods of optimization. Wiley
- Galerkin BG (1915) Series occurring in various questions concerning the elastic equilibrium of rods and plates. *Vestnik Inzhenerov Tech* 19:897–908

- Gauthier P, Thepaut J-N (2001) Impact of the digital filter as a weak constraint in the preoperational 4dvar assimilation system of météo-france. *Monthly Weather Rev* 129(8):2089–2102
- Ghil M, Cohn S, Tavantzis J, Bube K, Isaacson E (1981) Applications of estimation theory to numerical weather prediction. In: *Dynamic meteorology: data assimilation methods*. Springer, pp 139–224
- Gill PE, Murray W, Wright MH (1981) Practical optimization
- Grammeltvedt A (1969) A survey of finite difference schemes for the primitive equations for a barotropic fluid. *Monthly Weather Rev* 97(5):384–404
- Gratton S, Lawless AS, Nichols NK (2007) Approximate gauss-newton methods for nonlinear least squares problems. *SIAM J Optim* 18(1):106–132
- Griewank A, Walther A (2008) Evaluating derivatives: principles and techniques of algorithmic differentiation. SIAM
- Gustafsson B (1971) An alternating direction implicit method for solving the shallow water equations. *J Comput Phys* 7:239–254
- Hall MCG, Cacuci DG, Schlesinger ME (1982) Sensitivity analysis of a radiative-convective model by the adjoint method. *J Atmos Sci* 39(9):2038–2050
- Hollingsworth A, Uppala S, Klinker E, Burridge D, Vitart F, Onvlee J, De Vries JW, De Roo AD, Pfrang C (2005) The transformation of earth-system observations into information of socio-economic value in geoss. *Q J Royal Meteorol Soc* 131(613):3493–3512
- Ingleby NB (2001) The statistical structure of forecast errors and its representation in the met office global 3-d variational data assimilation scheme. *Q J Royal Meteorol Soc* 127(571):209–231
- Isaksen L, Bonavita M, Buizza R, Fisher M, Haseler J, Leutbecher M, Raynaud L (2010) Ensemble of data assimilations at ECMWF. *European Centre for Medium-Range Weather Forecasts*
- Janisková M, Thépaut J-N, Geleyn J-F (1999) Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. *Monthly Weather Rev* 127(1):26–45
- Kelley CT (1995) Iterative methods for linear and nonlinear equations. Number 16 in *Frontiers in applied mathematics*. SIAM
- Klinker E, Rabier F, Kelly G, Mahfouf J-F (2000) The ECMWF operational implementation of four-dimensional variational assimilation. III: experimental results and diagnostics with operational configuration. *Q J Royal Meteorol Soc* 126(564):1191–1215
- Lagrange JL (1762) Application de la méthode exposée dans le mémoire précédent à la solution des problèmes de dynamique différents. *Œuvres de Lagrange (1867–1892)* 1:151–316
- Lagrange JL (1761) Essai d'une nouvelle méthode pour de'terminer les maxima, et les minima des formules integrales indefinies
- Lauvernet C, Baret F, Hascoët L, Buis S, Le Dimet F-X (2008) Multitemporal-patch ensemble inversion of coupled surface-atmosphere radiative transfer models for land surface characterization. *Remote Sens Environ* 112(3):851–861
- Lawless AS (2010) A note on the analysis error associated with 3d-fgat. *Q J Royal Meteorol Soc* 136(649):1094–1098
- Le Dimet F-X, Souopgui I, Titaud O, Shutyaev V (2014) Toward the assimilation of images. *Non-linear Proces Geophys Discuss* 1:1381–1430
- Le Dimet F-X, Navon IM, Daescu DN (2002) Second-order information in data assimilation. *Monthly Weather Rev* 130(3):629–648
- Le Dimet F-X, Ngodock HE, Luong B, Verron J (1997) Sensitivity analysis in variational data assimilation. *J Meteorol Soc Jpn Ser 2*(75):135–145
- Le Dimet F-X, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus A* 38(2):97–110
- Li Z, Navon IM, Zhu Y (2000) Performance of 4D-Var with different strategies for the use of adjoint physics with the FSU global spectral model. *Monthly Weather Rev* 128(3):668–688
- Lions JL (1968) Contrôle optimal de systemes gouvernés par des équations aux dérivées partielles. Gauthier-Villars, Paris
- Lions JL (1971) Optimal control of systems governed by partial differential equations, vol 170. Springer

- Liu DC, Nocedal J (1989) On the limited memory BFGS method for large scale optimization. *Math Progr* 45(1–3):503–528
- Lorenc AC (1986) Analysis methods for numerical weather prediction. *Royal Meteorol Soc Q J* 112:1177–1194
- Lorenc AC (2003) The potential of the ensemble Kalman filter for NWP—a comparison with 4D-Var. *QJR Meteorol Soc* 129:3183–3203
- Lugbenberger DG (1984) Linear and nonlinear programming. Addison-Wesley, New York
- Lupu C, Cardinali C, McNally T (2014) Evaluation of observation impact and observation error covariance retuning. Presented at The World Weather Open Science Conference, Montreal Canada
- Marchuk GI (1982) Mathematical issues of industrial effluent optimization. *J Meteorol Soc Jpn* 60(1):481–485
- McLaughlin D, O'Neill A, Derber J, Kamachi M (2005) Opportunities for enhanced collaboration within the data assimilation community. *Q J Royal Meteorol Soc* 131(613):3683–3694
- Navon IM, De Villiers R (1986) GUSTAF: a Quasi-Newton nonlinear ADI FORTRAN IV program for solving the Shallow-Water equations with augmented Lagrangians. *Comput Geosci* 12(2):151–173
- Navon IM, Zou X, Derber J, Sela J (1992) Variational data assimilation with an adiabatic version of the nmc spectral model. *Monthly Weather Rev* 120(7):1433–1446
- Navon IM (1981) Implementation of a posteriori methods for enforcing conservation of potential enstrophy and mass in discretized shallow-water equations models. *Monthly Weather Rev* 109(5):946–958
- Navon IM, De Villiers R (1983) Combined penalty multiplier optimization methods to enforce integral invariants conservation. *Monthly Weather Rev* 111(6):1228–1243
- Navon IM, Legler DM (1987) Conjugate-gradient methods for large-scale minimization in meteorology. *Monthly Weather Rev* 115(8):1479–1502
- Park SK, Županski D (2003) Four-dimensional variational data assimilation for mesoscale and storm-scale applications. *Meteorol Atmos Phys* 82(1–4):173–208
- Parrish DF, Derber JC (1992) The National Meteorological Center's spectral statistical-interpolation analysis system. *Monthly Weather Rev* 120(8):1747–1763
- Pontryagin LS, Boltyanskii VG, Gamkrelidze RV, Mishchenko E (1962) The mathematical theory of optimal processes (international series of monographs in pure and applied mathematics). Interscience, New York
- Powell MJD (1982) Nonlinear optimization. In: Powell MJD (ed) NATO conference series. Series II: systems science, proceedings of the NATO Advanced Research Institute, held at Cambridge (UK), vol 1. Academic Press, London
- Rabier F (2005) Overview of global data assimilation developments in numerical weather-prediction centres. *Q J Royal Meteorol Soc* 131(613):3215–3233
- Rabier F, Järvinen H, Klinker E, J-F Mahfouf J-F, Simmons A (2007) The ecmwf operational implementation of four-dimensional variational assimilation. I: experimental results with simplified physics. *Q J Royal Meteorol Soc* 126(564):1143–1170
- Ritz W (1908) Über eine neue Methode zur Lösung gewisser Variationsprobleme der mathematischen Physik. *J Reine Phys Math* 135:1–61
- Saad Y (1994) Sparsekit: a basic tool kit for sparse matrix computations. Technical Report, Computer Science Department, University of Minnesota
- Saad Y (2003) Iterative methods for sparse linear systems, 2nd edn. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA
- Sasaki Y (1958) An objective analysis based on the variational method. *J Meteorol Soc Jpn* 36(3):77–88
- Sasaki Y (1969) Proposed inclusion of time variation terms, observational and theoretical, in numerical variational objective analysis. *J Meteorol Soc Jpn* 47(2):115–124
- Sasaki Y (1970a) Numerical variational analysis formulated under the constraints as determined by longwave equations and a low-pass filter. *Monthly Weather Rev* 98(12):884–898

- Sasaki Y (1970b) Some basic formalisms in numerical variational analysis. *Monthly Weather Rev* 98(12):875–883
- Sasaki Y (1976) Variational design of finite-difference schemes for initial value problems with an integral invariant. *J Comput Phys* 21(3):270–278
- Sasaki Y, Gu P, Yan L (1955) A fundamental study of the numerical prediction based on the variational principle. *J Meteorol Soc Jpn* 33(6):262–275
- Shanno DF, Phua KH (1980) Remark on algorithm 500-a variable method subroutine for unconstrained nonlinear minimization. *ACM Trans Math Softw* 6:618–622
- Shutyaev V, Gejadze I, Copeland GJM, Le Dimet F-X (2012) Optimal solution error covariance in highly nonlinear problems of variational data assimilation. *Nonlinear Process Geophys* 19(2):177–184
- Simmons AJ, Hollingsworth A (2002) Some aspects of the improvement in skill of numerical weather prediction. *Q J Royal Meteorol Soc* 128(580):647–677
- Ştefănescu R, Navon IM (2013) POD/DEIM Nonlinear model order reduction of an ADI implicit shallow water equations model. *J Comput Phys* 237:95–114
- Ştefănescu R, Sandu A, Navon IM (2015) POD/DEIM reduced-order strategies for efficient four-dimensional variational data assimilation. *J Comput Phys* 295:569–595
- Stephens JJ (1966) A variational approach to numerical weather analysis and prediction. PhD thesis, Texas A & M University, College-Station, TX, 77863
- Stephens JJ (1968) Variational resolution of wind components. *Monthly Weather Rev* 96:229–231
- Trémölet Y (2007) Incremental 4d-var convergence study. *Tellus A* 59(5):706–718
- Trémölet Y (2008) Computation of observation sensitivity and observation impact in incremental variational data assimilation. *Tellus A* 60(5):964–978
- Tsuyuki T (1997) Variational data assimilation in the Tropics using precipitation data. Part III: assimilation of SSM/I precipitation rates. *Monthly Weather Rev* 125(7):1447–1464
- Veersé F, Thepaut J-N (1998) Multiple-truncation incremental approach for four-dimensional variational data assimilation. *Q J Royal Meteorol Soc* 124(550):1889–1908
- Vieux BE, LeDimet F, Armand D (1998) Inverse problem formulation for spatially distributed river basin model calibration using the adjoint method. *EGS, Annales Geophysicae, Part II, Hydrology, Oceans and Atmosphere, Supplement II to 16*:C501
- Wahba G (1975) Smoothing noisy data with spline functions. *Numerische Mathematik* 24(5):383–393
- Wahba G (1981a) Some new techniques for variational objective analysis on the sphere using splines, hough-functions, and sample spectral data. In: Proceedings of 7th conference on probability and statistics in the atmospheric sciences, Monterey, California, pp 213–216
- Wahba G (1981b) Spline interpolation and smoothing on the sphere. *SIAM J Sci Stat Comput* 2(1):5–16
- Wahba G, Wendelberger J (1980) Some new mathematical methods for variational objective analysis using splines and cross validation. *Monthly Weather Rev* 108(8):1122–1143
- Washington WM, Duquet RJ (1963) An objective analysis of stratospheric data by Sasaki's method. Department of Meteorology, Penn. State University, University Park, PA, 16802:23
- Whitaker JS, Hamill TM (2002) Ensemble data assimilation without perturbed observations. *Monthly Weather Rev* 130(7):1913–1924
- Wu L, Le Dimet F-X, De Reffye P, Hu B-G, Cournède P-H, Kang M-Z (2012) An optimal control methodology for plant growth-case study of a water supply problem of sunflower. *Math Comput Simul* 82(5):909–923
- Županski D, Mesinger F (1995) Four-dimensional variational assimilation of precipitation data. *Monthly Weather Rev* 123(4):1112–1127
- Županski D, Županski M, Rogers E, Parrish DF, DiMego GJ (2002) Fine-resolution 4DVAR data assimilation for the Great Plains tornado outbreak of 3 May 1999. *Weather Forecast* 17(3):506–525

Data Assimilation for Coupled Modeling Systems

Milija Županski

Abstract Coupled numerical models address the interaction between processes in the atmosphere, ocean, land surface, biosphere, chemistry, cryosphere, and hydrology. Including interaction between such processes can potentially extend the predictability and eventually help in reducing the uncertainty of the prediction. Coupled data assimilation is a branch of data assimilation that deals with coupled modeling systems. In this article the fundamentals of coupled data assimilation are described. Challenges of coupled data assimilation are addressed in terms of the variational and ensemble methods, with implications for hybrid data assimilation methods. Several illustrative examples of coupled data assimilation of a single observation with realistic regional coupled modeling systems are included as well.

1 Introduction

Use of coupled modeling systems is important for improving the predictability of coupled processes. The components of coupled modeling system may include atmosphere, ocean, land surface, biosphere, cryosphere, hydrology, chemistry, aerosol, as well as other processes of interest. The role of data assimilation for coupled modeling systems (e.g., coupled data assimilation) is also important as it has a capability to utilize the information from various observation types across the components to improve the prediction by coupled models.

Due to their increased complexity, coupled modeling systems are potentially more sensitive to the initial conditions than the individual components, and thus require special attention in data assimilation (Sakaguchi et al. 2012). In general one

M. Županski (✉)

Cooperative Institute for Research in the Atmosphere, Colorado State University, Fort Collins, CO 80523-1375, USA

e-mail: milija.zupanski@colostate.edu

can distinguish three main regimes of coupled data assimilation systems: (a) uncoupled, (b) weakly coupled, and (c) strongly coupled.

Uncoupled data assimilation implies that each system is completely independent: the model and data assimilation for each individual component is done separately from each other. Using as an example atmosphere and hydrology, the atmospheric model forecast and data assimilation are used to produce the precipitation, which can be then used as forcing in the hydrology model with its own, independent data assimilation system.

Weakly coupled data assimilation allows some interaction between components. Typically, this means that data assimilation is performed completely separately for each component, but their initial conditions and parameters are fed back into the coupled model. In terms of sequential data assimilation, which consists of the forecast and the analysis steps, the weakly coupled data assimilation implies coupling in the forecast step, but no coupling in the analysis step. Most importantly, there is no use of the cross-component part of the forecast error covariance in data assimilation.

Strongly coupled data assimilation means that both the coupled forecast and coupled data assimilation are done in a single system that combines the information from all components. The cross-component part of the forecast error covariance is used in data assimilation, thus the forecast and observations of each component have a potential to influence all other components. In this paper we focus on strongly coupled data assimilation, as it offers a more profound, but also a more natural way of combining the information from models and observations.

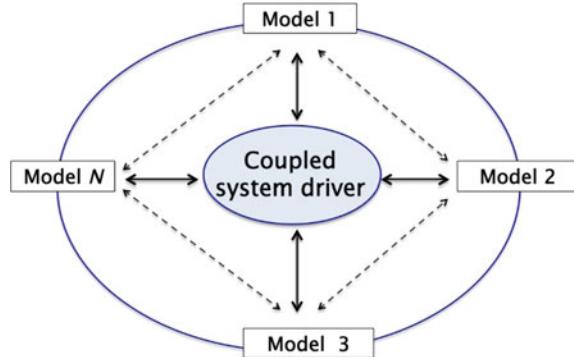
Although limited in number, the coupled data assimilation systems have already been developed (Sugiura et al. 2008; Rasmy et al. 2012; Han et al. 2013). They are mostly based on variational data assimilation, as most operational weather and climate data assimilation systems are variational, but there are also few ensemble coupled data assimilation systems being developed (Zhang et al. 2007; Tardif et al. 2014).

The paper is organized as follows. We begin by describing the motivation for coupled data assimilation in Sect. 2, followed by the challenges of coupled data assimilation in Sect. 3. An example of the two-component coupled data assimilation is presented in Sect. 4, followed by realistic applications of coupled data assimilation to investigate the coupled forecast error covariance and uncertainty interaction in Sect. 5. The summary and future development are presented in Sect. 6.

2 Motivation

Main motivation for developing coupled data assimilation is to improve the knowledge and/or improve the prediction of a coupled modeling system, in terms of the state and its uncertainty. The benefit of having a combined modeling system, as opposed to having the individual systems only, is that the model components are

Fig. 1 Schematic representation of an N -component coupled modeling system. The *full lines* represent the mandatory interaction between the model component and the driver, while the *dashed lines* represent the optional interaction between components



acting as “constraints” that restrict possible adjustments of the state to those that are “allowed” by other system components. In other words, it is anticipated that the analysis adjustment in the coupled data assimilation is in better physical agreement with the real world than the analysis adjustments in data assimilation of individual components.

A coupled modeling system is schematically represented in Fig. 1. While the interaction between the driver and the model components is always present, the interaction between model components is optional, related to the degree of dependence between relevant physical processes.

Another useful view of coupled data assimilation that supports its use can be described in terms of Shannon information theory (Shannon and Weaver 1949). Recall that the entropy is defined as

$$H(X) = - \int_x p(x) \log p(x) dx \quad (2.1)$$

where p denotes the probability density function (pdf) and x is a random variable. The entropy is closely associated with uncertainty, which is important for data assimilation. In coupled systems we are interested in the joint entropy of at least two processes, X_1 and X_2 ,

$$H(X_1, X_2) = - \int_{x_1} \int_{x_2} p(x_1, x_2) \log p(x_1, x_2) dx_1 dx_2 \quad (2.2)$$

where $p(x_1, x_2)$ is the joint pdf. An important quantification of the information exchange in coupled systems can be defined in terms of mutual information

$$I(X_1; X_2) = - \int_{x_1} \int_{x_2} p(x_1, x_2) \log \left[\frac{p(x_1, x_2)}{p(x_1)p(x_2)} \right] dx_1 dx_2 \quad (2.3)$$

where $p(x_1)$ and $p(x_2)$ are the marginal probability density functions. Mutual information measures the information shared by multi-component coupled system. It can be conveniently represented in terms of entropy (e.g., Cover and Thomas 2006)

$$I(X_1; X_2) = H(X_1) + H(X_2) - H(X_1, X_2) \quad (2.4)$$

Since $H(X_1, X_2) \leq H(X_1) + H(X_2)$, with equality sign true for independent processes X_1 and X_2 , the mutual information is non-negative with zero value corresponding to independent processes. Given that the motivation of developing coupled systems modeling is to improve the interaction between related physical processes, the component processes of the coupled system have well developed dependence and are intrinsically characterized by positive mutual information. This implies that the information from one component enhances the information about other components. This is quite important for data assimilation of coupled systems since it effectively reduces the dimension of the forecast error covariance and allows cross-component impacts during data assimilation, as will be shown in detail in the next sections.

When coupling exists, one can further partition it in terms of its strength: low intensity coupling referring to having only marginal interaction between the components, and high intensity coupling implying a profound interaction between the components. This qualitative distinction can be in principle quantified by calculating the mutual information of coupled modeling system, i.e. using their prior and joint probability density functions. Using as an example the Gaussian pdf, one can define the mutual information of multivariate Gaussian probability distribution (Silva and Quiroz 2003)

$$I(X_1, X_2) = -\frac{1}{2} \ln \frac{\det(\Sigma)}{\det(\Sigma_{11}) \det(\Sigma_{22})} \quad (2.5)$$

where Σ , Σ_{11} , and Σ_{22} are the joint, and the marginal covariances of variables X_1 and X_2 , respectively, and \det denotes the determinant of a matrix. Since the joint covariance is

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{pmatrix} \quad (2.6)$$

and (e.g., Arellano-Valle et al. 2012)

$$\det(\Sigma) = \det(\Sigma_{11}) \det(\Sigma_{22}) \det(I - \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^T) \quad (2.7)$$

one can define the mutual information as

$$I(X_1, X_2) = -\frac{1}{2} \ln [\det(I - \Sigma_{11}^{-1} \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{12}^T)] \quad (2.8)$$

The above equation implies that the mutual information can be calculated from the auto- and cross-components of the coupled system error covariance. It is also clear that, if the cross-component covariance does not exist in Eq. (2.6) ($\Sigma_{12} = 0$), the mutual information is zero.

Depending on the coupled system's complexity and dimensions, in some applications it may be possible to calculate the exact value of the mutual information according to (2.8), or an approximate value using only the diagonal elements of the matrices. By defining a mutual information threshold for low/high intensity of coupling, it may be possible to have a quantitative measure of the strength of coupling.

3 Challenges

There are numerous challenges of coupled data assimilation, some of which will be discussed here. They include: control variable, spatiotemporal scales, forecast error covariance, high state dimensions, and non-Gaussian errors.

3.1 *Control Variable*

In principle, control variable refers to modeling system parameters that impact its prediction. In particular, the prediction has to be sensitive to the choice of these parameters in the spatiotemporal limits of the data assimilation window. In practical data assimilation one can define control variable different from the state variable. For example, in some data assimilation systems it is assumed that the control variable includes stream function and velocity potential, while the state (model) variable includes the east-west and north-south wind components (e.g., u and v , respectively). However, for simplicity of presentation, we will assume that the control variable is a simple subset of the state variable, without requiring any transformation.

In geophysical modeling based on using partial differential equations the control variable typically includes (i) initial conditions, (ii) lateral boundary conditions (for regional systems), (iii) model empirical parameters, and (iv) model and observation operator biases (systematic errors). Therefore, for the k -th modeling component the control variable can be written as

$$x_k = (x_k^{ic} \quad x_k^{bias} \quad x_k^{par})^T \quad k = 1, \dots, N \quad (3.1)$$

where the superscripts *ic*, *bias*, and *par* refer to the initial conditions, bias, and empirical parameters, respectively, and N is the number of coupled system components. Note that this formulation includes both the single and the coupled systems, $N=1$ corresponding to the single system, and $N \geq 2$ corresponding to the coupled system. The general form of coupled data assimilation control variable is

$$x = (x_1 \quad x_2 \quad \dots \quad x_N)^T. \quad (3.2)$$

The coupled data assimilation challenge comes from the fact that practical data assimilation for individual components does not include all possible control parameters as in (3.2). Most often the control variable is defined as the initial conditions. However, there are data assimilation systems that are focused on adjusting empirical parameters (e.g., climate), empirical parameters and initial conditions (hydrology), model bias (carbon cycle, biosphere), rather than the initial conditions only. This creates a need for developing a coupled data assimilation system that is general enough so that it can assimilate any type of control variable. In practice, the control variable would include relevant control variables for each component. For example, if the coupled system includes the following components: atmospheric—denoted *a*, carbon transport—denoted *c*, land surface—denoted *l*, and hydrological—denoted *h*, one could define control variable as

$$x = (a^{ic} \quad c^{bias} \quad l^{ic} \quad l^{par} \quad h^{par})^T. \quad (3.3)$$

Such control variable may be a very efficient way of updating the given coupled system, but it does require a mathematical apparatus that can handle it in practice. If the available algorithmic structure of the coupled data assimilation does not allow the inclusion of all these variables, one may reduce the list to include only the feasible parameters. A better solution would be to update the coupled data assimilation algorithm by adding the capability to augment the control variable.

3.2 Forecast Error Covariance

The forecast error covariance is intrinsically related to the choice of control variable. The forecast error covariance represents the uncertainty of the control variable and thus includes all control variable components. Forecast error covariance has a fundamental role in data assimilation (e.g., Bannister 2008a, b) since all analysis adjustments are projected onto the subspace of this matrix. There are numerous papers addressing the role of error covariance in variational and ensemble data assimilation (Hollingsworth and Lonnberg 1986; Derber and Bouttier 1999; Buehner 2005; Bello Pereira and Berre 2006; Berre and Desrozier 2010).

Given a general coupled control variable (3.2), the forecast error covariance is a matrix

$$P_f = \begin{pmatrix} P_{11} & P_{12} & \cdots & P_{1N} \\ P_{12}^T & P_{22} & & P_{2N} \\ \vdots & & \ddots & \vdots \\ P_{1N}^T & P_{2N}^T & \cdots & P_{NN} \end{pmatrix} \quad (3.4)$$

where the subscripts refer to the coupled system component. Note that, according to (3.1), each of the inputs in (3.4) is also a matrix. For example

$$P_{ij} = \begin{pmatrix} P_{ij}^{ic,ic} & P_{ij}^{ic,bias} & P_{ij}^{ic,par} \\ (P_{ij}^{ic,bias})^T & P_{ij}^{bias,bias} & P_{ij}^{bias,par} \\ (P_{ij}^{ic,par})^T & (P_{ij}^{bias,par})^T & P_{ij}^{par,par} \end{pmatrix} \quad (3.5)$$

where i and j define the system components. It is clear that defining the elements of such complex matrix becomes challenging.

In variational data assimilation, the error covariance is modeled using some previous knowledge about correlations and variances. Knowing a priori the correlations between initial conditions, empirical model parameters, and model biases is very difficult. In ensemble data assimilation the correlations are obtained directly from the ensemble forecast, without the need for modeling the correlation functions. However, ensemble data assimilation requires at least some knowledge about the true correlations in order to know if low-dimensional approximation of the ensemble error covariance is acceptable.

Such challenges exist even in assimilation of individual components; in coupled systems this issue is magnified and thus made more difficult. The most challenging aspect of defining the error covariance in coupled data assimilation is to define the cross-component correlations, since the cross-variable correlations, and in particular the cross-component correlations are the least known.

3.3 High-Dimensional State Vector

The dimension of the state vector can considerably impact the design and performance of coupled data assimilation. Realistic atmospheric data assimilation systems have the control variable of the order of hundreds of millions. Adding chemistry component, for example, can further double the dimension of state vector since there may be several chemical species that are adjusted in data assimilation. Similarly, the aerosol can also add considerably to the total state dimension since it is defined at all model grid points, and there can be a dozen of aerosol species, such as dust, sea salt, black carbon, etc. Some other components, such as land surface, may

add only a smaller dimension to the augmented state since typically there are fewer soil layers than atmospheric layers.

In general, coupled data assimilation system has to be able to deal with such high dimensions. This may imply that new approximations may be necessary when extending the single-component assimilation to the coupled data assimilation. Practical data assimilation algorithms often include approximations such as dual-resolution (i.e. coarse resolution adjoint or ensemble and a high-resolution control), or a related incremental variational approach with additional linearization. In cases when one coupled component is more nonlinear than others, or there is a highly nonlinear observation operator of one of the components, it may be necessary to re-validate the assumptions in the context of the coupled system.

3.4 *Non-gaussian Errors*

It is well known that some control variables can have non-Gaussian errors, although typical data assimilation is based on Gaussian error assumption. If there is a small number of such variables, this may impact the results of assimilation only locally. However, this may become more relevant for coupled systems, as one component can have all non-Gaussian variables. For example, typical atmospheric chemistry and aerosol variables are strictly positive definite, which effectively excludes the Gaussian distribution with infinite tails. Therefore, in case of coupled atmosphere-chemistry-aerosol data assimilation, there may be a need to revisit the Gaussian error assumptions. In more extreme cases, it may be necessary to design a mixed Gaussian-non-Gaussian data assimilation system to accommodate all coupling components.

3.5 *Spatiotemporal Scales*

Another potential issue in coupled data assimilation is related to characteristic spatiotemporal scales of control variables. For example, atmospheric variables likely have shorter temporal scales than ocean variables, or compared to some chemistry variables such as stratospheric ozone. In the coupled system forecast, this can be resolved by using different time steps for different components of the system. For data assimilation, however, this may be more complicated as it should involve the coupled forecast error covariance, given its fundamental role in data assimilation.

In variational data assimilation the modeled error covariance has to include some knowledge of spatiotemporal scales in order to produce a dynamically balanced increments of both components. Since the knowledge of such cross-component correlations is very limited in general and possibly situation-dependent, the use of complete cross-component correlations in practical variational data assimilation is

not very likely. The use of ensemble error covariance may be simpler since the information about different spatiotemporal scales is already included in the coupled forecast models used in ensemble forecasting. However, since both of these approaches can produce only an approximate error covariance, there is still a need for verifying the validity of the coupled error covariance in terms of the spatiotemporal scales. The resolution may be more intuitive in data assimilation with four-dimensional error covariance, since than one can impose the temporal aspect of error covariance more directly.

4 Two-Component Coupled System Data Assimilation

The role of coupled forecast error covariance is now examined in the context of an idealized two-component coupled modeling system. Let define the two state components as x_1 and x_2 , thus forming a two dimensional state vector of the coupled system

$$x = (x_1 \quad x_2)^T \quad (4.1)$$

In this system each component is represented by a value at a single grid point. For the land-atmosphere system, for example, one grid point would be located in the atmosphere and one in the soil. For the atmosphere-chemistry coupled system, both components can represent the value at the same grid point. The forecast step of such system includes a coupled prediction model (denoted m)

$$x^n = m(x^{n-1}) \quad (4.2)$$

where the superscript n defines time. We are now interested in performing the analysis at time n . In coupled data assimilation, the cost function formally appears the same as in any other data assimilation. Assuming sequential data assimilation for simplicity, the cost function for a general two-component system can be formally defined in terms of its components

$$\begin{aligned} J(x_1, x_2) = & \frac{1}{2} \begin{pmatrix} x_1 - x_1^f \\ x_2 - x_2^f \end{pmatrix}^T \begin{pmatrix} P_{11} & P_{12} \\ P_{12}^T & P_{22} \end{pmatrix}^{-1} \begin{pmatrix} x_1 - x_1^f \\ x_2 - x_2^f \end{pmatrix} + \\ & + \frac{1}{2} [y_1 - h_1(x_1)]^T R_1^{-1} [y_1 - h_1(x_1)] + \frac{1}{2} [y_2 - h_2(x_2)]^T R_2^{-1} [y_2 - h_2(x_2)] \end{aligned} \quad (4.3)$$

where h is the nonlinear observation operator, y is the observation and R is the observation error covariance. The subscripts 1 and 2 refer to the first and second state vector components, respectively. It is implied in (4.3) that forecast error covariance can have cross-component correlations, but that observation errors are

independent. A general linear solution to the problem (4.3) can be obtained after setting $\nabla J = 0$ (e.g., Lorenc 1986)

$$x^a = (I + P_f H^T R^{-1} H)^{-1} (x^f + P_f H^T R^{-1} y) \quad (4.4)$$

with

$$x^a = \begin{pmatrix} x_1^a \\ x_2^a \end{pmatrix} \quad x^f = \begin{pmatrix} x_1^f \\ x_2^f \end{pmatrix} \quad y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \quad (4.5)$$

where H is the Jacobian of the observation operator. Now let assume that there is only a single observation of one of the components, say the component 1, defined at that grid point. Then the cost function (4.3) becomes

$$\begin{aligned} J(x_1, x_2) = & \frac{1}{2} \begin{pmatrix} x_1 - x_1^f \\ x_2 - x_2^f \end{pmatrix}^T \begin{pmatrix} (\sigma_f^2)_1 & \rho_{12} \\ \rho_{12} & (\sigma_f^2)_2 \end{pmatrix}^{-1} \begin{pmatrix} x_1 - x_1^f \\ x_2 - x_2^f \end{pmatrix} + \\ & + \frac{1}{2} [y_1 - x_1]^T (\sigma_o^2)_1^{-1} [y_1 - x_1] \end{aligned} \quad (4.6)$$

where σ denotes the standard deviation, the cross-component error covariance (directly related to the correlation between the coupling components) is denoted ρ , and indexes f and o refer to the forecast and the observation, respectively. The analysis solution (4.4) and (4.5) for the cost function (4.6) in terms of the components is

$$x_1^a = \frac{1}{1 + \varepsilon_1^2} x_1^f + \frac{\varepsilon_1^2}{1 + \varepsilon_1^2} y_1 \quad (4.7)$$

$$x_2^a = x_2^f + \frac{1}{1 + \varepsilon_1^2} \frac{\rho_{12}}{(\sigma_o^2)_1} (y_1 - x_1^f) \quad (4.8)$$

where ε represents the ratio between forecast and observation standard deviations

$$\varepsilon_1 = \frac{(\sigma_f)_1}{(\sigma_o)_1}. \quad (4.9)$$

The solution (4.7) and (4.8) illustrates several important aspects of coupled data assimilation. When observing only one (the first) component, the analysis is identical to the de-coupled assimilation of the first component (e.g., (4.7)). This means that, although the coupled covariance has cross-component correlations, the analysis of the first component does not benefit from the coupling. Although the second component is not observed, the analysis of the second component is impacted by the observation of the first component (e.g., (4.8)). A closer inspection reveals that this was possible only because of the non-zero cross-component error

covariance ρ_{12} , assuming $y_1 \neq x_1^f$. This is a very important result of coupled data assimilation, as it indicates that it is possible to change the analysis of a coupled component even if it is not observed. There are often instances when one component, e.g., land surface, is not well observed, while the other component, e.g., atmosphere, is well observed. The above results suggest that the land surface control variables can still be adjusted in the analysis. Note that this would not be possible in the stand-alone land surface analysis system.

5 Structure of Coupled Forecast Error Covariance

In this section we conduct single observation experiments in order to understand and illustrate the structure of coupled forecast error covariance. As suggested above, the true power of coupled data assimilation comes from the cross-component correlations. They allow a more efficient use of observations by impacting all control variables, and thus produce a more balanced change of control variables. The structure of forecast error covariance can be assessed using a single observation experiment setup, where only one observation at a single point is assimilated (e.g., Thepaut et al. 1996; Whitaker et al. 2009). We apply the Maximum Likelihood Ensemble Filter (MLEF—Županski 2005; Županski et al. 2008) as a coupled data assimilation system. In all experiments we use 32 ensembles. Since this system is employing the ensemble coupled error covariance, there is no need to model the cross-components as they are created automatically by the ensemble forecasting.

As a first example we consider a land-atmosphere coupling by using the NASA Unified Weather Research and Forecasting (WRF) model (Peters-Lidard et al. 2015) that has an implicit coupling between Noah land surface model and the WRF atmospheric component. The horizontal model resolution of the parent domain is 27 km with the nest at 9 km. There are 31 vertical layers in the atmosphere, and 4 vertical layers in the soil. The control variable includes atmospheric variables (perturbation surface pressure, perturbation potential temperature, perturbation geopotential, horizontal winds, specific humidity) with clouds (cloud ice, cloud snow, cloud rain, cloud water, graupel) as well as land surface variables (soil moisture, soil temperature). We assimilate a pseudo-observation of cloud rain at 700 hPa and evaluate its impact in the analysis, in particular on land surface variables.

The analysis increments ($x^a - x^f$) for cloud rain and soil moisture are shown in Fig. 2. One can note that auto-correlation of cloud rain is adequately represented by the ensemble, indicating the maximum near the observed location and the response in the lower troposphere spreading down to surface (Fig. 2a). From the coupled data assimilation point of view, the analysis increments of soil variables are more interesting (e.g., Fig. 2b). First, one can see that the soil moisture analysis increments are non-zero, implying the existence of the cross-component correlation between cloud rain and soil moisture. The impact of cloud rain pseudo-observation has spread into all soil layers, with stronger impact at the near-surface layers.

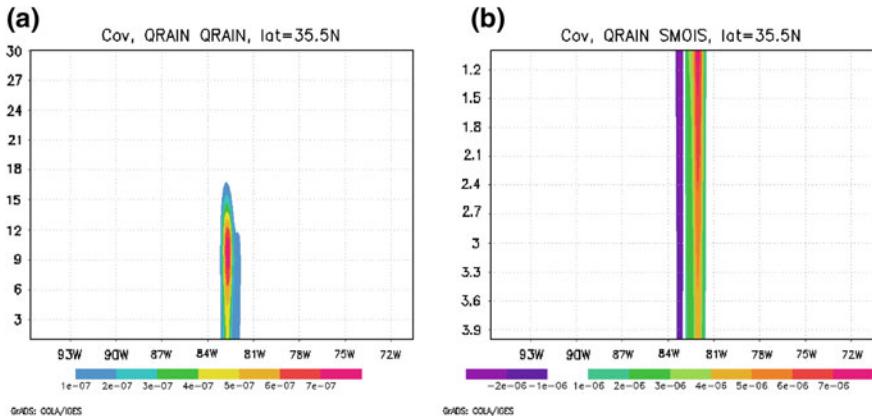


Fig. 2 Vertical cross-section of analysis increments $x^a - x^f$ for **a** cloud rain (kg/kg), and **b** soil moisture (kg/kg), as a response to a single pseudo-observation of cloud rain at 700 hPa. Note that the *vertical* index increases downwards for soil moisture, thus defines the depth. The *vertical* location of single observation is near level 11

A possible physical interpretation of these results is that an increase of cloud rain at 700 hPa induces an increase of moisture at the surface, which eventually spreads into the soil causing an increase of soil moisture. A careful investigation of the location of these impacts shows that the analysis increments of soil moisture lags the analysis increment of the atmospheric variable by about 60 km, with a shift towards the east. This suggests that the forecast error does not represent an instantaneous response, which would be characterized by a response at the same location. Rather, it reflects the interactions between atmosphere and soil as represented by the coupled modeling system, in this case characterized by a delay of land surface response. This potentially relates to the issue of spatiotemporal scale interaction in the coupled system (Sect. 3.5), but certainly requires further evaluation.

As a second example we consider a coupled atmosphere-chemistry model WRF-Chem (Grell et al. 2005). The control variables include the standard atmospheric variables and five species of chemical constituents (o3, so2, so3, no2, no3). The single observation experiments that examine the structure of the coupled atmosphere-chemistry error covariance have been published in Park et al. (2015). Here we present several additional cross-component correlations that confirm the inherent complexity of the coupled forecast error covariance. As in Park et al. (2015), we assimilate the Ozone Monitoring Instrument (OMI) total column ozone at a single point. In Fig. 3 we present the additional cross-component correlation that such atmosphere-chemistry coupled system describes, in terms of specific humidity (Fig. 3a) and east-west wind component (Fig. 3b). One can notice that both atmospheric components have well-defined analysis increments in both vertical and horizontal directions. Specific humidity has somewhat smaller magnitudes than the wind. It is also interesting to note that analysis responses for atmospheric

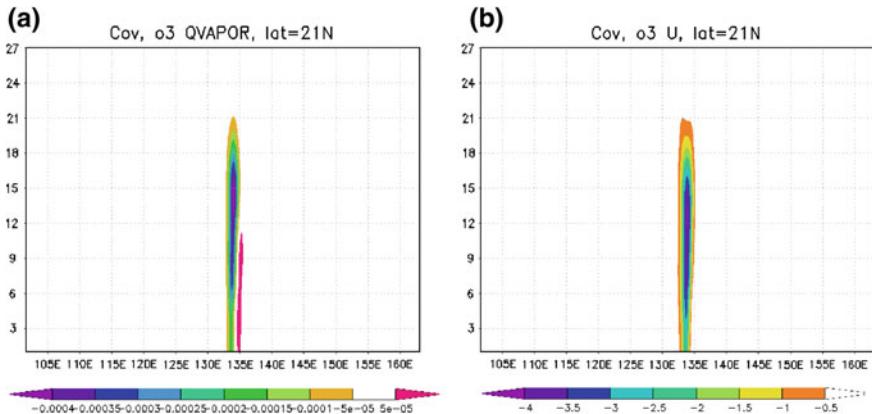


Fig. 3 Vertical cross-section of analysis increments $x^a - x^f$ for **a** specific humidity (kg/kg), and **b** east-west wind component (m/s), as a response to a single observation of total column ozone. The single observation location is near level 24

variables are located in middle and lower troposphere, while the maximum ozone response is at 250 hPa (e.g., Park et al. 2015). These analysis increments confirm that the cross-component covariance holds important information that could potentially benefit coupled atmosphere-chemistry data assimilation.

Lastly, we consider the experiment with coupled atmosphere-aerosol-chemistry model, which is the WRF-Chem model with the Goddard Chemistry Aerosol Radiation and Transport (GOCART) aerosol model (Chin et al. 2000) that includes the prediction of sulfates, black carbon, organic carbon, sea salt and dust. In particular, we investigate the correlations between ozone and dust. These correlations are largely unknown, but this example suggests that they do contain information that could be made useful in coupled data assimilation. The augmented control variable includes atmospheric variables, chemistry variables, and the aerosol variables (dust) at 0.5, 1.4, 2.4, 4.5 and 8.0 μm . As in the previous example we assimilate a single point OMI total column ozone observation. The analysis increments for selected dust variables are shown in Fig. 4. The cross-component correlation between ozone and dust indicates that the maximum analysis response of dust to ozone observation is at 300-400 hPa. One can again notice well-defined analysis increments, which suggests that the cross-component correlations are not a noise; rather, they appear to be a signal with important information about the coupled system uncertainties. A closer inspection of Fig. 4a, b shows that, although the responses are similar, larger dust particles have the maximum response at lower levels than small particles. Also, the magnitude of the increments is larger for smaller dust particles.

In overview, the above results indicate that the structure of coupled ensemble forecast error covariance is complex and contains important cross-component correlations, with potential benefit for coupled data assimilation.

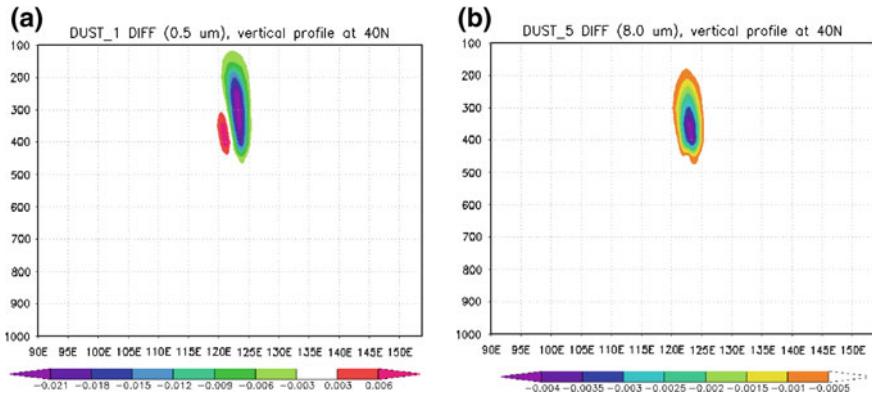


Fig. 4 Vertical cross-section of analysis increments $x^a - x^f$ in response to a single OMI total column ozone observation for **a** dust-1 (0.5 μm), and **b** dust-5 (8.0 μm)

6 Summary and Future

Coupled data assimilation is an important component of the coupled system modeling. It holds a great potential for improving the forecast of various Earth science components, as well as for better understanding of the coupled systems' state and uncertainty.

The coupled data assimilation formally represents a system very similar to single component data assimilation, however with increased difficulty. Main challenges that make coupled data assimilation difficult are associated with the structure of the forecast error covariance, the augmented control variable, the increased state dimensions, potentially non-Gaussian errors, and the interactions between the coupled components characterized by different spatiotemporal scales.

A simple two-component coupled system analysis indicates the relevance of the cross-component correlations. In particular, the cross-component correlations have a potential to increase the utility of observations in data assimilation by spreading the information throughout the components. The conducted single observation assimilation experiments confirm that the structure of cross-component correlations is complex and clearly related to the dynamical links between the coupled components and their control variables.

Numerous challenges of coupled data assimilation are still remaining, in particular related to the use of hybrid variational-ensemble systems, as the cross-component correlations of hybrid error covariance, coming from both the variational and ensemble methodologies, need to be reconciled. However, given the relevance and the increased interest in performing the forecasts with coupled modeling systems, the role of coupled data assimilation and its benefits will likely steadily increase.

Acknowledgements The author gratefully acknowledges support from the NASA Modeling, Analysis, and Prediction (MAP) Program Grant NNX13AO10G, the NASA Precipitation Measurement Mission (PMM) Program Grant NNX10AG92G, and the National Science Foundation Collaboration in Mathematical Geosciences Grant 0930265. The author would also like to acknowledge the computational support of NASA Advanced Supercomputing (NAS), and extend gratitude to the computing support from Yellowstone provided by NCAR's Computational and Information System Laboratory, sponsored by the National Science Foundation.

References

- Arellano-Valle RB, Contreras-Reyes JE, Genton MG (2012) Shannon entropy and mutual information for multivariate skew-elliptical distributions. *Scand J Statistics* 40:42–62
- Bannister RN (2008a) A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances. *Q J R Meteorol Soc* 134:1951–1970
- Bannister RN (2008b) A review of forecast error covariance statistics in atmospheric variational data assimilation. II: modelling the forecast error covariance statistics. *Q J R Meteorol Soc* 134:1971–1996
- Belo Pereira MB, Berre L (2006) The use of an ensemble approach to study the background error covariances in a global NWP model. *Mon Weather Rev* 134:2466–2489
- Berre L, Desroziers G (2010) Filtering of background error variances and correlations by local spatial averaging: a review. *Mon Weather Rev* 138:3693–3720
- Buehner M (2005) Ensemble derived stationary and flow dependent background error covariances: evaluation in a quasi-operational NWP setting. *Q J R Meteorol Soc* 131:1013–1043
- Chin M, Rood RB, Lin S-J, Muller JF, Thompson AM (2000) Atmospheric sulfur cycle in the global model GOCART: model description and global properties. *J Geophys Res* 105:24671–24687
- Cover TM, Thomas JA (2006) *Elements of information theory*. 2nd edn. John Wiley & Sons, Hoboken, New Jersey, 776 pp
- Derber J, Bouttier F (1999) A reformulation of the background error covariance in the ECMWF global data assimilation system. *Tellus* 51A:195–221
- Grell GA, Peckham SE, Schmitz R, McKeen SA, Frost G, Skamarock WC, Eder B (2005) Fully coupled “online” chemistry within the WRF model. *Atmos Environ* 39:6957–6975
- Han G, Wu X, Zhang S, Li W (2013) Error covariance estimation for Coupled Data Assimilation using a lorenz atmosphere and a simple pycnocline ocean model. *J Clim* 26:10218–10231
- Hollingsworth A, Lonnberg P (1986) The statistical structure of short- range forecast errors as determined from radiosonde data. Part I *Wind Field* *Tellus* 38A:111–136
- Lorenc A (1986) Analysis methods for numerical weather prediction. *Q J R Meteorol Soc* 112:1177–1194
- Park SK, Lim S, Županski M (2015) Structure of forecast error covariance in coupled atmosphere–chemistry data assimilation. *Geosci Model Dev* 8:1315–1320
- Peters-Lidard CD, Kemp EM, Matsui T, Santanello JA Jr, Kumar SV, Jacob JP, Clune T, Tao W-K, Chin M, Hou A, Case JL, Kim D, Kim K-M, Lau W, Liu Y, Shi J-J, Starr D, Tan Q, Tao Z, Zaitchik BF, Zavodsky B, Zhang SQ, Županski M (2015) Integrated modeling of aerosol, cloud, precipitation and land processes at satellite-resolved scales. *Environ Model Softw.* 67:149–159
- Rasmy M, Koike T, Kuria D, Mirza CR, Li X, Yang K (2012) Development of the Coupled Atmosphere and Land Data Assimilation System (CALDAS) and Its Application Over the Tibetan Plateau. *IEEE Trans Geosci Rem Sen* 50:4227–4242
- Sakaguchi K, Zeng X, Brunke MA (2012) The hindcast skill of the CMIP ensembles for the surface air temperature trend. *J Geophys Res* 117:D16113. doi:[10.1029/2012JD017765](https://doi.org/10.1029/2012JD017765)

- Shannon CE, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, 144 pp
- Silva C, Quiroz A (2003) Optimization of the atmospheric pollution monitoring network at Santiago de Chile. *Atmos Environ* 37:2337–2345
- Sugiura N, Awaji T, Masuda S, Mochizuki T, Toyoda T, Miyama T, Igarashi H, Ishikawa Y (2008) Development of a four-dimensional variational coupled data assimilation system for enhanced analysis and prediction of seasonal to interannual climate variations. *J Geophys Res* 113:C10017. doi:[10.1029/2008JC004741](https://doi.org/10.1029/2008JC004741)
- Tardif R, Hakim GJ, Snyder C (2014) Coupled atmosphere–ocean data assimilation experiments with a low-order climate model. *Clim Dyn* 43:1631–1643
- Thepaut J-N, Courtier P, Belaud G, Lemaitre G (1996) Dynamical structure functions in a four-dimensional variational assimilation: a case study. *Q J R Meteorol Soc* 122:535–561
- Whitaker JS, Compo GP, Thepaut J-N (2009) A comparison of variational and ensemble-based data assimilation systems for reanalysis of sparse observations. *Mon Weather Rev* 137:1991–1999
- Zhang S, Harrison MJ, Rosati A, Wittenberg A (2007) System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon Weather Rev* 135:3541–3564
- Županski M (2005) Maximum likelihood ensemble filter: theoretical aspects. *Mon Weather Rev* 133:1710–1726
- Županski M, Navon IM, Županski D (2008) The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. *Q J R Meteorol Soc* 134:1039–1050

Representer-Based Variational Data Assimilation Systems: A Review

Boon S. Chua and Liang Xu

Abstract This chapter reviews developments in representer-based variational data assimilation systems over the past 15 years. Data assimilation systems with representer-based algorithms are routinely used in operational and research centers for producing four-dimensional atmospheric and oceanic analyses and prediction. The systems reviewed in this chapter are the Inverse Ocean Modeling (IOM) system, the Naval Research Laboratory Atmospheric Variational Data Assimilation System-Accelerated Representer (NAVDAS-AR) system, the Navy Coastal Ocean Model 4D-Var (NCOM 4D-Var) system, and the Regional Ocean Modeling System 4D-Var (ROMS 4D-Var) system. These systems are mature operational or semi-operational weak-constraint, four-dimensional variational data assimilation systems. The emphasis here is on providing brief reviews with the key references related to the implementation and applications of these systems. Readers interested in early developments of representer-based systems (before 2002) are encouraged to look at Chua and Bennett (2001) and Bennett (2002).

1 Introduction

This chapter reviews recent developments associated with the application of the representer-based method (Bennett 1992) to weak-constraint, four-dimensional variational data assimilation (W4D-Var) systems, an approach which is the foundation of the so-called dual form of variational data assimilation (Courtier 1997). The four-dimensional variational data assimilation (4D-Var) is an estimation method that minimizes a quadratic cost function in the weighted least-squares sense

B.S. Chua (✉)

Science Applications International Corporation, Monterey, CA, USA
e-mail: boon.chua.ctr.my@nrlmry.navy.mil

L. Xu

Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

between a model state $x(t_0)$, at initial time t_0 , and a background field x_b (a prior estimate), and observations y . The cost function J is written

$$J[x(t_0)] = (x(t_0) - x_b)^T \mathbf{B}^{-1} (x(t_0) - x_b) + [y - H(x)]^T \mathbf{R}^{-1} [y - H(x)], \quad (1)$$

where H is the observation operator (also known as the forward operator), and \mathbf{B} and \mathbf{R} are the initial background and observation error covariance matrices, respectively. If the initial condition and observation errors are normally distributed with covariances \mathbf{B} and \mathbf{R} , the observation errors are unbiased, and the background field x_b is equal to the statistical mean of $x(t_0)$, then the minimizer of J is the maximum likelihood estimate of $x(t_0)$. Typically 4D-Var, strictly speaking strong constraint 4D-Var (also known as S4D-Var), also assumes that the dynamical model used is error free. Consequently, it requires that the four-dimensional analysis satisfies the model exactly. It is clear, however, that atmospheric and oceanic models are far from perfect (i.e. error free) because they contain other sources of error which must be considered. Specifically, there are errors in models such as boundary conditions and forcings. Weak-constraint (see Sasaki 1970) four dimensional variational data assimilation is a generalization of 4D-Var which permits one to estimate these additional errors, denoted f . The above cost function (1) is naturally generalized to

$$J[x(t_0), f] = (f - f_b)^T \mathbf{F}^{-1} (f - f_b) + (x(t_0) - x_b)^T \mathbf{B}^{-1} (x(t_0) - x_b) + [y - H(x)]^T \mathbf{R}^{-1} [y - H(x)] \quad (2)$$

where f_b is the values of the model forcing fields that are available, and \mathbf{F} is the model forcing error covariance matrix.

The dual formulation of variational data assimilation has been widely utilized in many data assimilation modeling studies in the atmosphere and ocean. The systems reviewed in this chapter are (1) the Inverse Ocean Modeling (IOM) system (Bennett et al. 2008), a Graphical User Interface (GUI)-driven system, (2) NAVDAS-AR system (Xu et al. 2005, Rosmond and Xu 2006, Xu et al. 2007, and Chua et al. 2009), an operational W4D-Var atmospheric data assimilation system for the United States Navy Global Environmental Model (NAVGEN) (Hogan et al. 2014), (3) NCOM 4D-Var system (Ngodock and Carrier 2013, 2014a), a W4D-Var ocean data assimilation system developed for the Navy Coastal Ocean Model (NCOM) (Martin 2000; Barron et al. 2006), and (4) the ROMS 4D-Var system (Moore et al. 2011a), a unique community ocean data assimilation system, developed for the Regional Ocean Modeling System (ROMS) (<http://www.myroms.org/>).

The outline of this paper is as follows. Section 2 describes the implementation and applications of the aforementioned representer-based variational data assimilation systems. The chapter ends with a summary in Sect. 3.

2 Systems

In this section we emphasize the implementation and applications of four representer-based variational data assimilation systems—the IOM in Sect. 2.1, NAVDAS-AR in Sect. 2.2, NCOM 4D-Var in Sect. 2.3, and finally ROMS 4D-Var in Sect. 2.4. Scientific results obtained from these systems are not discussed here. They may be found in the references cited.

2.1 *IOM*

The IOM system is a GUI-driven system that was developed for configuring, constructing, and running weak-constraint, four-dimensional variational data assimilation for any dynamical model and observing system (see Bennett et al. 2008).

2.1.1 Implementation

The system was designed to configure, construct, and run weak-constraint, four-dimensional variational data assimilation for any dynamical model and any observing system. The user needs only to provide the model and the observing system methods, together with an interpolation scheme that relates the model numerics to the observer’s coordinates. The model dynamics and the observing system methods may be nonlinear. All other model-dependent elements of the Inverse Ocean Modeling assimilation algorithm (see Bennett et al. 2008), including model adjoint generation and prior and posterior errors estimation, have been derived and coded as templates in Parametric FORTRAN (Erwig et al. 2007). This programming language has been developed specifically for the IOM system. The IOM system generates conventional FORTRAN code for each of the algorithm elements, such as the adjoints of the user’s dynamical model and observation operators, using the model information entered by the user via a GUI, and the developer provided Parametric FORTRAN templates. The IOM is also used to configure and run various W4D-Var assimilations. The progress and results from each assimilation test is monitored through the IOM GUI.

2.1.2 Applications

The system is a modular system for constructing and running weak constraint four-dimensional variational data assimilation for any linear or nonlinear dynamical model and observing array.

The IOM has been applied to four different ocean models with widely varying model characteristics (Muccino et al. 2008). The models are (1) the Primitive Equations Z-coordinate-Harmonic Analysis of Tides (PEZ-HAT) (Zaron and Egbert 2006), (2) the Regional Ocean Modeling System (ROMS) (Di Lorenzo et al. 2007), (3) the Advanced Circulation model in 2D (ADCIRC-2D) (Muccino and Luo 2005), and (4) the Spectral Element Ocean Model in 2D (SEOM-2D) (Levin et al. 2006). These models have been used in conjunction with the IOM system, to investigate a wide variety of scientific problems including tidal, wind-driven, and mesoscale ocean circulations. In general, the assimilation of ocean observations with the IOM system provides a better estimate of the ocean state than the model prediction alone (see Muccino et al. 2008).

The IOM template language, Parametric FORTRAN, has wider applications in scientific computing (see Erwig et al. 2007). It has been successfully applied to derive NCOM tangent linear and adjoint models (Ngodock and Carrier 2013).

2.2 NAVDAS-AR

NAVDAS-AR is an operational W4D-Var atmospheric data assimilation system that produces dynamically consistent operational global atmospheric analysis for NAVGEM (Hogan et al. 2014).

2.2.1 Implementation

The system described here is a weak-constraint four-dimensional variational data assimilation system to provide the atmospheric analysis for Navy's global numerical weather prediction model (see Xu et al. 2005, 2007; Rosmond and Xu 2006; Chua et al. 2009, 2013). The implementation of NAVDAS-AR follows the accelerated representer method of Xu and Daley (2002). NAVDAS-AR estimates atmospheric analysis simultaneously with bias predictor coefficients in the variational radiance bias correction method (Dee 2004). NAVDAS-AR solves a sequence of linearized weighted least-squares minimization problems. Instead of directly minimizing the four-dimensional cost function numerically, the system solves a large set of linear equations using the flexible conjugate gradient solver (Notay 2000) which shown to have a better convergence property over the standard conjugate gradient solver (see Chua et al. 2009).

One of the unique properties of NAVDAS-AR is that the adjoint of the data assimilation system can be easily constructed by simply changing the order of subroutine calls in the original data assimilation code (also known as the code for the forward data assimilation problem) due to the self-adjointness of the solver used in the representer-based algorithm (Xu et al. 2006). The adjoint of NAVDAS-AR system, a key component of the adjoint-based observation impact method as described by Baker and Daley (2000) and Langland and Baker (2004), has been

developed by Xu et al. (2006). The adjoint of NAVDAS-AR has been used for calculating and monitoring the impact of observations on the short-range forecast error of the Navy's global NWP in real-time.

In an effort to further improve the quality of overall analysis of NAVDAS-AR by introducing some flow-dependent information into the initial background error covariance, NAVDAS-AR-hybrid (a variant of NAVDAS-AR system) has been developed. It linearly combines the static NAVDAS-AR initial background error covariance with a flow-dependent covariance derived from an 80-member ensemble to improve the quality of the initial background error covariance. The ensemble members are generated using the ensemble transform technique (Bishop and Zoth 1999) with a three-dimensional variational data assimilation (3D-Var)-based estimate of analysis error variance (see Kuhl et al. 2013).

2.2.2 Applications

NAVDAS-AR is an operational weak-constraint four-dimensional variational atmospheric data assimilation system for the US Navy. NAVDAS-AR has been as the atmospheric data assimilation system for NAVGEM (Hogan et al. 2014). It was also used as the operational data assimilation system for the US Navy Operational Global Atmospheric Prediction System (NOGAPS; Hogan and Rosmond 1991). The conventional observations assimilated include the ones from land surface stations, radiosondes, dropsondes and pilot balloons, aircraft, and buoys. Besides assimilating the available conventional observations, it also assimilates many satellite observations. They include, for example, Advanced Microwave Sounder Unit-A (AMSU-A), Microwave Humidity Sounder (MHS), the Defense Meteorological Satellite Program (DMSP) Special Sensor Microwave Imager/Sounder (SSMIS), Infrared Atmospheric Sounder Interferometer (IASI), Atmospheric Infra-Red Sounder (AIRS), radio occultation from receivers using the Global Navigation Satellite System (GNSS) satellite systems, atmospheric motion vectors (AMVs) derived from both polar-orbiting and geostationary satellites, ocean surface winds from scatterometers and microwave imagers, and integrated water vapor from microwave imagers. NAVDAS-AR routinely assimilates over 3 million observations within each 6-hour data assimilation window.

The adjoint of NAVDAS-AR system has been applied by Daescu and Langland (2013a, b) to evaluate the forecast sensitivity with respect to the specification of the observation error covariance \mathbf{R} and initial background error covariance \mathbf{B} with NOGAPS. The adjoint-based method approach has also been used to provide guidance on the weighting of the radiance data in the data assimilation system based on observation-error variance estimates derived from an a posteriori diagnosis on the forecast impact. The information extracted from both error covariance \mathbf{R} and \mathbf{B} diagnosis is necessary for designing a parameter-tuning procedure that is effective in reducing the short-range forecast errors (see Daescu and Langland 2013a).

NAVDAS-AR-hybrid system has been applied to NOGAPS using operational model resolution and the operational observational dataset to evaluate its

performance relative to the NAVDAS-AR system (see Kuhl et al. 2013). In general, NAVDAS-AR-hybrid system significantly reduces the forecast error across a wide range of variables and regions.

NAVDAS-AR system with a high-altitude version of NAVGEM has been applied to upper stratosphere and mesosphere radiances data assimilation (see Hoppel et al. 2013, Ruston et al. 2015). The system has been used to investigate the ability of the SSMIS UAS to characterize the mesosphere (see Ruston et al. 2015). NAVGEM analysis with SSMIS UAS observations assimilated agrees with the atmospheric profiles retrieved from the Microwave Limb Sounder (MLS) and the Sounding of the Broadband Emission Radiometry (SABER).

2.3 *NCOM 4D-Var*

NCOM 4D-Var system is a W4D-Var ocean data assimilation system developed for NCOM (Martin 2000, Barron et al. 2006), a United States Navy operational coastal ocean model.

2.3.1 Implementation

NCOM is a baroclinic, hydrostatic, free surface, primitive equation discretized on an orthogonal curvilinear coordinates in the horizontal and a hybrid generalized vertical coordinates in the vertical (Martin 2000, Barron et al. 2006). Parametric FORTRAN (Erwig et al. 2007) was successfully used to generate tangent linear and adjoint models of NCOM (Ngodock and Carrier 2013). NCOM 4D-Var is a weak-constraint four-dimensional variational ocean data assimilation system based on the indirect representer method as described by Bennett (2002) and Chua and Bennett (2001). The implementation of NCOM 4D-Var is detailed by Ngodock and Carrier (2013, 2014a).

2.3.2 Applications

NCOM 4D-Var system described by Ngodock and Carrier (2013, 2014a) has been applied in various coastal ocean regions, including the Monterey Bay (Ngodock and Carrier 2014a, b) and the Gulf of Mexico (Carrier et al. 2014; Muscarella et al. 2015).

NCOM 4D-Var has been used to assimilate observations collected during the second Autonomous Ocean Sampling Network (AOSN II) field experiment in Monterey Bay into NCOM (see Ngodock and Carrier 2014b). The observations collected during AOSN II include sea surface temperature (SST) and sea surface height (SSH) from satellites, as well as subsurface observations from gliders deployed during the field experiment. Data assimilation experiments are carried out

with both strong and weak constraints, and results are compared against independent observations. In general, NCOM 4D-Var improves the model simulation; and that its weak constraint version produces lower analysis errors than the ones produced its strong constraint version.

NCOM 4D-Var configured for the Gulf of Mexico has been used to investigate the impact of assimilating surface velocity observations, derived from the 300 drifters released in the Gulf of Mexico by The Consortium for Advanced Research on Transport of Hydrocarbon in the Environment (CARTHE) during the summer 2012 Grand Lagrangian Deployment (GLAD) experiment, on NCOM performances. The assimilated velocity observations inferred from the drifters markedly improves NCOM model temperature, salinity, SSH, and velocity Eulerian forecast skill (see Carrier et al. 2014). The lagrangian forecast skill is also assessed (see Muscarella et al. 2015) using separation distance and angular differences between simulated and observed trajectory positions. The assimilated drifter velocities substantially improves the model forecast shape and position of a Loop Current ring.

2.4 ROMS 4D-Var

ROMS 4D-Var system is a unique community ocean data assimilation system (<http://www.myroms.org/>) that supports three variants of 4D-Var: a primal formulation of incremental strong constraint 4D-Var (I4D-Var), a dual formulation based on a physical-space statistical analysis system (4D-PSAS), and a representer-based 4D-Var (R4D-Var) (see Moore et al. 2011a). Here, the emphasis will be on the implementation and applications of its R4D-Var sub-system.

2.4.1 Implementation

The system described here is the representer-based 4D-Var (R4D-Var) system of ROMS. ROMS is a baroclinic, free surface, hydrostatic primitive equations discretized on a terrain following coordinate grid in vertical and an orthogonal curvilinear coordinates in the horizontal (Shchepetkin and McWilliams 2005).

The implementation of ROMS R4D-Var follows the indirect representer method of Chua and Bennett (2001). In ROMS R4D-Var, a Lanczos formulation of the restricted \mathbf{B} -preconditioned conjugate gradient (RBCG) method called RBLanczos (see Gürol et al. 2014) is used to solve the large linear system. The solver RBLanczos has demonstrated to have a superior convergence property over standard conjugate gradient solver (see Gürol et al. 2014).

ROMS 4D-Var system implements a suit of diagnostic tools (Moore et al. 2011a) included the adjoint of the R4D-Var system, a feature that is unique in the oceanographic modeling community. The adjoint of the system provides the ability to quantify the impact of individual observation and observation type on the analysis and forecast cycle.

2.4.2 Applications

The community ROMS R4D-Var system described by Moore et al. (2011a) has been widely applied in coastal and shelf-sea regions, including the California Current System (Moore et al. 2011b, c, 2013) and the New York Bight (Zhang et al. 2010).

The ROMS R4D-Var system has been applied to ROMS configured for the California Current System (CCS) to compare its performance relative to two other ROMS 4D-Var algorithms: I4D-Var and 4D-PSAS (see Moore et al. 2011b). The observations available are mostly from satellite platforms in the form of SST and SSH, and subsurface in situ observations from Argo floats, conductivity temperature depth devices (CTD), and expendable bathythermographs (XBT). The three assimilation approaches converge to the same ocean circulation estimate when using the same prior information. The adjoint of the entire 4D-Var system has been explored the sensitivity of the coastal transport to changes in the observations and the observation array (see Moore et al. 2011c). The ROMS R4D-Var system has also been used to compute a sequence of historical analyses for the California Current System for the period spans 1980–2011 (Moore et al. 2013).

The ROMS R4D-Var system has been used for observing strategy evaluation in an effort to build an integrated observation and modeling system for the New York Bight (Zhang et al. 2010). Specifically, the representer-based system identifies a set of proposed tracks for an autonomous coastal glider that is better for predicting horizontal salt flux within the Hudson Shelf Valley in a forecast period of two days.

3 Summary

Our aim in this chapter is to provide a brief review of the recent developments in representer-based operational or semi-operational variational data assimilation systems. The systems reviewed here are being used to perform analyses and prediction in the atmosphere and ocean. One such system, NAVDAS-AR, is currently in operational use (Hogan et al. 2014) for numerical weather prediction (NWP). Other systems, NCOM 4D-Var and ROMS 4D-Var have features that are comparable to operational NWP system, and are routinely used to compute regional ocean analyses and prediction in near real-time.

Acknowledgements This work is partially supported by the Chief of Naval Research through the NRL Base Program, PE 0601153N. The authors gratefully acknowledge Professor Andrew Bennett, a pioneer in the field of representer method for variational data assimilation. The authors also gratefully acknowledge late Professor Yoshi Sasaki, the developer of the weak-constraint variational data assimilation technique.

References

- Baker NL, Daley R (2000) Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Q J R Meteorol Soc* 126:1431–1454
- Barron CN, Birol Kara A, Martin PJ, Rhodes RC, Smedstad L (2006) Formulation, implementation and examination of vertical coordinate choices in the Global Navy Coastal Ocean Model (NCOM). *Ocean Model* 11:347–375
- Bennett AF (1992) Inverse methods in physical oceanography. Cambridge University Press, New York, 346 pp
- Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press, New York, 234 pp
- Bennett AF, Chua BS, Pflaum BL, Erwig M, Fu Z, Loft RD, Muccino JC (2008) The inverse ocean modeling system. I: Implementation. *J Atmos Oceanic Technol* 25:1608–1622
- Bishop CH, Toth Z (1999) Ensemble transformation and adaptive observations. *J Atmos Sci* 56:1748–1765
- Carrier M, Ngodock H, Smith S, Muscarella P, Jacobs G, Özgökmen T, Haus B, Lipphardt B (2014) Impact of assimilating ocean velocity observations inferred from lagrangian drifter data using the NCOM-4DVAR. *Mon Weather Rev* 142:1509–1524
- Chua BS, Bennett AF (2001) An inverse ocean modeling system. *Ocean Model* 3:137–165
- Chua BS, Xu L, Rosmond T, Zaron ED (2009) Preconditioning representer-based variational data assimilation systems: application to NAVDAS-AR. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications. Springer, Berlin/Heidelberg, pp 307–319
- Chua BS, Zaron ED, Xu L, Baker N, Rosmond T (2013) Recent applications in representer-based variational data assimilation. In: Park SK, Xu L (eds) Data assimilation in atmospheric, oceanic and hydrologic application, vol 2. Springer, Berlin, pp 287–301
- Courtier P (1997) Dual formulation of four-dimensional variational assimilation. *Q J R Meteorol Soc* 123:2449–2461
- Daescu DN, Langland RH (2013a) Error covariance sensitivity and impact estimation with adjoint 4D-Var: theoretical aspects and first applications to NAVDAS-AR. *Q J R Meteorol Soc* 139:226–241
- Daescu DN, Langland RH (2013b) The adjoint sensitivity guidance to diagnosis and tuning of error covariance parameters. In: Park SK, Xu L (eds) Data assimilation in atmospheric, oceanic and hydrologic application, vol 2. Springer-Verlag, Berlin, pp 205–232
- Dee DP (2004) Variational bias correction of radiance data in the ECMWF system. Proceedings of the ECMWF workshop on assimilation of high spectral resolution sounders in NWP. Reading, UK, pp 97–112
- Di Lorenzo E, Moore A, Arango H, Chua B, Cornuelle BD, Miller AJ, Powell B, Bennett A (2007) Weak and strong constraint data assimilation in the inverse regional ocean modeling system (ROMS): development and application for a baroclinic coastal upwelling system. *Ocean Model* 16(3–4):160–187
- Erwig M, Fu Z, Pflaum BL (2007) Parametric fortran: program generation in scientific computing. *J Softw Maint Evol* 19:155–182
- Gürol S, Weaver AT, Moore AM, Piacentini A, Arango HG, Gratton S (2014) B-preconditioned minimization algorithms for variational data assimilation with the dual formulation. *Q J R Meteorol Soc* 140:539–556
- Hogan T, Rosmond T (1991) The description of the navy operational global atmospheric prediction system's spectral forecast model. *Mon Weather Rev* 119:1786–1815
- Hogan TF, Liu M, Ridout JA, Peng MS, Whitcomb TR, Ruston BC, Reynolds CA, Eckermann SD, Moskaitis JR, Baker NL, McCormack JP, Viner KC, McLay JG, Flatau MK, Xu L, Chen C, Chang SW (2014) The navy global environmental model. *Oceanography* 27(3):116–125
- Hoppel KW, Eckermann SD, Coy L, Nedoluha GE, Allen DR, Swadley S, Baker NL (2013) Evaluation of SSMS upper atmosphere sounding channels for high-altitude data assimilation. *Mon Weather Rev* 141:3314–3330

- Kuhl DD, Rosmond TE, Bishop CH, McLay J, Baker NL (2013) Comparison of hybrid ensemble/4DVar and 4DVar within NAVDAS-AR data assimilation framework. *Mon Weather Rev* 141:2740–2758
- Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus* 56A:189–201
- Levin JC, Haidvogel DB, Chua B, Bennett AF, Iskandarani M (2006) Euler-Lagrange equations for the spectral element shallow water system. *Ocean Model* 12:348–377
- Martin PJ (2000) Description of the navy coastal ocean model version 1.0. NRL Report NRL/FR/7322/00/9962/, 45 pp
- Moore AM, Arango HG, Broquet G, Powell BS, Zavala-Garay J, Weaver AT (2011a) The regional ocean modeling system (ROMS) 4-dimensional variational data assimilation systems. Part I: System overview and formulation. *Prog Oceanogr* 91:34–49
- Moore AM, Arango HG, Broquet G, Edwards CA, Veneziani M, Powell BS, Foley D, Doyle J, Costa D, Robinson P (2011b) The regional ocean modeling system (ROMS) 4-dimensional variational data assimilation systems. Part II: performance and application to the California Current System. *Prog Oceanogr* 91:50–73
- Moore AM, Arango HG, Broquet G, Edwards CA, Veneziani M, Powell BS, Foley D, Doyle J, Costa D, Robinson P (2011c) The regional ocean modeling system (ROMS) 4-dimensional variational data assimilation systems. Part III: Observation impact and observation sensitivity in the California Current System. *Prog Oceanogr* 91:74–94
- Moore AM, Edwards C, Fiechter J, Drake P, Arango HG, Neveu E, Gürol S, Weaver AT (2013) A 4D-Var analysis system for the california current: a prototype for an operational regional ocean data assimilation system. In: Park SK, Xu L (eds) *Data assimilation in atmospheric, oceanic and hydrologic application*, vol 2. Springer, Berlin, pp 345–366
- Muccino JC, Luo H (2005) Picard iterations for a finite element shallow water equation model. *Ocean Model* 10:316–341
- Muccino JC, Arango H, Bennett AB, Chua BS, Cornuelle B, DiLorenzo E, Egbert GD, Hao L, Levin J, Moore AM, Zaron ED (2008) The inverse ocean modeling system. II: Applications. *J Atmos Oceanic Technol* 25:1623–1637
- Muscarella PA, Carrier M, Ngodock H, Smith S, Lipphardt B, Kirwan AD, Huntley H (2015) Do assimilated drifter velocities improve lagrangian predictability in an operational ocean model? *Mon Weather Rev* 143:1822–1832
- Ngodock H, Carrier M (2013) A weak constraint 4D-Var assimilation system for the navy coastal ocean model using the representer method. In: Park SK, Xu L (eds) *Data assimilation in atmospheric, oceanic and hydrologic application*, vol 2. Springer, Berlin, pp 367–390
- Ngodock H, Carrier M (2014a) A 4DVAR System for the navy coastal ocean model. Part I: System description and assimilation of synthetic observations in monterey bay. *Mon Weather Rev* 142:2085–2107
- Ngodock H, Carrier M (2014b) A 4DVAR System for the navy coastal ocean model. Part II: Strong and weak constraint assimilation experiments with real observations in monterey bay. *Mon Weather Rev* 142:2108–2117
- Notay Y (2000) Flexible conjugate gradients. *SIAM J Sci Comput* 22:1444–1460
- Rosmond T, Xu L (2006) Development of NAVDAS-AR: nonlinear formulation and outer loop tests. *Tellus* 58A:45–58
- Ruston B, Baker N, Swadley S, Hoppel K (2015) Assimilation in the upper stratosphere and mesosphere: role of radiances. In: *Proceedings of ECMWF seminar on use of satellite observations in numerical weather prediction*, Reading, UK, 8–12 Sept 2014
- Sasaki Y (1970) Some basic formulations in numerical variational analysis. *Mon Weather Rev* 98:875–883
- Shchepetkin AF, McWilliams JC (2005) The regional oceanic modeling system (ROMS): a split explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Model* 9:347–404
- Xu L, Daley R (2002) Data assimilation with a barotropically unstable shallow water system using representer algorithms. *Tellus* 54A:125–137

- Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR: formulation and initial tests of the linear problem. *Tellus* 57A:546–559
- Xu L, Langland R, Baker N, Rosmond T (2006) Development of the NRL 4D-Var data assimilation adjoint system. *Geophys Res Abs* 8:8773
- Xu L, Rosmond T, Goerss J, Chua B (2007) Toward a weak constraint operational 4D-Var system: application of the Burger's equation. *Meteorol Z* 16:767–776
- Zaron ED, Egbert GD (2006) Verification studies for a z-coordinate primitive-equation model: tidal conversion at a mid-ocean ridge. *Ocean Model* 14:257–278
- Zhang WG, Wilkin JL, Levin JC (2010) Towards building an integrated observation and modeling system in the New York Bight using variational methods. Part II: Representer-based observing system evaluation. *Ocean Model* 35:134–145

Adjoint-Free 4D Variational Data Assimilation into Regional Models

M. Yaremchuk, P. Martin, G. Panteleev, C. Beattie and A. Koch

Abstract The ongoing trend towards parallelization in computer technologies propels ensemble methods toward the forefront of data assimilation studies in geo-physics. Of particular interest are ensemble techniques which do not require the development of tangent linear numerical models and their adjoints for optimization. These “adjoint-free” methods detect effective search directions for optimization through direct perturbation of the numerical model across carefully chosen sets of states. Optimization proceeds by minimizing the cost function within the sequence of subspaces spanned by these perturbations. In this chapter, an adjoint-free variational technique (a4dVar) is described and demonstrated in an application estimating initial conditions of two numerical models: the Navy Coastal Ocean Model (NCOM), and the surface wave model (WAM). It is shown that a4dVar is capable of providing forecast skill similar to that of conventional 4dVar at comparable computational expense while being less susceptible to excitation of ageostrophic modes that are not supported by observations. Prospects of further development of the a4dVar methods are discussed.

M. Yaremchuk (✉) · P. Martin
Naval Research Laboratory, Stennis Space Center, Bay st louis, MS 39529, USA
e-mail: max.yaremchuk@nrlssc.navy.mil

G. Panteleev
International Arctic Research Center,
University of Alaska, Fairbanks, AK 99783, USA

C. Beattie
Department of Mathematics, Virginia Tech, Blacksburg, VA 24061, USA

A. Koch
Department of Marine Science, University of Southern Mississippi,
Hattiesburg, MS 39529, USA

1 Introduction

As the speed of a single processor reached its physical limit of around 3GHz, the general trend in computer development in the last decade has moved from chips containing several cores to ones with tens or even tens of thousands of cores. In addition, multi-core chips mixed with simultaneous multithreading, memory-on-chip, and special-purpose heterogeneous cores promise further performance and efficiency gains in processing problems which can be efficiently split into parallel subtasks. In that respect, the maximum improvement that can be achieved in running atmospheric and oceanic models is limited by the number of grid points, n_ℓ , that can be attributed to a single core without incurring significant performance loss from inter-core communication.

These new capabilities and the increase in computational power they have produced have stimulated the development of *ensemble methods* for data assimilation. In contrast to adjoint-based 4d variational (4dVar) methods which run the numerical model and its adjoint in sequence in order to compute the cost function gradient, ensemble methods directly aggregate ensemble perturbations to acquire information on the cost function gradient and Hessian structure. In that respect, the ensemble-based 4dVar techniques offer significant parallel performance advantages, replacing the sequence of forward/adjoint model runs with up to M^2/n_ℓ parallel subtasks that involve only the forward model.

A related advantage that ensemble approaches offer is that they are *nonintrusive*, offering the opportunity to treat the numerical model as a black box and thus avoid the burdensome development and maintenance of tangent linear and adjoint codes required by 4dVar methods. Employing this property, Anderson and co-workers (Anderson et al. 2009; Hoteit et al. 2013) developed the Data Assimilation Research Testbed (DART) system on the basis of the widely used Ensemble Kalman filter (EnKF).

There has been recent, significant progress in extending EnKF techniques into the particle filtering framework (Hoteit et al. 2012) and in coupling EnKF techniques with both 3d- and 4d-variational methods (e.g., Županski 2005; Liu et al. 2008; Zhang et al. 2009). Of particular interest for the adjoint-free approach that we present here has been the development of the Maximum Likelihood Ensemble Filter (Županski 2005) based on the explicit computation of the square root of the Hessian matrix restricted to a subspace spanned by ensemble members.

The merging of ensemble approaches with variational techniques has developed along two lines: (a) improvement of the background error covariances (BECs) through the introduction of ensemble-based estimates and their hybrid generalizations (Clayton et al. 2013; Kuhl et al. 2013), and (b) searching for the optimal solution within the subspaces spanned by the leading error modes of the BECs. This second line of approach has been pursued by many authors in the last decade (Liu et al. 2008; Zhang et al. 2009; Zhang and Zhang 2012; Trevisan et al. 2010) and implicitly assumes that the BEC structure is well described by these (possibly localized) BEC modes. More recently, the performance of a family of adjoint-free

methods (4dEnVar) based on a formulation by Liu et al. (2008) has been compared with the standard 4dVar techniques in the framework of idealized experiments with the Lorenz-05 model (Fairbairn et al. 2014). These results show significantly better performance of 4dEnVar for moderate-length assimilation windows with low-density observations. Desroziers et al. (2014) demonstrated a close relationship between the 4dEnVar and 4dVar state space formulations and compared various implementations of 4dEnVar with 4dVar in an idealized setting.

The developments described above mostly deal with meteorological applications, where ensembles are supported by significantly higher data densities than are available in oceanographic applications. High data density allows one to obtain reasonably good estimates of BECs from the ensemble using truncated representation of the localization matrices and to efficiently compute the cost function gradient directly from ensemble perturbations (Liu et al. 2009; Tian and Xie 2012). A significant advantage of such an approach is the absence of the necessity to develop and maintain tangent linear and adjoint codes and the flexibility that results in adapting to various dynamical constraints.

In the ocean, ensemble-based BEC estimates tend to be less accurate, and one has to rely on *ad hoc* BEC representations (Mirouze and Weaver 2010; Yaremchuk and Sentchev 2012). Without reliable correlation information, the development of an efficient adjoint-free assimilation method also becomes more problematic as one must select a small number of reliable perturbations with more care. Early attempts to develop practical a4dVar algorithms in oceanography were limited to predetermined low-dimensional subspaces spanned either by the reduced-order approximations of the model Green's functions (Stammer and Wunsch 1996; Menemenlis and Wunsch 1997), or by the dominant principal component vectors (EOFs) associated with the model statistics (Qui et al. 2007; Hoteit 2008). In fact, the 4dEnVar technique pursues a similar, but more general approach, parameterizing the search subspace by Schur products of the ensemble members with the eigenvectors of the reduced-order representation of the localization matrix.

In this chapter, we give an overview of recent developments in adjoint-free methods of data assimilation using both ensemble-generated and *ad hoc* BEC models, and illustrate the basic principles of the latter approach using an idealized optimization problem constrained by linear dynamics. We describe a particular approach to adjoint-free 4dVar, referred to here as a4dVar (cf., Yaremchuk et al. 2009), which we then apply in Sect. 3 to the assimilation of hydrographic surveys and velocity observations collected in the Adriatic Sea in 2006. Assimilation is constrained by the state-of-the-art Navy Coastal Ocean Model (NCOM) and a4dVar results are compared with those obtained by means of the traditional 4dVar technique. In Sect. 4 the a4dVar method is tested with simulated data constrained by a spectral surface wave model and the forecast skill of the optimized solution is compared with one delivered by an operational method based on sequential assimilation of significant wave height. Section 5 completes the chapter with a summary and discussion of the prospects for adjoint-free methods in general, and a4dVar in particular.

2 Variational Data Assimilation

2.1 Adjoint Methods

Consider the 4dVar method as solving as the following linear discrete least-squares problem constrained by model dynamics in a small vicinity of the model's background trajectory \mathbf{x}_b^n :

$$J = \frac{1}{2} \left[\mathbf{x}^{0\top} \mathbf{B}^{-1} \mathbf{x}^0 + \sum_{n=0}^N (\mathbf{H}_n \mathbf{x}^n - \mathbf{d}^n)^\top \mathbf{R}_n^{-1} (\mathbf{H}_n \mathbf{x}^n - \mathbf{d}^n) \right] \rightarrow \min_{\mathbf{x}^0}. \quad (1)$$

Here n enumerates observation times, \mathbf{B} is the error covariance matrix of \mathbf{x}_b^0 which describes the (Gaussian) error statistics of the model state at $n = 0$, \mathbf{H}_n are the model-data projection operators, \mathbf{d}^n are the misfits $\mathbf{y}_n^o - \mathbf{H}_n \mathbf{x}_b^n$ between observations \mathbf{y}_n^o and the corresponding background model values, \mathbf{R}_n are the observation error covariances, and $^\top$ denotes transposition. We will denote the dimension of the discretized model state vector \mathbf{x} by M and the total number of observations by M_d .

The optimal correction vector \mathbf{x}^n is governed by the recursive relationship $\mathbf{x}^n = \mathbf{M}_n \mathbf{x}^{n-1}$, where \mathbf{M}_n is the dynamical operator of the model linearized in the vicinity of the background trajectory, \mathbf{x}_b^n , across the time interval (t_{n-1}, t_n) , so that

$$\mathbf{x}^n = \mathbf{M}_n \mathbf{M}_{n-1} \dots \mathbf{M}_2 \mathbf{M}_1 \mathbf{x}^0. \quad (2)$$

To avoid the ensuing clutter of symbols, we introduce new notation: $\mathbf{c} \equiv \mathbf{x}^0$ for the control vector, $\mathbf{M}^n \equiv \mathbf{M}_n \dots \mathbf{M}_2 \mathbf{M}_1$ for the aggregated n -step propagator, $\overline{\mathbf{H}}_n \equiv \mathbf{R}_n^{-1/2} \mathbf{H}_n$, $\overline{\mathbf{d}}^n \equiv \mathbf{R}_n^{-1/2} \mathbf{d}^n$. We then drop the over-bars and so, taking (2) into account, the minimization problem (1) can be rewritten in terms of the optimal correction, \mathbf{c} , to the initial state:

$$J = \frac{1}{2} \left[\mathbf{c}^\top \mathbf{B}^{-1} \mathbf{c} + \sum_{n=0}^N (\mathbf{H}_n \mathbf{M}^n \mathbf{c} - \mathbf{d}^n)^\top (\mathbf{H}_n \mathbf{M}^n \mathbf{c} - \mathbf{d}^n) \right] \rightarrow \min_{\mathbf{c}}. \quad (3)$$

A 4dVar data assimilation method finds the minimum of J by solving the normal equations, expressed as:

$$\nabla_{\mathbf{c}} J = \mathbf{B}^{-1} \mathbf{c} + \sum_n \mathbf{M}^{n\top} \mathbf{H}_n^\top (\mathbf{H}_n \mathbf{M}^n \mathbf{c} - \mathbf{d}^n) = 0, \quad (4)$$

To further simplify discussion, introduce the following notation for the Hessian matrix, $\tilde{\mathbf{H}}$, and the right-hand side, \mathbf{b} ,

$$\tilde{\mathbf{H}} = \mathbf{B}^{-1} + \sum_n \mathbf{M}^{n\top} \mathbf{H}_n^\top \mathbf{H}_n \mathbf{M}^n; \quad \mathbf{b} = \sum_n \mathbf{M}^{n\top} \mathbf{H}_n^\top \mathbf{d}^n, \quad (5)$$

allowing us to rewrite the normal equations as $\tilde{\mathbf{H}}\mathbf{c} = \mathbf{b}$.

There are two major approaches to 4dVar data assimilation. The first one, the *state space approach*, iteratively solves (4) through a conjugate gradient descent or related algorithm, which on every iteration computes the gradient and then estimates an effective descent direction using information on the Hessian accumulated in previous iterations. This method is widely used in a number of community OGCMs (NEMO, ROMS), and in operational meteorology (ECMWF).

As may be seen from (4), this process must involve the application of both the model evolution operator, \mathbf{M}^n , and its transpose, $\mathbf{M}^{n\top}$ (the “adjoint model”). The numerical procedure of calculating the gradient involves two major steps:

- (1) Sequential calculation of $\mathbf{x}_i^n = \mathbf{M}^n \mathbf{c}_i$ (forward run of the tangent linear model), supplemented additionally by the calculation of the quantities $\mathbf{q}_i^n = \mathbf{D}_n \mathbf{x}_i^n - \mathbf{H}_n^\top \mathbf{d}^n$.
- (2) Accumulation of the products $\mathbf{M}^{n\top} \mathbf{q}_i^n$ conveniently performed in the reverse-time order because $\mathbf{M}^{n\top} = (\mathbf{M}_n \dots \mathbf{M}_1)^\top = \mathbf{M}_1^\top \dots \mathbf{M}_n^\top$ (backward-in-time integration of the adjoint model).

The sequential nature of this algorithm generally will limit parallel scalability.

A second approach to 4dVar data assimilation is the *observation space approach*—so called because the solution process is mapped into the space of observations instead of remaining solely in the state space. The framework for this may be developed by using the Sherman-Morrison-Woodbury formula to transform the Hessian inverse from having action defined directly in the state space to equivalent action defined in the (generally lower dimensional) observation space. This transformation tactic is closely related to “optimal interpolation” as it seeks the optimal solution in the form of a linear function of model-data misfits, leading also to a family of methods called “representer methods” (Bennett 2002; Rosmond and Xu 2006). In most geophysical applications there will be significant benefit in searching for a solution in the M_d -dimensional observation space as opposed to the much larger M -dimensional state space. The observation space method also has a certain advantage over the state space approach with respect to parallel computing efficiency, since the computation of M_d representers can be performed independently and in parallel.

The robustness of final estimators may be improved if one separates the aggregate background error covariance term into a component, \mathbf{B}_0 , associated with the uncertainty of the initial state, and components, \mathbf{B}_n , associated with the uncertainties of the model equation and forcing. In effect, one replaces $\mathbf{c}^\top \mathbf{B}^{-1} \mathbf{c}$ in the cost function (3) with $\mathbf{c}^\top \mathbf{B}_0^{-1} \mathbf{c} + \sum_{n \geq 1} \mathbf{e}^{n\top} \mathbf{B}_n^{-1} \mathbf{e}^n$ which now involves model errors, $\mathbf{e}^n = \mathbf{x}^n - \mathbf{M}_{n-1} \mathbf{x}^{n-1}$. The normal equations in this case are more complicated than

(4), but some numerical advantages accrue in approaching the resulting computational task using the representer method (Bennett 2002; Rosmond and Xu 2006). Minimization of the (non-linear) cost function in the observation space involves multiple convolutions with the (generally nonsparse) matrices \mathbf{B}_n , making the method sometimes more computationally expensive than the state space approach. This method has been implemented in ROMS (Moore et al. 2011) as an optional feature, and in the Naval Research Laboratory for both atmospheric (Xu and Rosmond 2004; Xu et al. 2005) and oceanic (Ngodock and Carrier 2014) data assimilation systems.

2.2 Adjoint-Free Methods

As the name suggests, adjoint-free methods perform minimization of the cost function without using linearized models and their adjoints. This is achieved by the direct assessment of cost function sensitivity through an ensemble of parallel model runs using perturbed control parameters. Assuming that control parameter perturbations capture the dominant modes of the background error statistics, the ensemble of model trajectories that is produced can be used to estimate the dynamically consistent evolution of the background error covariance, which is implicitly performed by the 4dVar algorithm during optimization.

Currently, the most developed adjoint-free technique is the \mathbf{B} -preconditioned state space approach proposed by Liu et al. (2008, 2009), referred to as 4dEnVar in literature. The major idea is to seek the 4dVar solution in the subspace spanned by the model perturbations. This makes the method equivalent to the observation space 4dVar with the only difference that the search is executed in \mathbb{R}^{km} , where m is the ensemble size and k is the number of eigenmodes in the covariance localization matrix used to diversify search directions (Hamill et al. 2001; Liu et al. 2009). Currently, the method has been successfully tested with real data (Liu et al. 2013) and in Meteo France/UK Met Office (Fairbairn et al. 2014; Desroziers et al. 2014) in a more theoretical context.

Although 4dEnVar has shown promise, the method has some deficiencies which may hinder its use, especially in oceanographic practice, where observations are not as plentiful as in atmospheric practice and, as a consequence, ensembles may be insufficiently accurate in approximating the background error statistics. One may also note that the 4dEnVar minimization process is still based on the sequential computation of gradients used in the course of building the optimal solution; this could lead to a performance bottleneck in massively parallel computing environments.

We discuss another method of adjoint-free minimization based on projecting $\tilde{\mathbf{H}}$ onto the subspace spanned by ensemble perturbations. The approach was first utilized in the Maximum Likelihood Ensemble Filter (Županski 2005) and later extended to an adjoint-free 4dVar variational algorithm that we will refer to as a4dVar (Yaremchuk et al. 2009; Panteleev et al. 2015). The technique involves direct minimization of the cost function in a sequence of $\tilde{\mathbf{H}}$ -orthogonal subspaces and requires an efficient algorithm for computing the action of $\mathbf{B}^{-1/2}$ on a vector which nicely fits

the approach in heuristic BEC modeling using polynomials of the diffusion operator (Yaremchuk et al. 2013; Yaremchuk and Sentchev 2012). Although the a4dVar formulation guarantees its convergence in M/m iterations, practical feasibility requires obtaining a reasonable degree of accuracy in solving the normal equation within several dozen iterations. This is achieved by restricting the basis vectors of the search subspaces to be smooth, implicitly assuming that the leading eigenvectors of $\tilde{\mathbf{H}}^{-1}$ have this property and that the rhs \mathbf{b} of the normal equation will have a sizable projection onto this “smooth manifold”.

To illustrate these ideas, consider a simple problem of retrieving the initial field of tracer concentration $c(\mathbf{x}, 0)$ from observations at some distant time T . The tracer evolution is governed by

$$\partial_t c + \mathbf{u} \nabla c - \mu \Delta c = f(\mathbf{x}, t) \quad (6)$$

in a closed rectangular 49×91 domain Ω (Fig. 1) with the boundary condition $\eta(\partial\Omega, t) = 0$. Equation (6) is discretized on a regular grid using simple first-order explicit time-stepping, upwind advection, and a standard 5-point stencil for the Laplacian with unit steps in temporal and spatial directions. The velocity $\mathbf{u} = (u, v)$ at any space-time location is defined by $u = -0.2 + 0.01v$; $v = -0.1 + 0.01\eta$, where η is the white noise on unit interval. The forcing f is generated by setting $f(\mathbf{x}, t) = 0.001\eta$ in every point of the space-time grid. The coefficient μ is set to 10^{-5} , so that diffusion is largely determined by the numerics.

The simulated data experiment is set as follows: Given the initial tracer distribution $\hat{c} = c(\mathbf{x}, 0) = \exp[-(\mathbf{x} - \mathbf{x}_0)^2/9]$ with $\mathbf{x}_0 = (70, 35)$ (bell-shaped disturbance in Fig. 1a), the model is integrated for $T = 200$ time steps to obtain the final distribution $c(\mathbf{x}, T)$ shown by contours in the same panel. Notice that the initial disturbance almost completely dispersed and migrated to $\mathbf{x}_T \sim (25, 15)$. After that, $c(\mathbf{x}, T)$ is sampled at 200 points shown in Fig. 1a, and the numbers obtained are used to reconstruct \hat{c} by minimizing the cost function (3) under the dynamical constraint (6) with an inverse background error covariance defined by

$$\mathbf{B}^{-1} = \left[\mathbf{I} - \frac{a^2}{2} \Delta \right]^2 \quad (7)$$

where \mathbf{I} is the identity operator in state space and $a = 1.5$ is the decorrelation scale. With the definition (7) at hand, it is easy to compute the action $\tilde{\mathbf{H}}^{1/2}$ on a control vector and perform $\tilde{\mathbf{H}}$ -orthogonalization (see Appendix).

For the purpose of comparison, the cost function is minimized using the state-space 4dVar technique and two versions of a4dVar, which differ in the method of building the search directions (SDs). The number m of SDs (ensemble size) in both a4dVar versions is set to 10. The first version specified SDs as a sequence of tens of eigenvectors of \mathbf{B} in descending order of eigenvalue magnitude. To specify search directions for the second a4dVar method, 200 observations were split into $m = 10$ equal groups so an observation operator \mathbf{H}_j for the j th search direction is sampling

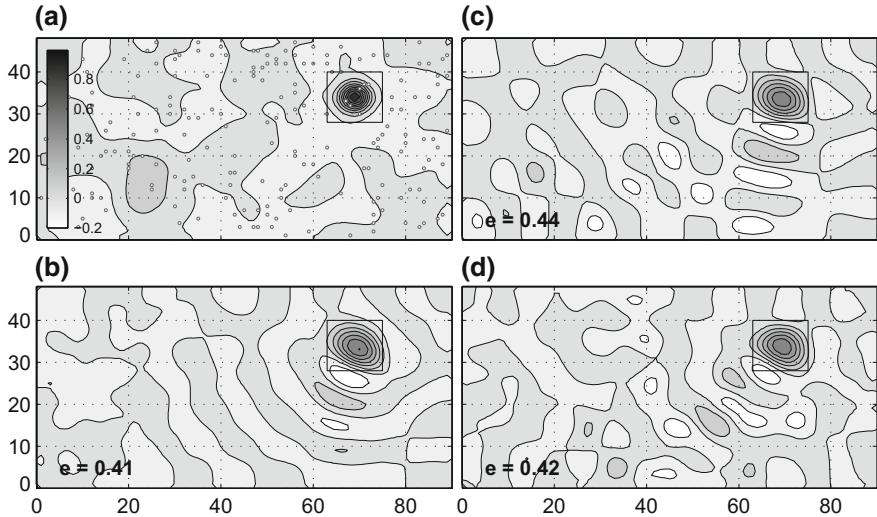


Fig. 1 Reconstruction of the initial condition of the tracer field by 4dVar (b) and a4dVar (c, d) techniques. Composite map of the tracer field evolution being reconstructed is shown in panel a with the initial position of the reconstructed feature (Gaussian eddy at $x = (70, 35)$) superimposed on the tracer field (contours) at the observation time ($t = 200$). Circles denote observation points. The errors in approximation of the true perturbation at $t = 0$ are shown in the left corner

a group of 20 distinct locations among those shown in Fig. 1a. SDs \mathbf{s}_j on the i th iteration were defined by

$$\mathbf{s}_j^i = (\mathbf{B}^{-1} + \mathbf{H}_j^T \mathbf{H}_j)^{-1} \mathbf{q}_j^i, \quad j = 1, \dots, 10 \quad (8)$$

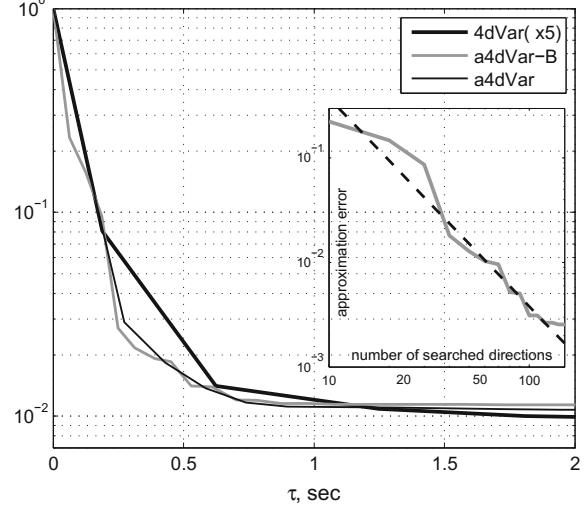
Optimal approximations \tilde{c} to \hat{c} obtained by 4dVar and a4dVar techniques are shown in Fig. 1b and Fig. 1c, d respectively). The quality of reconstruction was assessed by the parameter

$$e = \sqrt{\langle (\tilde{c} - \hat{c})^2 \rangle / \langle \hat{c}^2 \rangle} \quad (9)$$

where angular brackets denote averaging over rectangles enveloping the reconstructed perturbation in Fig. 1. Comparison of Fig. 1b, c, d suggests that the a4dVar method is capable of providing a solution of the same quality with 4dVar, and that the general a4dVar strategy of minimizing J using a sequence of smooth $\tilde{\mathbf{H}}$ -orthogonal SDs may work well with various methods of generating the ensemble members.

In terms of computational expense, the 4dVar method provided approximately five times faster reduction of the cost function (Fig. 2) due to high efficiency of the adjoint model. In this simple case, an adjoint model run required the same amount of time as the direct model run. In real applications, the tangent linear and adjoint codes are several times more expensive to run and the a4dVar techniques may prove to be more competitive, as shown in Sect. 3.

Fig. 2 Reduction of the cost function against CPU time for 4dVar and a4dVar techniques. The 4dVar CPU time is multiplied by five to mimic larger CPU requirements of the state-of-the-art adjoint models. Inset: Convergence of the a4dVar-B solution (Fig. 1c) to the exact solution. Dashed line shows the convergence rate given by (12)



The well-known structure of \mathbf{B} provides an opportunity to assess the convergence rate of the ad4Var solution exposed in Fig. 1c. Assume that after k a4dVar iterations $m_s = km$ $\tilde{\mathbf{H}}$ -orthogonal directions have been already searched and the k th approximation \hat{c}_k to the optimal solution $\hat{c} = \tilde{\mathbf{H}}^{-1}\mathbf{b}$ have been found. Without loss of generality, the eigenvectors ϕ_i of \mathbf{B} could be normalized to satisfy $\phi_i^T \mathbf{B}^{-1} \phi_i = 1$, so that their (Euclidean) norm is equal to the associated eigenvalue σ_i . The magnitude e_{m_s} of the approximation error $\mathbf{e}_k = \hat{c} - \hat{c}_k$ with respect to the norm induced by the inverse covariance can be assessed by projecting \hat{c} on the *unexplored* directions:

$$e_{m_s} = \mathbf{e}_k^T \mathbf{B}^{-1} \mathbf{e}_k \leq \sum_{l>m_s} |\hat{c}^T \mathbf{B}^{-1} \phi_l|^2 \quad (10)$$

Furthermore, since the optimal solution $\hat{c} = \tilde{\mathbf{H}}^{-1}\mathbf{b}$ allows representation in the (dual) form $\hat{c} = \mathbf{B}\rho$ (ρ is the optimal linear combination of the representers), the upper bound of the terms under summation in (10) can be assessed by

$$|\hat{c}^T \mathbf{B}^{-1} \phi_l| = |\rho^T \phi_l| \leq \sigma_l (\rho^T \mathbf{B}^{-1} \rho)^{1/2} \quad (11)$$

Plugging (11) into (10) yields the following upper bound on the error magnitude:

$$e_{m_s} \leq \rho^T \mathbf{B}^{-1} \rho \sum_{l>m_s} \sigma_l \sim O(m_s^{-2}) \quad (12)$$

This estimate remains intact if we assess e_{m_s} with respect to the norm induced by the Hessian matrix. In the latter case, the right-hand side of (12) will be additionally multiplied by a scaling factor $\|\tilde{\mathbf{H}}\|/\|\mathbf{B}^{-1}\| > 1$.

Dependence of the distance between the 4dVar solution (Fig. 1b) and the consecutive approximations to the a4dVar solution (Fig. 1d) shown in the inset to Fig. 2, confirms the above estimate.

Similar experiments with a low-dimensional ($M = 1,922$) non-linear quasigeostrophic model were performed by Yaremchuk et al. (2009) who documented compatible performance of the 4dVar and a4dVar methods in the non-linear regime and certain advantages of the a4dVar approach in the cases of sparse and/or noisy observations. In the next sections we present the results of applying a4dVar to real and simulated data constrained by state-of-the-art numerical models.

3 a4dVar and 4dVar Assimilation of Real Data in the Adriatic Sea

3.1 Model and Data

The NCOM is a free-surface primitive-equation hydrostatic ocean model with σ coordinates in the upper layers and, optionally, fixed depths below a user-specified distance from the surface. Algorithms that comprise a NCOM computational kernel are described in Martin (2000); Barron et al. (2006). The model was configured at 3 km resolution on an 85×294 horizontal grid (Fig. 3) with 32 levels in the vertical. The top 22 σ levels follow the bathymetry, stretching from the surface to a fixed depth of 291 m, and 10 fixed-depth levels are used below 291 m. Initial and open boundary conditions for the sea surface height ζ , temperature T , salinity S , and horizontal velocities u, v were provided from the regional NCOM simulation (Martin et al. 2009). The model was forced by the river runoff and atmospheric fluxes derived from the regional atmospheric model with 8 km horizontal resolution. In the described assimilation experiments, initial conditions were used as control variables,

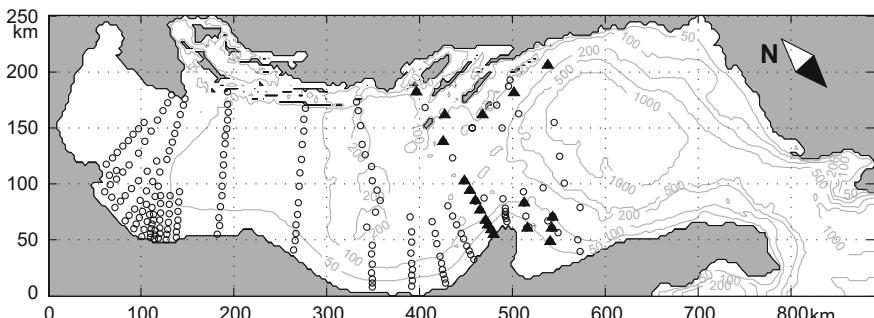


Fig. 3 Model domain with CTD stations (circles) and moorings (triangles) of the DART experiment. Gray contours (m) show bottom topography

i.e., the vector \mathbf{c} comprised all the grid point values of ζ, T, S, u, v at $n = 0$. With the given 3-dimensional grid and bathymetry, the inverse problem has $M = 1,493,570$ unknowns.

The first guess (background) values of \mathbf{c} were taken from the NCOM simulation described by Martin et al. (2009) and then adjusted to suppress temperature and salinity biases during the assimilation time interval (0.00 UTC on 08/14 to 0.00 UTC on 08/29/2006). After the adjustment, the horizontal-and-time average misfits between the background solution and TS observations did not exceed 0.02 °C and 0.005 psu, respectively.

Assimilated data were acquired in the course of the field experiment Dynamics of the Adriatic in Real Time (DART) (Martin et al. 2009; Burrage et al. 2009). In the present study, CTD and ADCP observations from August 14 to August 29, 2006 are used (Fig. 3). Temperature T and salinity S were measured at 219 CTD stations occupied in the northern and central parts of the basin. The total number of TS observations used in the assimilation is 9,650. Current velocities u, v were measured by 19 moored ADCPs in the depth range from 15 to 150 m at locations shown by triangles in Fig. 3. All the velocity data were detided and averaged over 29 twelve-hour intervals centered at the assimilation times t_n of 0 and 12 UTC. With the total number of the observed velocities 13,856 the dimension of the observation space was $M_d = 23,506$.

3.2 Assimilation Parameters

In the course of the experiments the parameters of the tested 4dVar and a4dVar systems were kept as close as possible to each other. However, due to the different formulations (observation space for NCOM 4dVar and state space for a4dVar), certain discrepancies remained in the shape of the background error covariance \mathbf{B} . In both algorithms \mathbf{B} is given by the product \mathbf{VCV} where \mathbf{V} is the diagonal matrix of the background error rms variances and \mathbf{C} is the respective correlation matrix.

In the 4dVar algorithm, the action of \mathbf{C} on a state vector is represented by the operator

$$\mathbf{C} \simeq \exp\left(\frac{1}{2}b^2\Delta\right), \quad (13)$$

which is implemented numerically by integrating the heat transfer equation (e.g., Weaver and Courtier 2001) with the decorrelation length scale $b = 9$ km. Since the matrix \mathbf{C} is rank-deficient, the second-order polynomial approximation to the exponent in (13) was used to define \mathbf{C}^{-1} in the a4dVar algorithm. Parameter a was set to $\sqrt{8/\pi b}$ to preserve the value of the integral decorrelation scale specified in 4dVar (e.g., Yaremchuk and Smith 2011). The rest of the assimilation parameters were identical for both the 4dVar and a4dVar assimilation systems.

The tested a4dVar method is based on Eq. 8 and outlined as follows:

0. Specify the dimension m_s of the search subspaces, their number k to be kept in memory for $\tilde{\mathbf{H}}$ -orthogonalization, the maximum number of iterations I , the perturbation magnitude ϵ and the background model trajectory \mathbf{x}_b^n . Set the iteration number i to zero, $\mathbf{c}_0 = 0$, and compute \mathbf{d}^n .

1. Compute $\mathbf{x}_i^n, J_i, \mathbf{Y}_i = \tilde{\mathbf{H}}^{1/2} \mathbf{c}_i$ and the search directions $\mathbf{s}_n^i, n = 0, \dots, N$ (Eq. 8).
2. Extract the m_s leading EOFs $\mathbf{p}_i^m, m = 1, \dots, m_s$ of the search directions to form the basis in the search subspace.
3. Perturb the initial conditions $\mathbf{c}_i \rightarrow \mathbf{c}_i + \epsilon \mathbf{p}_i^m$ and run (in parallel) the ensemble of m_s perturbed models, computing the respective perturbed values of J_i^m and \mathbf{Y}_i^m .
4. $\tilde{\mathbf{H}}$ -orthogonalize the search basis $\{\mathbf{p}_i^m\}$ with respect to at most k basis vectors obtained on the previous iterations and compute optimal corrections $\delta \mathbf{c}_i$ (see Appendix 1).
5. Set $\mathbf{c}_{i+1} = \mathbf{c}_i + \delta \mathbf{c}_i$.
6. If $i = I$ exit. Otherwise set $i \leftarrow i + 1$, then go to 1.

The stopping criteria for the iterative processes were selected as follows: For the 4dVar system the solution of the system for the representer coefficients was terminated after $n_t = 7$ iterations, when the accuracy of the conjugate gradient (CG) solver was, as a rule, better than 10^{-3} . With the value of $n_t = 7$, 8–10 outer loops were executed before the values of J started to increase. For the a4dVar system, the minimization was terminated when the total CPU time reached the value used by the respective 4dVar experiment. The number of ensemble members was kept constant at $m_s = 9$ through all the experiments.

3.3 Comparison with 4dVar

In the reported experiments we varied the length of the assimilation window from short (4 days, $N = 9$) to moderate (8 days, $N = 17$) and long (14 days, $N = 29$) duration. Performance of the assimilation algorithms was evaluated in three categories: the forecast skill at the end of the assimilation window (for $N = 9, 17$), the rate of convergence, and by qualitative inspection of the optimal model trajectories.

3.3.1 Convergence Rates and Computational Expense

To assess the rates of convergence, one has to have an ability to compare the reduction of the cost function with iterations, which is not straightforward for two reasons.

First, in the 4dVar algorithm considered here, the regularization term of the cost function can be evaluated only within the range of the correlation matrix defined by (13). To avoid the burden of restricting the a4dVar correlation matrix to the range of \mathbf{C} , we compared only the observational parts of the 4dVar and a4dVar cost functions (second term in Eq. (3)).

Second, the number of iterations required for convergence cannot be considered as an objective criterion because 4dVarV and a4dVar iterations are different in nature. Due to the non-linearity of the problem, an iteration (either 4dVar or a4dVar) performs minimization in the vicinity of the current (suboptimal) state, but 4dVar does that in the range of \mathbf{B} , whereas a4dVar minimizes in the subspace of a much smaller dimension spanned by \mathbf{p}^m . For that reason, iterations require quite different computational resources and should be compared in terms of CPU time. Figure 4 shows such a comparison by rescaling the horizontal axis with the total CPU time τ_a required by one a4dVar iteration. The value of τ_a was 11 times larger than the CPU time τ_m of a direct NCOM model run for a given experiment, i.e. $\tau_a \simeq 11\tau_m$. The major contribution to τ_a is given by the ensemble run (9 τ_m , p.3 in the layout of Sect. 3.2), while the master NCOM run (p.1) and operations listed in pp.2 and 4 require τ_m and $0.8\tau_m$, respectively. Overall, convergence was achieved at an expense of 60–70 iterations (650–800 NCOM runs).

As may be seen in Fig. 4, a single 4dVar iteration was approximately equivalent to 6–7 a4dVar iterations, or 70–80 direct model runs. This computational expense arises because sequential execution of the adjoint and tangent linear codes (inner loops of the CG solver) required around $11\tau_m$, whereas one 4dVar outer loop included seven inner loops to solve the system of linear equations for the representor coefficients.

Figure 4 shows that, in general, the tested a4dVar method is computationally comparable to the observation space 4dVar. Although the total CPU time required for reduction of J by the factor of 0.4 (attained after the first outer loop of the 4dVar) appears to be similar for the 4dVar and a4dVar methods, the a4dVar minimization noticeably slows down at subsequent iterations, especially for longer assimilation windows ($\tau = 8, 14$ days).

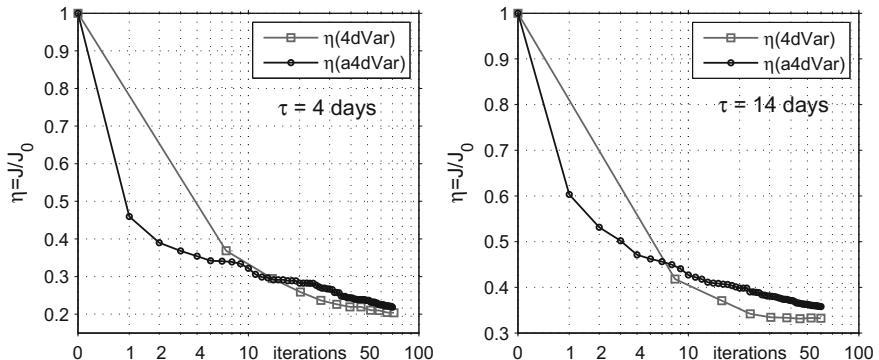


Fig. 4 Relative reduction η of the cost function with iterations (marked by *circles*) for different assimilation periods. The *horizontal axis* is scaled by the CPU time required for the a4dVar iteration. Squares label the 4dVar outer loops

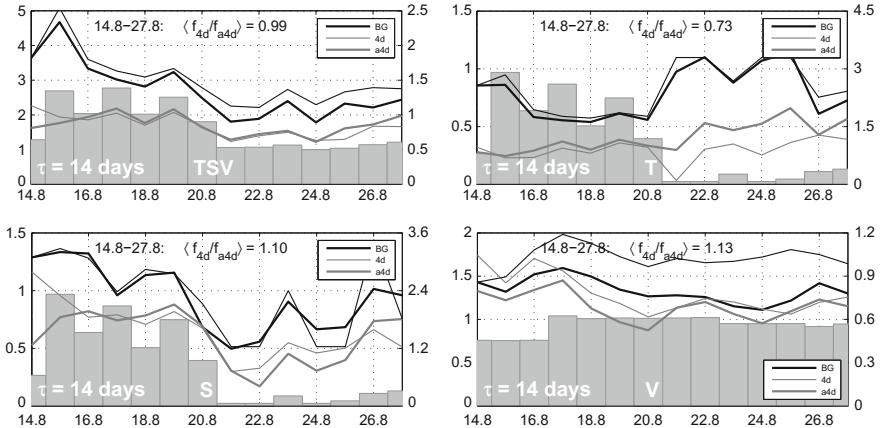


Fig. 5 Evolution of the root-mean-square model-data misfits f_q characterizing the background (BG, thick lines), 4dVar-optimized (4d, thin gray lines) and a4dVar-optimized (a4d) solutions. Thin black line shows the misfit with the background fields at $t = 0$ (persistence). The values of f_q are shown on the right axis of each panel. The left axis quantifies the number of the data points for each day in thousands (shown by gray shaded rectangles). The ratio of the mean values of f_q averaged over the assimilation window for the 4dVar and a4dVar methods is given

Figure 5 demonstrates the time evolution of the quantities

$$f_q^n = \left\langle \left[(\mathbf{H}_n \mathbf{x}_q^n - \mathbf{d}_q^n)^T (\mathbf{H}_n \mathbf{x}_q^n - \mathbf{d}_q^n) / n_q \right]^{1/2} \right\rangle \quad (14)$$

characterizing the daily averaged $\langle \rangle$ model-data misfits of the various state vector components before (black lines) and after (gray lines) optimization with a 14-day assimilation window (i.e., using all the available data). The subscript q takes the values of the labels in the mid-bottom parts of Fig. 5 which indicate the observed variables (temperature, salinity and velocity vector) for which the statistics f_q were computed, whereas n_q stands for the total number of respective observations taken at a given day. The upper left panel in Fig. 5 shows a remarkable similarity in the time evolution of the combined model-data misfit for the 4dVar- and a4dVar-optimized NCOM states. The a4dVar algorithm has, however, a noticeable tendency to provide a better fit at the beginning of the assimilation window, clearly visible in the lower panels for f_S and f_v . This can be explained by the above mentioned property of a4dVar to better retrieve optimal states at shorter integration times.

When separated into different components, behavior of f_T^n , f_S^n , and f_v^n reveals more differences. In particular, the 4dVar method provides a much better fit to the temperature data after August 20 (in the second half of the assimilation window), but appears to be 10–13% worse than a4dVar with respect to salinity and velocity.

A large contribution to a better salinity fit is given by the first two days of the a4dVar model trajectory (lower left panel in Fig. 5). However, certain gains relative to 4dVar are also observed at the end of the assimilation, which is quite opposite to the difference in the values of f_T .

Compared to f_T and f_S , the overall improvement of the model-data misfit is the smallest for velocity (lower right panel in Fig. 5), which was characterized by the observation errors of $\mathbf{R}^{1/2} \sim 7\text{--}10 \text{ cm/s}$ in the cost function. Several assimilation runs were made with significantly smaller (3–5 cm/s) errors, but they were found to be inconsistent with a posteriori statistics of the model-data misfits as the optimal cost function values in these cases were much larger than those obtained in the reported experiments. The a4dVar-optimized value of f_v is persistently smaller (as compared to 4dVar) during the entire assimilation period providing the 13 % better value in the 14-day average. This advantage could be partly attributed to the fact that the a4dVar search directions are derived from the most persistent patterns of the model-data misfits and therefore tend to be closer to the slowly evolving (geostrophically and hydrostatically balanced) modes of the flow.

The quality of the assimilated solutions was assessed for 4- and 8-day experiments using comparison with observations outside the respective assimilation windows. Evolution of the quantities f^n for the background and optimized solutions is shown in Fig. 6 for the 4-day assimilation experiment.

The general behavior of f is consistent with the one obtained in the 14-day experiment, showing persistently better 4dVar forecasts in temperature and the advantage of a4dVar in the salinity and velocity forecasts. The upper left panel in Fig. 6 summarizes the forecast skill and indicates that 4dVar slightly outperforms a4dVar, mostly because of the better temperature forecasts. On the other hand, the 4dVar-optimized

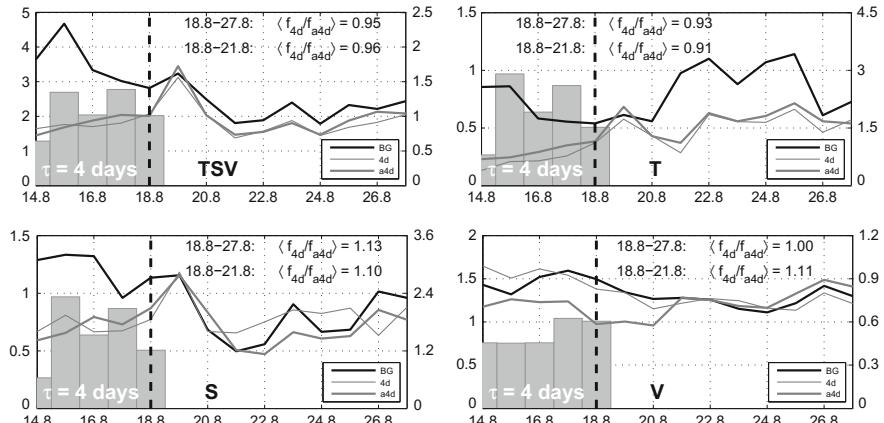


Fig. 6 Forecast skills f_s , f_v and f_t of the 4dVar and a4dVar-optimized solutions for the 4-day assimilation window. The relative number of the respective data points for each day is shown by gray shaded rectangles. Vertical dashed line show the time interval of data assimilation. Ratios of the mean f values averaged over the 3- and 9-days intervals are shown

salinity is characterized by very low forecast skill (lower left panel in Fig. 6), especially during August 21–25, when it was even farther away from the observations than the background forecast.

The 4dVar-optimized velocities show only small improvements compared to the background solution (lower right panel in Fig. 6). In contrast, the a4dVar-optimized velocities demonstrate 10–30% reduction of the model-data misfit within the assimilation window, which persists for up to three days (August 18–21) of the free model run. After August 21, the velocity mismatch of the background, a4dVar and 4dVar-optimized solutions are nearly identical. Qualitatively similar behavior of the forecast skill and its distribution among the state vector components was observed in the results of the 8-day assimilation experiment.

In general, the overall forecast skill provided by the a4dVar method appears to be comparable with that of the 4dVar (upper panel in Fig. 6), and in some aspects (such as short-term velocity forecast), the a4dVar technique provides noticeably better results. It should be noted that available observations could effectively constrain only a small part $M_d/M = 23,506/1,493,570 \sim 1.5\%$ of the model's degrees of freedom, so one should expect substantial differences in the small-scale structure of the optimal solutions obtained by two different methods.

3.3.2 Comparison of the Optimal Solutions

Temperature and velocity increments for the optimal states of the 14-day assimilation experiment are shown in Fig. 7. A certain coherence between the larger scale corrections to the background temperature field are clearly seen in the northern part of the model domain that is well covered by observations (cf. Fig. 1). The time-mean correlation coefficients ρ between the low-pass filtered temperature and salinity increments of the 4dVar and a4dVar solutions are 0.61 and 0.45, respectively if averaging is performed in the upper 200 m over the northern part of the domain. In the data-free region south of the 340 km mark, the correlations are substantially lower (respectively, 0.26 and 0.32) and lie below the 95 % confidence level of nonzero correlation (0.36). Similar values of ρ (0.59 and 0.32 in the northern and southern subregions, respectively) were obtained for the sea surface height field.

Velocity increments appear to have the lowest correlations among the model fields with time-averaged values of $\rho_v = 0.36, 0.27$ for the northern and southern subregions respectively. The lowest correlations ($\rho_v = 0.09, \rho_T = 0.21$, and $\rho_S = 0.12$) were observed in the data-free southern subregion during the first 4 days (8/14–8/18) of the assimilation. Such incoherence between the increments is caused by excessive ageostrophic activity (lower panel in Fig. 7) of the 4dVar solution at the beginning of the assimilation window. The ageostrophic mode disappears at the later times and does not affect the cost function because the southern subregion is virtually data-free, whereas smoothness constraints are imposed on the model fields only at the initial time.

The problem could be apparently solved by introducing balance constraints (e.g., Weaver et al. 2005) into the BEC definition at $n = 0$, which may not be necessary if

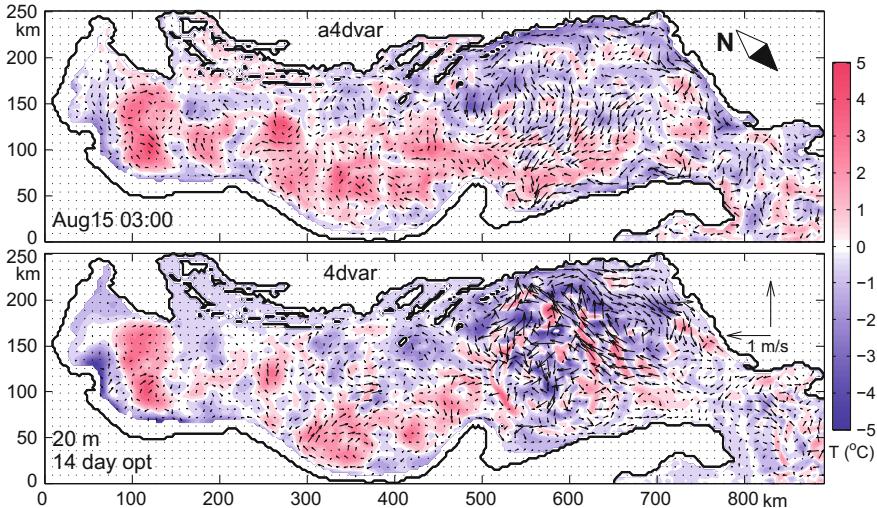


Fig. 7 Temperature and velocity differences between the background and optimized NCOM states at 20 m on August 15 03 UTC. Results of 4dVar and a4dVar optimizations are shown on the *left* and *right* panels respectively

the NCOM 4dVar were run in the weakly constrained mode, i.e., if model errors were prescribed throughout the entire assimilation window. For the purpose of comparison with a4dVar we ran the 4dVar system in the strongly constrained mode and the effect became visible after several 4dVar outer loops. It is remarkable that the a4dVar algorithm appears to be much less susceptible to excitation of the ageostrophic modes (upper panel in Fig. 7), possibly because the EOF-derived descent directions span subspaces characterized by slower time variation of the model trajectory and, therefore, tend to be closer to geostrophic and hydrostatic balance. It is quite likely that introduction of the balance constraints into \mathbf{B} will certainly improve the performance of both algorithms with a potentially larger benefit for the 4dVar case.

4 a4dVar Analysis of Simulated Wave Data in the Chukchi Sea

Spectral models simulating surface gravity waves in the ocean are challenging for application of 4dVar due to complexity of their numerics and non-local nature of the observational operators. Since only a few wave models have been supplied with (incompletely) linearized codes and their adjoints, operational forecasts are still performed using sequential techniques, mostly based on optimal interpolation (OI) of the significant wave height (SWH) data from satellites. In this section, we test the

performance of a4dVar technique under the dynamical constraints imposed by a spectral wave model (WAM 1988; Monbaliu et al. 2000) and compare the results of assimilation with a sequential method.

4.1 The WAM Model and Simulated Data

The WAM model performs time integration of the balance equation describing spectral density $F(\mathbf{x}, \mathbf{k}, t)$ for the wave component with the wavenumber $\mathbf{k} = (k_x, k_y)$ at the location $\mathbf{x} = (x, y)$:

$$\frac{\partial F}{\partial t} + \nabla \cdot (\mathbf{v}F) = \mathcal{S}(F, \mathbf{x}, \mathbf{k}, t), \quad (15)$$

where \mathcal{S} is the sum of source functions, composed primarily of wind-forced generation, dissipation and redistribution of the wave spectrum by non-linear wave-wave interactions (WAM 1988), $\nabla = \{\nabla_x, \nabla_k\}$ stands for the gradient in the horizontal and wavenumber coordinates and \mathbf{v} is the 4-component vector of the respective wave-propagation velocities depending on the ambient current and constrained by the dispersion relationship for linear surface waves $\sigma^2 = g|\mathbf{k}|\tanh|\mathbf{k}|h$, where σ is the wave angular frequency and $h(\mathbf{x})$ is the water depth. Given the appropriate initial/boundary conditions, ambient current and wind forcing, Eq. (15) is integrated numerically to produce evolution of the wave spectrum.

The model was configured in the domain shown in Fig. 8 with the spatial resolution of $\delta x = 9$ km. There were $m_x = 4412$ active grid points in horizontal and $m_k = 600$ grid points (24 directions at 15° resolution and 25 logarithmically spaced frequencies between 0.0314 and 0.3091 Hz) in the wavenumber space. The total length of the state vector was $M = m_x \times m_k = 2,647,200$.

Distance between the model states were assessed in terms of the correlation coefficient C and the normalized rms difference S between the spectra:

$$C(F) = \frac{\langle F'_1 F'_2 \rangle}{\sqrt{\langle F'^2_1 \rangle \langle F'^2_2 \rangle}}; \quad S(F) = \left[\frac{\langle (F_1 - F_2)^2 \rangle}{\sqrt{\langle F'^2_1 \rangle \langle F'^2_2 \rangle}} \right]^{1/2} \quad F' = F - \langle F \rangle \quad (16)$$

where angular brackets denote averaging in space, time, and over the wavenumbers. Similar coefficients were calculated to assess the differences between the scalar (SWH) and vector (wind speed) fields, with averaging performed just in space and time.

The general form of the cost function used in the data assimilation experiments was identical to (3) with the M -dimensional vector $\mathbf{c} = N(t_0) - N_b(t_0)$ describing the difference between the gridded model state $F(\mathbf{x}, \mathbf{k}, t_0)$ and the background (first guess) state $F_b(\mathbf{x}, \mathbf{k}, t_0)$ at the start of model integration t_0 . The first term in (3) was specified by

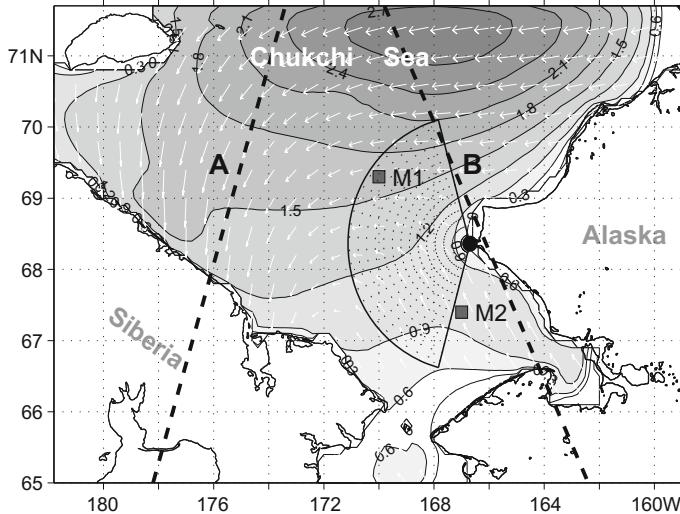


Fig. 8 Wind speed (white arrows) and significant wave height (contours, m) of the reference model solution at 0.00 UTC 09/20/2011 ($t = 0$). Mooring positions are shown by black squares. HFR location and coverage area are given by the black circle with a sector. SWH data are acquired along the radar beams shown by dotted lines within the sector. Dashed lines are the tested tracks of the Envisat satellite

$$\mathbf{c}^T \mathbf{B}^{-1} \mathbf{c} = W \sum_x [(\mathbf{I} - a^2 \Delta_x^2) \hat{Q} \mathbf{c}]^2 \quad (17)$$

where Δ_x stands for the Laplacian in horizontal coordinates and the operator \hat{Q} relates squared SWH with the spectral density through the following linear relationship:

$$Q^2(\mathbf{x}, t) = \hat{Q}F = 16 \sum_k F(\mathbf{x}, \mathbf{k}, t) d\mathbf{k}. \quad (18)$$

Here $d\mathbf{k}$ denotes the grid cell area in the wavenumber space and summation is done over the entire spectral grid. In Eq. (17), the regularization weight W was chosen to be inversely proportional to the squared mean of SWH in the background solution with the proportionality coefficient $\varepsilon_x = 0.01$. By setting $a = 2\delta x$ throughout the experiments, SWH variability at spatial scales below two horizontal grid steps (18 km) was heavily penalized. In the spectral subspace, Eq. (17) defines the inverse error covariance to have only one linearly independent column (specified by the components of \hat{Q}). As a consequence, spectral correlations at a given point are represented by $m_k \times m_k$ correlation matrix whose elements are equal to 1 (thus implying 100 % correlation between all the spectral components). This assumption is routinely used in the sequential algorithms assimilating SWH (e.g., Wittmann and Cummings 2004).

To perform the a4dVar experiments, the reference wave field was generated by integrating WAM from the state of rest for ten days under realistic forcing by the

winds taken for the period 11–20 of September, 2011. The reference initial state $F_r(x, k, t)$ shown in Fig. 8 was taken at the beginning of the last 9 h of the model run (0.00 to 9.00 on 09/20/2011). Synthetic data were picked from the reference solution and then used for its reconstruction by the a4dVar method.

In this study two types of simulated data are considered: moored observations of the wave spectra and SWH measurements from coastal HF radars and satellites.

Two tested mooring sites are shown in Fig. 8. Simulated data from the moorings were generated by multiplying the reference spectrum at any moment by the random factor $1 + \varepsilon\eta$, where η is the white noise with unit variance and $\varepsilon_m = 0.01$. The observational error covariance matrices \mathbf{R} for both moorings were diagonal with time-independent diagonal elements equal to $\langle \varepsilon F_r \rangle^2$. The respective observational operators \mathbf{H} picked the time varying WAM spectra every 15 min from the grid point nearest to the buoy location, providing $4m_k = 2,400$ observations per hour.

SWH observations were simulated by integrating the true spectrum (Eq. 18) in the apexes of the grid cell containing an HFR observation point followed by linear interpolation onto that point. After that, the SWH value was contaminated by random noise with the rms variance of 30 cm. HF radar observation points were located along the beams of the radar shown in Fig. 8. The above described HFR observation operator computed SWH values along the 25 beams every 15 min, providing information to 535 model grid points within the sector shown in Fig. 8 (2,140 observations per hour).

Synthetic satellite observations of sea surface roughness provided SWH data along the Envisat tracks shown in Fig. 8 with 9 km discretization (55 and 73 points for track A and B respectively). These data were assumed instantaneous and satellite passage occurred for both tracks after 2 h of model integration. The respective observation operator was similar to the one used for HFR, except for it picked SWH values at the sequence of WAM grid points closest to the sampling points along the tracks (i.e. no spatial interpolation was used). Satellite SWH observations were contaminated similarly to HFRs with the rms error variance of 30 cm.

The background model trajectory was obtained as follows: The reference solution was averaged in time and space and the resulting spatially homogeneous spectrum was used as initial condition for the background run. The run was forced by the winds which were different from those forcing the true solution. First, the true winds were horizontally smoothed to mimic the errors typical for reanalysis winds from meteorological centers that are usually available at a coarser ($0.25\text{--}1^\circ$) resolution and have to be interpolated on the fine resolution grid of a regional wave model. In the case considered, the smoothing was done by the isotropic Gaussian filter with the half-width of 25 km. After smoothing, the winds were rotated 35° counterclockwise to increase their distance from the reference vectors to $S_{wind} = 0.67$. The larger distance from the true forcing was needed for better assessment of the observation impact on the reconstruction of initial conditions, whose signature usually persists for 3–5 h in a typical wave model integration.

Synthetic observations of SWH and wave spectra were assimilated into WAM using the a4dVar technique described in Sect. 3. The WAM model was constrained by data during the first three hours of model integration and then integrated for six

hours to assess the improvement of the forecast skill. Performance of the method was quantified by calculating correlation coefficients C and normalized rms deviation S (Eq. 16) between the optimized and true solutions. These quantities were computed with time averaging over three time intervals: 0–3 h (assimilation period), and two forecast periods of 3–6, and 6–9 h.

On each a4dVar iteration five SDs were extracted from the EOF analysis of the 3 h model run constrained by the data. The ensemble model runs (p. 3 in the layout of Sect. 3.2) were executed in parallel and required 62 s of wall time per a4dVar iteration on five processors of the 2.3 GHz cluster.

4.2 Comparison with Sequential Method

To compare the a4dVar results with the traditional OI method, we used the 2d OI approach (Wittmann and Cummings 2004; Waters et al. 2013) in application to the SWH data: at the observation times the WAM model state was sequentially updated by the OI analysis of the SWH field, which was projected on the spectral components by multiplying the spectrum at a grid point by the ratio of the updated to predicted SWH values. The OI algorithm was configured with the same background error covariance \mathbf{B} , \mathbf{R}_n , \mathbf{H}_n and using the same reference and background solutions as the a4dVar method.

A series of OI and a4dVar experiments were conducted, involving assimilation of the data from five sources and their combinations: high-frequency radar (hereinafter denoted by HF), two moorings (a4dVar analyses only, locations shown in Fig. 8) and two Envisat tracks (A and B, Fig. 8). For comparison purposes, we conducted similar experiments with OI method assimilating only SWH data from satellites and/or HF radar. In the description below, these experiments are abbreviated by oHFA(B) and oHF respectively. With the exception of satellite tracks, all a4dVar assimilation experiments demonstrated significant improvement of the model state in terms of its proximity to the reference solution. The stopping criterion for optimization was reduction of the cost function gradient 1000 times, which usually occurred after 80–100 iterations. By that time the cost function was typically reduced 2–3 times.

Maps of deviations from the reference solution of the spatially averaged background and optimized spectra at $t = 0$ are shown in Fig. 9. In most of the a4dVar experiments, the initial error has been reduced to the values compatible with the wind forcing errors. The only exceptions were the results of optimal interpolation (Fig. 9b) and of the a4dVar assimilation of SWH data from a single satellite track (not shown): in these cases the optimized spectrum was only slightly different from the one produced by the background solution. For the OI case such a small correction can be explained by the fact that SWH data are weakly constrained by dynamics and can barely affect the shape and location of the spectra because they provide information only on their mean magnitude at a geographical position. Small spectral improvement of the a4dVar experiments with a single satellite track could be attributed to the small amount of data (55 SWH observations). As a consequence, the cost func-

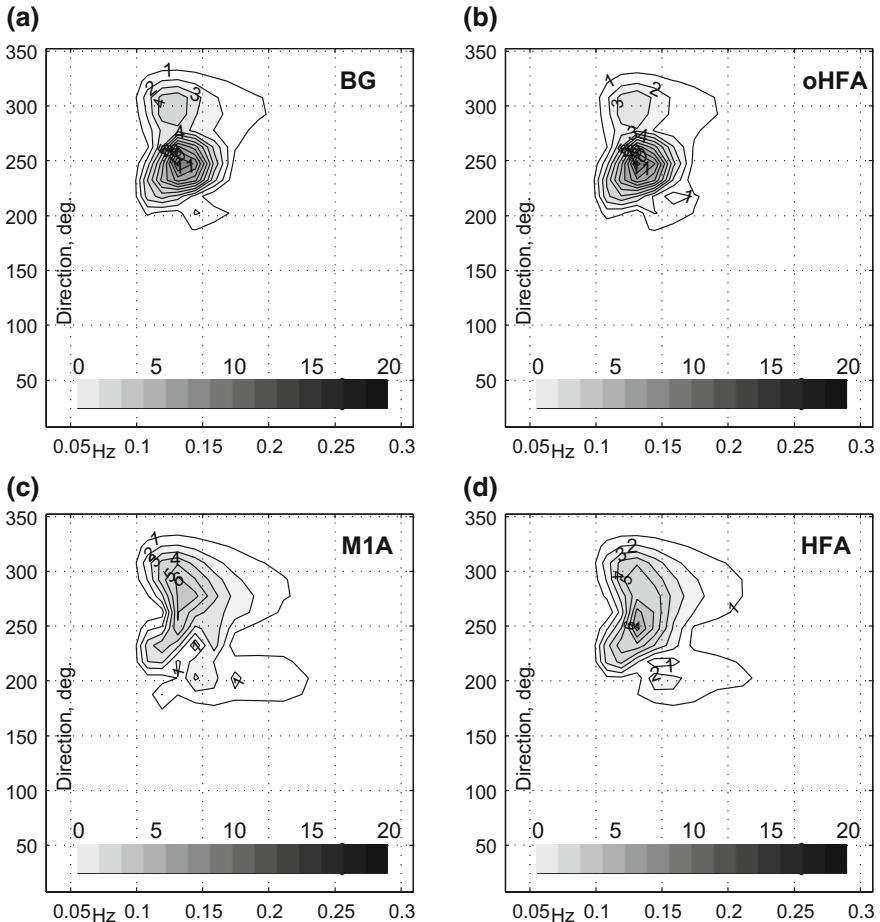


Fig. 9 Absolute difference between the horizontally averaged reference spectrum at $t = 0$ and **a** background spectrum, **b** oHFA-optimized, **c** M1A-optimized, and **d** HFA-optimized spectra

tion is dominated by the regularization term, which implies 100 % correlations in spectral space and is therefore capable of adjusting only the spectral magnitude.

These properties of the above mentioned assimilated solutions translate into their lower spectral forecast skill shown in Table 1, which also includes spectral errors from the other assimilation experiments. Abbreviations in the header of the Table correspond to the types of data used in the experiment (e.g., HFA corresponds to assimilation of the HF data and SWH data from the Envisat track A).

Direct measurement of the spectra by a single mooring (7,200 observation points, columns M1,M2) also provide only a moderate increase of the correlation coefficients C_{03} to 0.52 and decrease of S_{03} to 0.86 as compared to the background (BG) solution. This can be partly explained by the fact that assimilated spectra occupy a

Table 1 Normalized rms distances S and correlations C between the optimized and true solutions for the experiments with various types of data. Subscripts 03, 36 and 69 correspond to time averaging between 0–3, 3–6, and 6–9 h of model integration

	BG	HF	HFA	HFB	oHFA	M1	M1A	M1B	M2	M2A	M2B	M12	A	B
C_{03}	0.47	0.77	0.75	0.76	0.48	0.53	0.70	0.71	0.52	0.65	0.69	0.71	0.47	0.48
S_{03}	0.89	0.65	0.66	0.64	0.87	0.85	0.71	0.70	0.87	0.76	0.75	0.73	0.89	0.88
C_{36}	0.48	0.72	0.72	0.76	0.49	0.47	0.70	0.70	0.50	0.69	0.69	0.67	0.48	0.48
S_{36}	0.87	0.69	0.68	0.65	0.86	0.87	0.71	0.71	0.87	0.72	0.72	0.75	0.88	0.87
C_{69}	0.59	0.71	0.70	0.75	0.59	0.50	0.68	0.69	0.57	0.67	0.68	0.68	0.57	0.58
S_{69}	0.80	0.70	0.70	0.67	0.79	0.86	0.73	0.72	0.86	0.73	0.73	0.73	0.79	0.78

small part of spectral domain (at most 15–20 %, Fig. 9). As a consequence, the effective number of observations with useful (non-zero) information on the state of the wave field should be reduced 5–7 times down to \sim 1,500 data points on the total, which is compatible, by the way, to assimilating 3–7 spectral moments. Besides, mooring data do not provide any information on the spatial variability of the spectra, which appears to be crucial for the successful recovery of the reference state.

In that respect, it is remarkable that adding much less numerous satellite data to moored spectral observations improves the performance of the assimilation system considerably. Combining moored and satellite data provides 30–40 % growth of the correlation coefficients and 20–25 % drop of the normalized standard deviations from the reference spectrum (compare columns M1 and M2 with columns M1A(B) and M2A(B)). At the same time, Satellite SWH data do not add much new information to that containing in HFR observations (cf. columns HF and HFA(B)), which monitor the same integral quantity for the whole assimilation period (3 h) and cover a significant part of the model domain (Fig. 8).

Importance of the spatial coverage by observations is confirmed by the result of the experiment with assimilation of the spectra from two moorings: The values of C and S in this case demonstrate a considerable improvement and become compatible (column M12 in Table 1) with those achieved with the joint assimilation of spectra from the single mooring and satellite SWH data (columns M1A(B) and M2A(B)).

Inspection of Table 1 also shows that information from track A increases the efficiency of assimilating spectra from moorings, but to somewhat lesser extent than track B. This phenomenon can be partly explained by the fact that track A does not cover the region of the highest SWH and, therefore, provides less information on the magnitude of spatial variability of the wave field. Similarly, assimilation of the M2 data appears to be slightly less efficient than M1, which can be partly attributed to M2 position at the periphery of the domain.

Table 1 also indicates that instantaneous Envisat observations on a regional scale cannot provide a significant improvement to the background state, if they are not accompanied by continuous in situ measurements. At the same time, satellite data

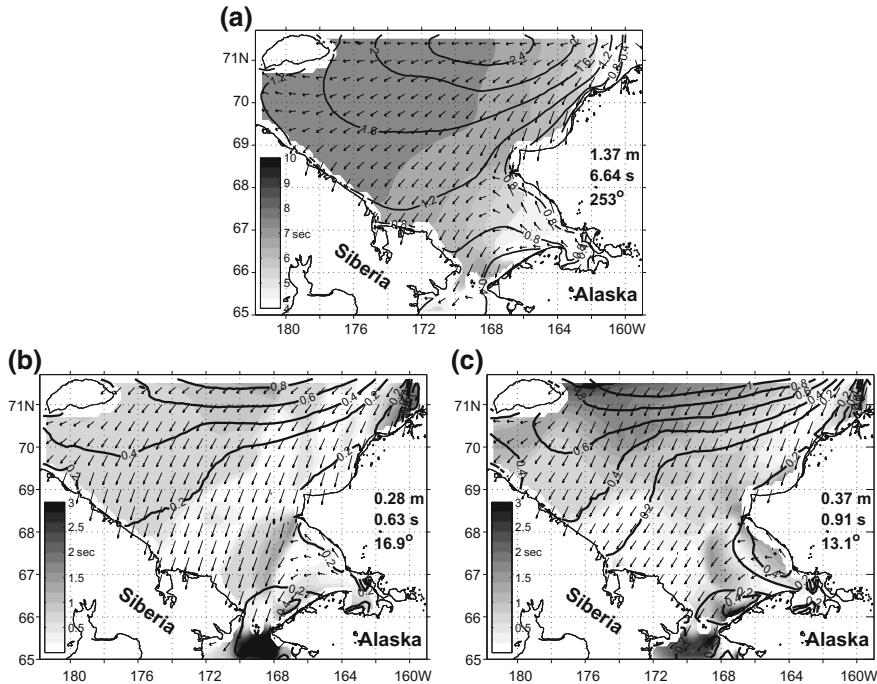


Fig. 10 Time-averaged (3–9 h) values of SWH (contours, m), peak period (shading, seconds) and wave direction (arrows) of the true state **a** and respective absolute differences of the HF-optimized **b** and oHf **c** solutions from the true state. The domain-averaged values of the fields are shown on the right

become quite valuable in complementing observations if the wave conditions are measured by a single mooring.

The forecast errors provided by the HFA data assimilation using OI and a4dVar techniques and averaged over the period 3–9 h are compared in Fig. 10 in terms of the horizontal distributions of the SWH, peak period and wave direction errors. It is seen, that a4dVar technique provides 30–50 % better forecast skill in terms of the SWH (0.28 vs 0.37 m) and peak period (0.63 vs 0.91 s). Although discrepancies in the peak period near the southern and eastern boundaries are comparable in both solutions, the a4dVar method demonstrates a significant advantage over OI in the northern Chukchi Sea and south of Cape Hope resulting in approximately 10 cm smaller SWH errors throughout the entire domain. A local maximum in the a4dVar peak period errors is also observed southwest of Cape Hope (Fig. 10b), that can be partly explained by a sharper gradient in the peak period field of the true solution (Fig. 10a).

The OI solution demonstrates a slightly better skill in forecasting the wave direction (the mean difference of 13.1° vs 16.9°). However, in the OI assimilation experiments with other types of SWH data this number varied within $13\text{--}13.3^\circ$ and was quite close to the respective characteristic (13.2°) of the background solution.

In general, our experiments have shown that the OI method tends to improve the amplitude of the spectrum, and has only a slight impact on its shape and position in the frequency-direction coordinates. In contrast, a4dVar technique is capable of improving these characteristics as well, since it performs optimization along the most persistent dynamical modes of the governing equation (15). This important property of the a4dVar algorithm provides a significantly better approximation of the reference solution and improved forecast skill.

5 Summary and Discussion

In this chapter we have shown feasibility of the a4dVar technique (Yaremchuk et al. 2009) in realistic applications and compared its performance with the observation space 4dVar and OI methods. It was shown that the a4dVar approach is capable of producing optimized solutions of similar quality to 4dVar with comparable computational expense. It was also found that the a4dVar technique is less susceptible to excitation of ageostrophic modes in the data-free regions if balance constraints are not imposed on the background error covariances.

The a4dVar technique employs square root factorization the inverse BEC and the possibility of inexpensive evaluation of the product $\tilde{\mathbf{H}}^{1/2}\delta\mathbf{c}$ during the integration of the ensemble of perturbed model trajectories. The technique of Hessian factorization was first proposed by Županski (2005) in the framework of minimizing the cost function within the subspace spanned by the ensemble members. It was later extended in Yaremchuk et al. (2009) to heuristic BEC models coupled with iterative ensemble updates produced by projections of the model-data misfits on a suitable “smooth” manifold generated by the low-pass filtering operators \mathbf{M}^n or \mathbf{B} .

Our experience shows that there exists a considerable freedom in generating the SDs as long as they are kept being spatially smooth and $\tilde{\mathbf{H}}$ -orthogonal. In particular, selecting the ensembles as eigenvectors of \mathbf{B} in the decreasing order of their eigenvalues proves to be equally efficient, at least in the simple linear setting considered in Sect. 2. In that respect, there is a considerable similarity between the a4dVar and the adjoint-free 4dEnVar method (Liu et al. 2008), which explicitly looks for optimized solution in the range of the localized approximation to the background error covariance. However, the 4dEnVar uses the ensemble to approximate the cost function gradient which is then used in the iterative optimization, whereas a4dVar directly employs the ensemble members to minimize the cost function in the respective subspace.

The ultimate goal of a search algorithm is to rapidly gain information on the Hessian structure, which helps to find the SD $\mathbf{s} = \tilde{\mathbf{H}}^{-1} \mathbf{b}$ towards the cost function minimum (note that \mathbf{s} is often nearly orthogonal to the local gradient). In that respect, ensemble methods can offer a significant advantage in their ability to perform *parallel* searches in *multiple SDs*, which can be competitive with the adjoint-based methods even in some cases where individual SDs may not appear to be as efficient as steepest descent or conjugate gradient directions.

Regarding linearization issues, a state-of-the-art GCM code is never fully differentiable and its (always approximate) adjoint usually requires several times more CPU/memory resources than the direct model run. This observation indicates that a4dVar approach could be even competitive with 4dVar even in terms of the total CPU time at small ensemble sizes. The present chapter demonstrates this compatibility in both a real-life scenario and a simplified linear application.

The a4dVar technique can be developed further by introducing flow-dependent covariances and better restricting the SDs to the slow-evolving (geostrophically and hydrostatically balanced) manifold. In application to atmospheric and oceanic modeling, the BEC matrix is can easily incorporate these balance constraints by representing the state vector in the form

$$\mathbf{x} = \begin{bmatrix} \mathbf{I}_1 & \mathbf{0} \\ \mathbf{L} & \mathbf{I}_2 \end{bmatrix} \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \quad (19)$$

where \mathbf{L} is the balance operator (Weaver et al. 2005), $\mathbf{x}_{1,2}$ are the unbalanced components of the state vector, and $\mathbf{I}_{1,2}$ are the identity matrices of the respective sizes. Under these constraints, $\mathbf{B}^{-1} = \langle \mathbf{x} \mathbf{x}^T \rangle^{-1}$ in the a4dVar formulation will take the form

$$\mathbf{B}_{bal}^{-1} = \begin{bmatrix} \mathbf{B}_1^{-1} + \mathbf{L}^T \mathbf{B}_2^{-1} \mathbf{L} & -\mathbf{L}^T \mathbf{B}_2^{-1} \\ -\mathbf{B}_2^{-1} \mathbf{L} & \mathbf{B}_2^{-1} \end{bmatrix}, \quad (20)$$

where \mathbf{B}_1^{-1} and \mathbf{B}_2^{-1} are the inverse covariances of \mathbf{x}_1 and \mathbf{x}_2 . Further improvements can be made by replacing the Laplacian in Eq. (7) with a more general expression (e.g., Weaver and Mirouze 2012; Yaremchuk and Nечаev 2013) introducing flow dependent structure into $\mathbf{B}_{1,2}^{-1}$ while keeping them square root factorisable. Note that spectral analysis of the background error covariance could be efficiently performed prior to the assimilation.

Alternatively, flow-dependence and cross-correlations could be introduced into BEC through its representation by the localized external ensembles, as it is done in 4dEnVar. This will require \mathbf{B} -preconditioning of the control variables, which will bring the method closer to the observation space 4dVar. In that respect, it is interesting to note a certain similarity between the a4dVar and observation space 4dVar: The latter method explicitly computes the Hessian projection on the observation space (representer matrix), whose computation is efficiently parallelized between M_d processors, making the method competitive with a4dVar in terms of scalability on massively parallel computers. This property brings observation space 4dVar

closer to the family of optimization algorithms capable of taking the advantage of massive parallelism. In our vision, such algorithms are getting higher priority under the current “parallelization trend” in the development of computer technologies. In a sense, the present situation is somewhat similar to the situation 30 years ago when the adjoint methods started coming into practice in response to the rapid increase of computer speed and memory.

In terms of the computational expense, the tested a4dVar technique appears comparable to 4dVar, mostly because of the excessive computational cost of tangent linear and adjoint codes that were, on average, several times more expensive than a direct run of the parent nonlinear model (which is a typical situation with state-of-the-art OGCMs, e.g., Oldenborgh et al. 1999). On massively parallel machines, the advantage of a4dVar will be more noticeable due to the limited parallel scalability of an OGCM code, be it original, adjoint, or tangent linear.

An important issue with the a4dVar technique is its extension to optimization of other sets of variables that may control the model trajectory, such as surface forcing fields. One of the possible solutions in this case augments the search subspaces (ocean model states) by the leading EOFs of the surface forcing error fields. This will require a better knowledge of error statistics of the atmospheric model used to force the ocean in a particular application. In view of recent rapid development of the observational systems and data acquisition techniques in the atmosphere, the issue of accessibility to the above mentioned statistics seems likely to be resolvable in the near term. Moreover, the a4dVar technique appears to be even more suitable for coupled ocean-atmosphere systems, where external forcing errors tend to play a lesser role at the time scale of a typical assimilation window.

A much larger computational advantage is evident when considering the wall time in a massively parallel environment, which formally allows a4dVar to search over multiple directions at a fraction of the wall time used by 4dVar to generate a steepest descent direction. In fact, in the experiments reported in Sect. 3, one a4dVar run was executed almost five times faster if all the ensemble members were run on separate nodes. This property of the a4dVar approach gives good prospects for its further development in sync with other types of ensemble data assimilation techniques that are based on relaxed communication requirements between processors. In our vision, rapidly decreasing prices of the massively parallel computers make finite differentiation in functional spaces more affordable, favoring development of ensemble methods of data assimilation, while investment in the development and maintenance of linearized codes and their adjoints may gradually become less practical.

Acknowledgements This study was supported by the Office of Naval Research (Program element 0602435N) as part of the project “Adjoint-free 4dVar for Navy ocean models”. Partial support from NSF grants PLR-1107925, PLR-1203740, DMS-1217156 and Einstein Foundation of Berlin is also acknowledged.

Appendix

The a4dVar method utilizes the technique employed by Źupanski (2005) in the Maximum Likelihood Ensemble Filter, which is based on the explicit inversion of the Hessian matrix in the subspace spanned by the model perturbations. In view of the definition (7), $\mathbf{B}^{-1/2}$ can be explicitly represented using the expression for the square root of the inverse error covariance:

$$\mathbf{B}^{-1/2} = \mathbf{V}^{-1}(\mathbf{I} - \frac{b^2}{2}\mathbf{A}) \quad (21)$$

which allows a symmetric Hessian factorization

$$\tilde{\mathbf{H}} = \tilde{\mathbf{H}}^{\top/2} \tilde{\mathbf{H}}^{1/2}, \quad (22)$$

where

$$\tilde{\mathbf{H}}^{\top/2} = [\mathbf{B}^{-1/2} \quad \mathbf{H}_0 \quad \mathbf{H}_1 \mathbf{M}^1 \quad \dots \quad \mathbf{H}_N \mathbf{M}^N] \quad (23)$$

is the Hessian square root.

For sufficiently small perturbations $\delta \mathbf{c}_m = \epsilon \mathbf{p}_m$, perturbations of the auxiliary vector

$$\delta \mathbf{Y}_m = \tilde{\mathbf{H}}^{1/2} \delta \mathbf{c}_m \quad (24)$$

are linear in $\delta \mathbf{c}_m$, so that computation of the dot products between the vectors $\delta \mathbf{Y}_m$ provides the inner product in the control space associated with the Hessian matrix

$$\delta \mathbf{Y}_1^{\top} \delta \mathbf{Y}_2 = \delta \mathbf{c}_1^{\top} \tilde{\mathbf{H}} \delta \mathbf{c}_2 = \langle \delta \mathbf{c}_1, \delta \mathbf{c}_2 \rangle_{\tilde{\mathbf{H}}}, \quad (25)$$

which can be used for $\tilde{\mathbf{H}}$ -orthogonalization of the search subspaces of the a4dVar algorithm.

We seek the optimal correction of the control variable \mathbf{c} in the search subspace \mathbb{S} spanned by \mathbf{p}_m :

$$\mathbf{c} \leftarrow \mathbf{c} + \sum_{l=1}^{m_s} s_l \mathbf{p}_l,$$

where the coefficients s_l satisfy for $m = 1, 2, \dots, m_s$,

$$\mathbf{p}_m^{\top} \left(\tilde{\mathbf{H}} \left(\mathbf{c} + \sum_{l=1}^{m_s} s_l \mathbf{p}_l \right) - \mathbf{b} \right) = 0. \quad (26)$$

This constitutes a Ritz-Galerkin projection of the normal system (4) to the search subspace, \mathbb{S} . Rearranging, we obtain the linear system of m_s equations in the m_s unknowns s_1, s_2, \dots, s_{m_s} :

$$\sum_{l=1}^{m_s} \mathbf{p}_m^T \tilde{\mathbf{H}} \mathbf{p}_l s_l = \mathbf{p}_m^T (\mathbf{b} - \tilde{\mathbf{H}} \mathbf{c}). \quad (27)$$

Substituting $\mathbf{p}_m = \delta \mathbf{c}_m / \varepsilon$ into (27), multiplying by ε^2 , using (22) and (24) yields

$$\sum_{l=1}^{m_s} \delta \mathbf{Y}_m^T \delta \mathbf{Y}_l s_l = \varepsilon \delta \mathbf{c}_m^T (\mathbf{b} - \tilde{\mathbf{H}} \mathbf{c}). \quad (28)$$

The right-hand side of (A.7) cannot be computed directly because evaluation of $\mathbf{b} - \tilde{\mathbf{H}} \mathbf{c}$ requires the adjoint code (Eq. 5). Nonetheless, for each m , $\delta \mathbf{c}_m^T (\mathbf{b} - \tilde{\mathbf{H}} \mathbf{c})$ can be calculated directly from the variations of the cost function $\delta J_m = J(\mathbf{c} + \delta \mathbf{c}_m) - J(\mathbf{c})$ induced by $\delta \mathbf{c}_m$:

$$\begin{aligned} \delta J_m &= \frac{1}{2} \delta \mathbf{c}_m^T \tilde{\mathbf{H}} \delta \mathbf{c}_m + \delta \mathbf{c}_m^T (\tilde{\mathbf{H}} \mathbf{c} - \mathbf{b}) \\ &= \frac{1}{2} \delta \mathbf{Y}_m^T \delta \mathbf{Y}_m - \delta \mathbf{c}_m^T (\mathbf{b} - \tilde{\mathbf{H}} \mathbf{c}). \end{aligned} \quad (29)$$

Thus, the coefficients for the optimal correction of the control variable \mathbf{c} within the search subspace \mathbb{S} are given as the solution to a linear system posed in terms of the quantities δJ_m and $\delta \mathbf{Y}_m$ computed by the a4dVar algorithm:

$$\sum_{l=1}^{m_s} \delta \mathbf{Y}_m^T \delta \mathbf{Y}_l s_l = \varepsilon \left(\frac{1}{2} \delta \mathbf{Y}_m^T \delta \mathbf{Y}_m - \delta J_m \right). \quad (30)$$

In the $\tilde{\mathbf{H}}$ -orthonormal coordinate system $\delta \mathbf{Y}_m^T \delta \mathbf{Y}_m = \varepsilon^2$, and Eq. (30) are simplified to

$$s_l = \sum_m \alpha_{lm} \left(\frac{\varepsilon}{2} - \frac{\delta J_m}{\varepsilon} \right), \quad (31)$$

where α_{lm} are the matrix elements of the linear transformation of the original basis $\delta \mathbf{c}_m$ that are obtained in the orthogonalization process.

For a differentiable numerical model and sufficiently small ε , the quadratic term in the right hand side of (30) is negligible. In the experiments reported in Sect. 3 we kept it in place since the value of ε was close to 0.01 and could not be reduced any further without affecting the rate of convergence. The relatively large limit on the value of ε was caused by a number of factors deteriorating the linear dependence between the magnitude of the model perturbations and ε . These factors include rounding errors (especially for temperature and salinity in the upper layers),

non-differentiable operators in the model code, particularly at the open boundary, and small-scale instabilities of the flow, especially prominent in the experiments with the 14-day assimilation window.

References

- Anderson JL, Hoar T, Raeder K, Liu H, Collins N, Torn R, Arellano A (2009) The data assimilation research testbed: a community facility. *Bull Am Meteorol Soc* 90:1283–1296
- Barron CN, Kara AB, Martin PJ, Rhodes RC, Smedstad LF (2006) Formulation, implementation and examination of vertical coordinate choices in the global Navy Coastal Ocean Model (NCOM). *Ocean Modell* 11:347–375
- Bennett AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press, pp 234. ISBN 0-521-81373-5
- Burrage DM, Book JW, Martin PJ (2009) Eddies and filaments of the Western Adriatic Current: Analysis and prediction, *J Mar Syst*, 78, S205–S226
- Clayton AM, Lorenc AC, Barker DM (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q J R Meteorol Soc*. doi: [10.1002/qj.2054](https://doi.org/10.1002/qj.2054)
- Desroziers G, Camino J-T, Loik Berre (2014) 4dEnVar: link with 4D state formulation of variational assimilation and different possible implementations. *Q J R Meteorol Soc* 140:2097–2110
- Fairbairn D, Pring SR, Lorenc AC, Roulstone I (2014) A comparison of 4dVar with ensemble data assimilation methods. *Q J R Meteorol Soc* 140:281–294
- Hamill TM, Whitaker JS, Snyder C (2001) Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon Weather Rev* 129:2776–2790
- Hoteit I (2008) A reduced-order simulated annealing approach for four-dimensional variational data assimilation in meteorology and oceanography. *Int J Numer Methods Fluids* 58:1181–1199
- Hoteit I, Luo X, Pham DT (2012) Particle Kalman filtering: a nonlinear Bayesian framework for ensemble Kalman filters. *Mon Weather Rev* 140:528–542
- Hoteit I, Hoar T, Gopalakrishnan G, Anderson J, Collins N, Cornuelle B, Kohl A, Heimbach P (2013) A MITgcm/DART ocean prediction and analysis system with application to the Gulf of Mexico. *Dyn Atmos Oceans* 63:1–23
- Kuhl DD, Rosmond TE, Bishop CH, McLay J, Baker N (2013) Comparison of hybrid ensemble/4dVar and 4dVar within the NAVDAS-AR data assimilation framework. *Mon Weather Rev* 141:2740–2758
- Liu C, Xiao Q, Wang B (2009) An ensemble-based four-dimensional variational data assimilation scheme. Part II: observing system simulation experiments with advanced research WRF (ARW). *Mon Weather Rev* 137:1687–1704
- Liu C, Xiao Q, Wang B (2013) An ensemble-based four-dimensional variational data assimilation scheme. Part III: Antarctic applications with advanced research WRF using real data. *Mon Weather Rev* 141:2721–2739
- Liu C, Xiao Q, Wang B (2008) An ensemble-based four-dimensional variational data assimilation scheme. Part I: Technical formulation and preliminary test. *Mon Weather Rev*, 136, 3363–3373
- Martin PJ, Book JW, Burrage DM, Rowley CD, Tudor M, (2009) Comparison of model-simulated and observed currents in the central Adriatic during DART, *J Geophys Res*, 114, C01S05. doi: [10.1029/2008JC004842](https://doi.org/10.1029/2008JC004842)
- Martin PJ (2000) A description of the Navy Coastal Ocean model version 1.0, NRL Rep. NRL/FR/7322 00-9962, Nav Res Lab, Stennis Space Cent, MS pp 42
- Menemenlis D, Wunsch C (1997) Linearization of an oceanic general circulation model for data assimilation and climate studies. *J Atmos Ocean Technol* 14:1420–1443
- Mirouze I, Weaver A (2010) Representation of correlation functions in variational data assimilation using an implicit diffusion operator. *Q J R Meteorol Soc* 136:1421–1443

- Monbaliu J, Padilla-Hernandez R, Hargreaves JC, Carretero Albiach JC, Luo W, Sclavo M, Gnther H (2000) The spectral wave model WAM, adapted for applications with high spatial resolution. *Coast Eng* 41:41–62
- Moore AM, Arango HG, Broquet G, Powell BS, Weaver AT, Zavala-Garay J (2011) The Regional Ocean Modeling System (ROMS) 4-dimensional variational data assimilation systems: Part I system overview and formulation. *Prog Ocean* 91(1):34–49
- Ngodock H, Carrier M (2014) A 4dVar system for the Navy Coastal Ocean Model. Part I: system description and assimilation of synthetic observations in Monterey Bay. *Mon Weather Rev* 142(6):2085–2107
- Oldenborgh GJ, Burgers G, Venzke S, Eckart C, Giering R (1999) Tracking down the ENSO delayed oscillator with an adjoint OGCM. *Mon Weather Rev* 127:1477–1495
- Panteleev G, Yaremchuk M, Rogers E (2015) Adjoint-free variational data assimilation into a regional wave model. *J Atmos Ocean Technol*, 32. (in press)
- Qui C, Shao A, Wei L (2007) Fitting model fields to observations by using singular value decomposition: an ensemble-based 4dVar approach. *J Geophys Res* 112:D11105. doi:[10.1029/2006JD007994](https://doi.org/10.1029/2006JD007994)
- Rosmond T, Xu L (2006) Development of the NAVDAS-AR: non-linear formulation and outer loop tests. *Tellus* 58A:45–58
- Stammer D, Wunsch C (1996) The determination of the large-scale circulation of the Pacific Ocean from satellite altimetry using model Green's functions. *J Geophys Res* 101:18409–18432
- The WAMDI Group (1988) The WAM model—a third generation wave prediction model. *J Phys Ocean* 18:1775–1810
- Tian X, Xie Z (2012) Implementations of a square-root ensemble analysis and a hybrid localization into the POD-based ensemble 4dVar. *Tellus A* 64:1–10. doi:[10.3402/tellusa.v64i0.18375](https://doi.org/10.3402/tellusa.v64i0.18375)
- Trevisan A, DiIsidoro M, Talagrand O (2010) Four-dimensional variational assimilation in the unstable subspace and the optimal subspace dimension. *Q J R Meteorol Soc* 136:487–496
- Waters J, Wyatt LR, Wolf J, Hines A (2013) Data assimilation of partitioned HF radar wave data into Wavewatch III. *Ocean Modell* 72:17–31
- Weaver AT, Deltel C, Machu E, Ricci S, Daget N (2005) A multi-variate balance operator for variational data assimilation. *Q J R Meteorol Soc* 131:3605–3625
- Weaver AT, Mirouze I (2012) On the diffusion equation and its application to isotropic and anisotropic correlation modeling in variational assimilation. *Q J R Meteorol Soc* 138. doi:[10.1002/qj.1953](https://doi.org/10.1002/qj.1953)
- Weaver AT, Courtier P (2001) Correlation modelling on a sphere using a generalized diffusion equation. *Q J R Meteorol Soc* 127:18151846
- Wittmann PA, Cummings JA (2004) Assimilation of altimeter wave measurements into WAVEWATCH III. In: 8th International Workshop on Wave Hindcasting and Forecasting, North Shore, Oahu, Hawaii, 14–19 November 2004
- Xu L, Rosmond T (2004) Formulation of the NRL atmospheric variational data assimilation system—accelerated representer (NAVDAS-AR), NRL/MR/7532-36. Naval Research Laboratory, pp 28
- Xu L, Rosmond T, Daley R (2005) Development of the NAVDAS-AR: formulation and initial tests of the linear problem. *Tellus* 57A:546–559
- Yaremchuk M, Nechaev D, Panteleev G (2009) A method of successive corrections of the control subspace in the reduced-order variational data assimilation. *Mon Weather Rev* 137:2966–2978
- Yaremchuk M, Carrier M, Smith S, Jacobs G (2013) Background error correlation modeling with diffusion operators. In: Park SK, Xu L (eds) *Data Assimilation for Atmospheric, Oceanic and Hydrological Applications*, vol. 2, Springer, pp 177–203
- Yaremchuk M, Nechaev D (2013) Covariance localization with the diffusion-based correlation models. *Mon Weather Rev* 141:848–860
- Yaremchuk M, Sentchev A (2012) Multi-scale correlation functions associated with polynomials of the diffusion operator. *Q J R Meteorol Soc* 138:1948–1953

- Yaremchuk M, Smith S (2011) On the correlation functions associated with polynomials of the diffusion operator. *Q J R Meteorol Soc* 137:1927–1932
- Zhang F, Zhang M, Hansen JA (2009) Coupling ensemble Kalman filter with four-dimensional variational data assimilation. *Adv Atmos Sci* 26:19
- Zhang M, Zhang F (2012) E 4dVar: coupling an ensemble Kalman filter with four-dimensional variational data assimilation in a limited-area weather prediction model. *Mon Weather Rev* 140:587–600
- Županski M (2005) Maximum likelihood ensemble filter: theoretical aspects. *Mon Weather Rev* 133:1710–1726

Convergence of a Class of Weak Solutions to the Strong Solution of a Linear Constrained Quadratic Minimization Problem: A Direct Proof Using Matrix Identities

S. Lakshmivarahan

Abstract In this note we provide a direct proof of convergence of the sequence of penalty function based weak solutions to the strong solution of a quadratic minimization problem with linear constraints using two well-known matrix identities.

1 Introduction

Let $A \in \mathbb{R}^{n \times n}$ be a symmetric positive definite (SPD) matrix, $b \in \mathbb{R}^n$ and $c \in \mathbb{R}$. Consider the minimization of

$$J(x) = \frac{1}{2}x^T Ax - b^T x + c \quad (1.1)$$

when $x \in \mathbb{R}^n$ lies in the affine subspace defined by

$$d = Bx \quad (1.2)$$

where $d \in \mathbb{R}^m$ and $B \in \mathbb{R}^{m \times n}$ with $1 \leq m < n$. It is assumed that B is of full rank, that is, $\text{Rank}(B) = m$.

This constrained minimization problem in (1.1) and (1.2) is usually solved in one of two ways as (a) a strong constraint problem using Lagrangian multiplier method and (b) a weak constraint counterpart using the penalty function method (Sasaki 1970). For completeness, we provide a short resume of these two methods.

Dedicated to the memory of Professor Yoshi Sasaki.

S. Lakshmivarahan (✉)

School of Computer Science, University of Oklahoma, Norman, OK 73019, USA
e-mail: varahan@ou.edu

(a) **Strong constraint version:** Define the Lagrangian

$$L(x, \lambda) = J(x) + \lambda^T(d - Bx) \quad (1.3)$$

where $\lambda \in \mathbb{R}^m$ is the undetermined Lagrangian multiplier. A necessary condition for the minimum of the unconstrained problem in (1.3) (Luenberger 1984; Nash and Sofer 1996; Bazaraa et al. 2006) is given by

$$\begin{aligned} \nabla_x L(x, \lambda) &= Ax - b - B^T \lambda = 0 \\ \nabla_\lambda L(x, \lambda) &= d - Bx = 0 \end{aligned} \quad (1.4)$$

Solving (1.4), it can be verified that the minimizer x_s and λ_s are given by

$$\lambda_s = (BA^{-1}B^T)^{-1}(d - BA^{-1}b) \quad (1.5)$$

and

$$x_s = A^{-1}b + A^{-1}B^T(BA^{-1}B^T)^{-1}(d - BA^{-1}b) \quad (1.6)$$

which is the sum of the unconstrained minimum $A^{-1}b$ and the correction term arising from the linear constraints (1.2). Recall that $BA^{-1}B^T$ is the transformation of the SPD matrix A^{-1} onto the subspace generated by the columns of the full rank matrix B . Hence, $BA^{-1}B^T$ is also SPD.

(b) **Weak constraint version:** Let $\alpha > 0$ be a real parameter and define a penalty function

$$P(x, \alpha) = J(x) + \frac{\alpha}{2}(d - Bx)^T(d - Bx) \quad (1.7)$$

A necessary condition for the minimum of (1.7) is given by

$$\nabla_x P(x, \alpha) = Ax - b + \alpha B^T(Bx - d) = 0 \quad (1.8)$$

Solving (1.8), the minimizer $x(\alpha)$ is given by

$$x(\alpha) = x_1(\alpha) + x_2(\alpha) \quad (1.9)$$

with

$$x_1(\alpha) = (A + \alpha B^T B)^{-1}b \quad (1.10)$$

and

$$x_2(\alpha) = (A + \alpha B^T B)^{-1} \alpha B^T d \quad (1.11)$$

It stands to reason to expect that

$$\lim_{\alpha \rightarrow \infty} x(\alpha) = x_s \quad (1.12)$$

That is, the weak solution converges to the strong solution as the penalty parameter α increases without bound.

In this note, we prove the convergence in (1.12) using the well-known Sherman-Morrison-Woodbury (SMW) formula and a matrix identity derived from it. Appendix A contains a version of this formula and the matrix identity we propose to use in our proof.

We hasten to add that the general version of the above convergence result for the case of minimizing $f(x)$ when $h(x) = 0$, where $f: \mathbb{R}^n \rightarrow \mathbb{R}$ and $h: \mathbb{R}^n \rightarrow \mathbb{R}^m$ are sufficiently smooth, is contained in many textbooks—Chap. 12 in Luenberger (1984), Chap. 16 in Nash and Sofer (1996) and Chap. 9 in Bazaraa et al. (2006), to name a few. The special case where $f(x)$ is a quadratic in x and $h(x)$ is affine in x is tailor made for the application of the matrix identities as shown in Sect. 2.

2 Proof of Convergence in (1.12)

Applying the SMW formula in (A.1) in Appendix A to the inverse on the right hand side of (1.10) (by setting $H = B$, $\Sigma_v^{-1} = \alpha I_m$ and $\Sigma_x^{-1} = A$ in (A.1)) we obtain

$$\begin{aligned} (A + \alpha B^T B)^{-1} &= A^{-1} - A^{-1} B^T [BA^{-1} B^T + \alpha^{-1} I_m]^{-1} BA^{-1} \\ &\rightarrow A^{-1} - A^{-1} B^T (BA^{-1} B^T)^{-1} BA^{-1} \text{ as } \alpha \rightarrow \infty \end{aligned} \quad (2.1)$$

Now combining (2.1) with (1.10), it follows that

$$x_1^* = \lim_{\alpha \rightarrow \infty} x_1(\alpha) = A^{-1} b - A^{-1} B^T (BA^{-1} B^T)^{-1} BA^{-1} b \quad (2.2)$$

Applying the matrix identity (A.3) in Appendix A to the inverse on the right-hand side of (1.11) (by setting $H = B$, $\Sigma_v^{-1} = \alpha I_m$ and $\Sigma_x^{-1} = A$ in (A.3)) we get

$$\begin{aligned} (A + \alpha B^T B)^{-1} \alpha B^T &= A^{-1} B^T [BA^{-1} B^T + \alpha^{-1} I_m]^{-1} \\ &\rightarrow A^{-1} B^T (BA^{-1} B^T)^{-1} \text{ as } \alpha \rightarrow \infty \end{aligned} \quad (2.3)$$

Combining (2.3) with (1.11), it follows that

$$x_2^* = \lim_{\alpha \rightarrow \infty} x_2(\alpha) = A^{-1}B^T(BA^{-1}B^T)^{-1}d \quad (2.4)$$

It can be verified that x_s in (1.6) is indeed given by

$$x_s = x_1^* + x_2^* \quad (2.5)$$

and hence the claim.

3 Applications

For completeness, we mention two standard applications of the problem in (1.1) and (1.2). First, a special case of (1.1) with $A = I$, the identity matrix, $b = 0$ and $c = 0$ along with (1.2) arises as an underdetermined linear, least squares problem (Chap. 5 in Lewis et al. 2006). Second, consider

$$Q(x) = \frac{1}{2}(z - Hx)^T(z - Hx) \quad (3.1)$$

which represents the miss-fit measured by the sum of the squared errors between observations, $z \in \mathbb{R}^p$ and its model counterpart, Hx where $H \in \mathbb{R}^{p \times n}$ and $x \in \mathbb{R}^n$ when x is required to satisfy a constraint of the form (1.2) which could represent the basic geostrophic constraint or mass continuity (Sasaki 1970; Lewis and Lakshmivarahan 2008).

Acknowledgements We wish to express our thanks to John Lewis for his interest in this work and to Trung Nguyen and Junjun Hu for their help with typesetting this paper.

Appendix A

Let $H \in \mathbb{R}^{m \times n}$, $\Sigma_x \in \mathbb{R}^{n \times n}$ and $\Sigma_v \in \mathbb{R}^{m \times m}$ where $m < n$, and Σ_x and Σ_v are non-singular matrices. Then the well-known Sherman-Morrison-Woodbury (SMW) formula is given by Golub and Van Loan (1989)

$$[H^T \Sigma_v^{-1} H + \Sigma_x^{-1}]^{-1} = \Sigma_x - \Sigma_x H^T [H \Sigma_x H^T + \Sigma_v]^{-1} H \Sigma_x \quad (A.1)$$

or equivalently as

$$[H\Sigma_x H^T + \Sigma_v]^{-1} = \Sigma_v^{-1} - \Sigma_v^{-1} H [H^T \Sigma_v^{-1} H + \Sigma_x^{-1}]^{-1} H^T \Sigma_v^{-1} \quad (\text{A.2})$$

Multiplying both sides of (A.2) on the left by $\Sigma_x H^T$ and simplifying (Chap. 17, Lewis et al. (2006) we obtain the matrix identity

$$\Sigma_x H^T [H\Sigma_x H^T + \Sigma_v]^{-1} = [H^T \Sigma_v^{-1} H + \Sigma_x^{-1}]^{-1} H^T \Sigma_v^{-1} \quad (\text{A.3})$$

References

- Bazaraa MS, Sherali HD, Shetty CM (2006) Nonlinear programming: theory and algorithm. John Wiley & Sons, New York, USA
- Golub G, Van Loan C (1989) Matrix computation. Johns Hopkins University Press, Baltimore, MD, USA
- Lewis JM, Lakshmivarahan S (2008) Sasaki's pivotal contribution: calculus of variations applied to weather map analysis. *Mon Weather Rev* 136:3553–3567
- Lewis JM, Lakshmivarahan S, Dhall SK (2006) Dynamic data assimilation: a least squares approach. Cambridge University Press, New York, USA
- Luenburger DG (1984) Introduction to linear and nonlinear programming. Addison-Wesley, Reading, MA, USA
- Nash SG, Sofer A (1996) Linear and Nonlinear Programming. McGraw Hill, New York, USA
- Sasaki Y (1970) Numerical variational analysis with weak constraint and application to surface analysis of severe storm gust. *Mon Weather Rev* 98:899–910

Information Quantification for Data Assimilation

Sarah King, Wei Kang, Liang Xu and Nancy L. Baker

Abstract In this paper we discuss the application of observability as a measurement of observation quality for data assimilation in numerical weather prediction (NWP). Observability is a measure of well-posedness of a dynamical system and provides a flexible framework to address questions in data assimilation. We review the concept of observability for differential equations and high dimensional numerical models. We discuss the relationship of observability to observation impact. We conclude with a discussion of various applications of observability to data assimilation including optimal sensor placement and data thinning.

1 Introduction

There are many outstanding questions in data assimilation related to observations. For example: can we tell in advance the impact gaining or losing an instrument will have on our ability to determine the current systems state? What types of new observations would assist in improving the quality of our assimilation? If we can only use a limited number of observations which ones should we use? Currently there are partial answers to these questions based on observation impacts (Baker and Daley 2000) or expected error (Majumdar et al. 2001) for various data assimilation methodologies used in numerical weather prediction (NWP). For instance, observation impacts (Langland and Baker 2004) only determine the effect of observations after the fact. A different approach developed in the last few years, inspired by some fundamental concepts in the fields of control theory and dynamical systems, could help answer these questions.

Observability has long been used in the control theory community as a measure of how well we can infer information on a system state based on sensor and non-sensor

S. King (✉) · L. Xu · N.L. Baker
Naval Research Laboratory, Monterey, CA, USA
e-mail: Sarah.King@nrlmry.navy.mil

W. Kang
Applied Mathematics Department, Naval Postgraduate School, Monterey, CA, USA

knowledge. In the context of data assimilation it provides us with a measure of how well we can “see” what is going on in numerical models. In Kang and Xu (2009a, b) the concept of observability was introduced for large-scale nonlinear systems such as atmospheric models. The empirical gramian matrix was introduced in Krener and Ide (2009), which is an effective computational method. In Kang and Xu (2012) the problem of finding the optimal sensor location was studied based on the idea of maximizing the observability. The results were illustrated using an example of Burgers’ equation. The numerical experiments show that the data collected at optimal sensor locations improves the accuracy of data assimilation. Similar results were achieved for a shallow water model in King et al. (2014). To make computations efficient for large scale systems, in King et al. (2013), an algorithm to compute observability was developed in which the tangent linear model was employed to calculate the empirical gramian matrix. Some issues on the theoretical foundation of observability for partial differential equations were proved in Kang and Xu (2014).

In this paper, we conduct a comprehensive study of several issues and applications of observability for data assimilation. After a brief introduction of basic concepts, such as sensitivity, observation impact, and observability, the problem of finding optimal sensor locations is defined and a computational algorithm is introduced. We illustrate these concepts using a system of shallow water equations. To validate the effectiveness of observability based optimal sensor locations for data assimilation, the observation impact is computed. The shallow water example shows that optimal sensor locations improve the observation impact and data assimilation accuracy. In addition to sensor locations and observation impact, we also explore the idea of using observability as a metric in data thinning. Based on the determinant of the observability gramian and a quasi-Newton optimization method, the optimal sampled data set is compared to the data from other thinning methods. The comparison is based on the output of a 4D-Var data assimilation for the shallow water equations.

2 Review of Observability

Observability is an inherent property of the dynamical system and not of the estimation or assimilation algorithm and has implications for the performance of assimilation algorithms. As such it cannot be altered by different assimilation systems but may be modified by changing the dynamical system. For the ease of our discussion we consider the linear time-invariant system, i.e. $x_k = M^k x(0)$, given by

$$\begin{aligned} x(k+1) &= Mx(k) \\ y(k) &= Hx(k) \end{aligned} \tag{1}$$

where x describes the system state, y describes an observation process, M is a linear model, and H is the observation operator. Consider the set of observations $\{y(0), y(1), y(2), y(3), \dots\}$ where

$$\begin{aligned}
y(0) &= Hx(0) \\
y(1) &= Hx(1) = HMx(0) \\
y(2) &= Hx(2) = HM^2x(0) \\
y(3) &= Hx(3) = HM^3x(0) \\
&\vdots
\end{aligned}$$

which may be rewritten as

$$\begin{bmatrix} H \\ HM \\ HM^2 \\ HM^3 \\ \vdots \end{bmatrix} x(0) = \begin{bmatrix} y(0) \\ y(1) \\ y(2) \\ y(3) \\ \vdots \end{bmatrix}.$$

If the matrix

$$\mathcal{O} = [H^T, M^T H^T, (M^2)^T H^T, (M^3)^T H^T, \dots]^T \quad (2)$$

is full rank then we can uniquely determine $x(0)$ and we say (M, H) is observable. As this definition suggests observability can be seen as a measure of well-posedness. To further demonstrate the relationship of observability to data assimilation consider the filtering problem

$$\begin{aligned}
\hat{x}(k+1) &= M\hat{x}(k) + K(y(k) - \hat{y}(k)) \\
\hat{y}(k) &= H\hat{x}(k)
\end{aligned} \quad (3)$$

where \hat{x} is the estimated state and K is the gain. The estimation error $e(k+1) = x(k+1) - \hat{x}(k+1)$ is equivalently given by

$$e(k+1) = (M - KH)e(k) \quad (4)$$

which approaches zero if the eigenvalues of $M - KH$ are in the left half plane. Note that (3) is equivalently written as

$$x_a = x_b + K(y - Hx_b)$$

with estimation error, or analysis error, $e = x_{truth} - x_a$. Let us assume for a moment (M, H) not fully observable, i.e., \mathcal{O} has zero eigenvalues. There is a similarity transform Q such that

$$\hat{M}^T = Q^{-1}M^T Q = \begin{bmatrix} M_1 & M_{21} \\ 0 & M_2 \end{bmatrix} \text{ and } \hat{H}^T = Q^{-1}H^T = \begin{bmatrix} H_1 \\ 0 \end{bmatrix}.$$

Then

$$\begin{aligned} Q^{-1}(M - KH)^T Q &= Q^{-1}M^T Q - Q^{-1}H^T K^T Q \\ &= \begin{bmatrix} M_1 & M_{12} \\ 0 & M_2 \end{bmatrix} + \begin{bmatrix} H_1 \\ 0 \end{bmatrix} [K_1, K_2] \end{aligned}$$

where $[K_1, K_2] = KQ$. So if (M, H) is not fully observable then the eigenvalues of M_2 are unaffected by K and if the eigenvalues of M_2 are not stable the estimation error will not vanish. However, if we have full state observability there exists a gain K where $e(k) \rightarrow 0$ as $k \rightarrow \infty$. To summarize, observability gives us a sufficient condition for the convergence of a filter. Note that typically variational forms of assimilation have a filter form so this is applicable to them as well.

2.1 Comparison to Observation Sensitivity and Observation Impact

Sensitivity and observation impacts are properties of the data assimilation system and the observed states, and not solely the dynamics. The quantitative value of an observation in data assimilation varies based on the content of the observation and the assimilation method used. Observation sensitivity measures the potential effect of an observation while observation impact measures the realization of the effect of an observation. Sensitivity, as the gradient of the atmospheric analysis with respect to observations, was introduced in Baker and Daley (2000) as a method for targeted observations. Observation impacts estimate the effect an observation has on the short-range forecast error and is routinely used to evaluate observation performance (Langland and Baker 2004). For our discussion we will review these concepts in order to compare them to observability. Assume the analysis or state estimate is given by

$$x_a = x_b + K(y - Hx_b) \quad (5)$$

where x_a is the analysis and x_b is the background with the Kalman gain

$$K = P_b H^T (H P_b H^T + R)^{-1}.$$

We can rewrite (5) as

$$x_a = (I - KH)x_b + Ky. \quad (6)$$

Differentiating (6) with respect to the observations yields

$$\frac{\partial x_a}{\partial y} = K^T$$

providing the sensitivity of the analysis to the observations. The sensitivity of a scalar function J with respect to the observations is obtained via

$$\frac{\partial J}{\partial y} = \frac{\partial J}{\partial x_a} \frac{\partial x_a}{\partial y} = K^T \frac{\partial J}{\partial x_a} = (HP_b H^T + R)^{-1} H P_b \frac{\partial J}{\partial x_a} \quad (7)$$

where

$$\frac{\partial J}{\partial x_a} = M^T \frac{\partial J}{\partial x}.$$

The observation sensitivity (7) may identify regions of high sensitivity to be used for targeted observations. To measure observation impact we consider the error in a forecast x_f measured against an analysis, x_t , at time t

$$e_f = \langle (x_f - x_t), C(x_f - x_t) \rangle \quad (8)$$

where C is the energy norm weights. Observation impact is computed in the context of two forecasts: the first forecast at time g is the prior or background for a later forecast at time f which is a corrected forecast based on new initial conditions from an assimilation cycle. The difference in these forecasts at a verifying time t arises from the inclusion of observations and as such we expect the forecast at time f to contain less error. The difference in error between two forecasts is defined as

$$\Delta e_f^g = e_f - e_g \quad (9)$$

and we expect $e_f < e_g$ and so this value should be negative. If we consider the case of (3) with error modeled by (4) this amounts to the effect of KH ; the gain in the model space on the forecast error. Our interest is in the forecast error so we define our scalar measure J in (7) as $J_f = \frac{1}{2}e_f$ then

$$\frac{\partial J_f}{\partial x_f} = C(x_f - x_t).$$

We may use M^T to map the gradient to the verification time t . To estimate (9) in terms of sensitivities we begin with

$$\Delta e_f^g = \left\langle (x_f - x_g), \frac{\partial J_f}{\partial x_f} + \frac{\partial J_g}{\partial x_g} \right\rangle. \quad (10)$$

The difference between the forecast trajectories is the analysis increment $x_a - x_b$ which evolves linearly so we may write (10) as

$$\begin{aligned}
\delta e_f^g &= \left\langle (x_a - x_b), \frac{\partial J_f}{\partial x_a} + \frac{\partial J_g}{\partial x_b} \right\rangle \\
&= \left\langle K(y - Hx_b), \frac{\partial J_f}{\partial x_a} + \frac{\partial J_g}{\partial x_b} \right\rangle \\
&= \left\langle y - Hx_b, K^T \left(\frac{\partial J_f}{\partial x_a} + \frac{\partial J_g}{\partial x_b} \right) \right\rangle
\end{aligned}$$

and noting

$$\frac{\partial J_f^g}{\partial y} = K^T \left(\frac{\partial J_f}{\partial x_a} + \frac{\partial J_g}{\partial x_b} \right)$$

gives the relation

$$\delta e_f^g = \left\langle y - Hx_b, \frac{\partial J_f^g}{\partial y} \right\rangle. \quad (11)$$

The observation impact is the improvement in the estimation error between two forecasts attributable to the Kalman gain whereas observability limits how much we can improve the estimation error with the Kalman gain.

3 Partial Observability

The previous conventional definition of observability based on the observability matrix (2) characterizes the concept for the full system state. However, if the dimension of a system is very high, it may not be possible or desirable to achieve observability in the entire state space. In general a family of solutions of a PDE is infinite dimensional so it is perhaps more appropriate to look at a subspace that consists of finite number of modes.

Partial observability provides a quantitative assessment of observability on a low dimensional subspace. This quantitative metric has the advantage that it also indicates of the degree of observability. The definition of observability in terms of dynamic optimization was first introduced in Kang and Xu (2009a, b) and extended to systems defined by PDEs in Kang (2011), Kang and Xu (2014), and King et al. (2013). We now consider the nonlinear, time varying, dynamical system defined by the PDE

$$\begin{aligned}
\dot{u}(t) &= f(u, x, t), \quad t \in [0, T] \\
u(x, 0) &= u_0
\end{aligned} \quad (12)$$

with output denoted by

$$y(t) = h(u(x, t)) \quad (13)$$

where $u(\cdot, t)$ is, for any fixed t , in a function space. It represents the state of the system or model and is measured by sensors. The output $h(\cdot)$ is the sensor measurement. The variation of the measured variable $y(t)$ under the variation of u_0 , the initial conditions, is

$$J(u_0, \delta u_0) = \delta u_0^T P_1 \delta u_0 + ||y(\cdot; u_0 + \delta u_0) - y(\cdot; u_0)||_{P_2} \quad (14)$$

where P_1 and P_2 are weight matrices. The norm in output space is defined as

$$||y(\cdot)||_{P_2} = \int_0^T y(t)^T P_2 y(t) dt.$$

Let

$$W = \text{span}(w_1, \dots, w_s)$$

be the space for estimation which is a finite dimensional subspace of the state space. For a solution to (12) and (13), we have a best estimate $u_W \in W$, which is defined to be the state $u_0 + \delta u_0$ in W that minimizes J in (14). If $u_W - u_0$ is small, then it is practically good enough to achieve strong observability in W and ignore the rest of the state space. The selection of the dimension of W is based on a number of factors, such as required accuracy, which we will discuss in relation to our shallow water equations example in Sect. 4. Assume that W has a norm denoted by $|| \cdot ||_W$. We define the partial observability as follows.

Definition 3.1 Let $\rho > 0$ be a positive number. Then the number ϵ is defined as

$$\begin{aligned} \epsilon^2 &= \min_{\delta u_0} J(u_0, \delta u_0) \\ \text{subject to } &||\delta u_0||_W = \rho, \quad \delta u_0 \in W \end{aligned} \quad (15)$$

then the ratio ρ/ϵ is called the unobservability index (Kang and Xu 2009a).

The ratio ρ/ϵ measures the sensitivity of output y to the variation of the state around u_0 . From an estimation viewpoint this implies that if the maximum sensor error is ϵ , the worst possible estimation error of the initial condition u_0 is ρ . Therefore a small ρ/ϵ implies strong observability of u_0 . Solving the problem of dynamic minimization in (15) is challenging. We approximate (15) by using an empirical covariance matrix. The variation of δu_0 on the sphere $||\delta u_0||_W = \rho$ can be represented by

$$\delta u_0 = \sum_{i=1}^s c_i w_i$$

where $\{w_i\}_{i=1}^s$ is an orthonormal basis and

$$\sum_{i=1}^s c_i^2 = \rho^2.$$

Using the output variation in the form of $\delta u_0 = \rho w_i$, an empirical gramian can be defined, which is denoted by G . Details about the gramian can be found in Appendix A. Then we have

$$\begin{aligned} \epsilon^2 &\approx \min_{\sum c_k^2 = \rho^2} [w_1 \ w_2 \ \dots \ w_s] G [w_1 \ w_2 \ \dots \ w_s]^T \\ &= \sigma_{\min} \rho^2 \end{aligned} \quad (16)$$

where σ_{\min} is the smallest eigenvalue of G . This approach is closely related to conventional control theory. In fact, for linear systems the gramian G is equal to the standard observability gramian (Kang et al. 2013) if W spans the entire state space. The gramian matrix depends on W and its norm. For the shallow water equations, which will be discussed in the following section, W and $\|\cdot\|_W$ are defined in a similar way as in King et al. (2013). Some details are given in Appendix A.

4 Observability Optimization

In this section we focus on specific applications of observability in data assimilation. Observability can be used to discuss the merits of current observing networks and identify weaknesses of the current networks (King et al. 2014). It can also be used to determine the potential of new observations. We discuss the use of observability primarily as a metric for optimization focusing on the placement of targeted observations and data thinning for high resolution observations. We use the shallow water equations (SWEs) governed by

$$\frac{\partial v}{\partial t} = -v \frac{\partial v}{\partial x} - g \frac{\partial h}{\partial x} \quad (17)$$

$$\frac{\partial h}{\partial t} = -v \frac{\partial h}{\partial x} - h \frac{\partial v}{\partial x} \quad (18)$$

where v is the velocity, h is the height, and g is gravity. To discretize (17)–(18) we use a staggered (“Arakawa C”) grid. For the time integration we use a leap-frog scheme with zero velocity at the boundary. We start with an exponential initial condition and as we move forward in time two waves form and move away from each other, see Fig. 1. For longer time intervals these waves reflect back. Optimizing the observability of a system allows us to improve the conditioning and estimation error of a system by modifying H in (5). To optimize the observability of stationary configurations we decrease the ratio ρ/ϵ by increasing ϵ which leads to the max-min problem

$$\max_{\lambda} \min_{\delta u_0} J(u_0, \delta u_0, \lambda) \text{ subject to } \|\delta u_0\| = \rho \quad (19)$$

where $u = (v, h)$ and λ represents the sensor locations. This may also be formulated to accommodate moving sensor platforms to determine the optimal trajectory by

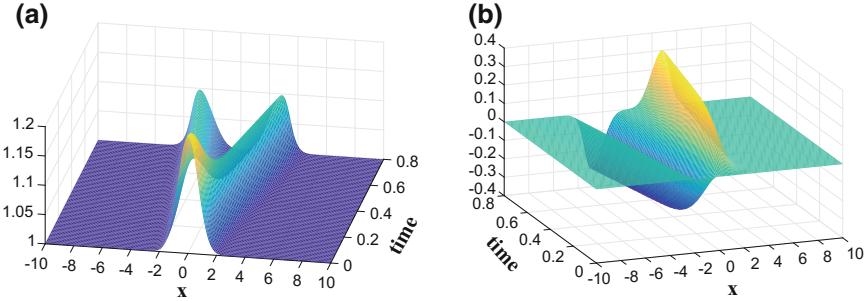


Fig. 1 The solutions for **a** the height h and **b** the velocity v after for 80 time steps

solving the max-min problem with an additional constraint governing the motion of the sensors given by

$$\begin{aligned} & \max_{\lambda} \min_{\delta u_0} J(u_0, \delta u_0, \lambda) \\ \text{subject to } & \frac{d\lambda}{dt} = \eta(\lambda(t)) \text{ for all } t \in [t_o, t_0 + T] \text{ and } \|\delta u_0\| = \rho. \end{aligned}$$

While these optimization problems may be solved iteratively for small dimension problems it is very costly for high dimensional problems as this could require many model evaluations. The tangent linear model (TLM), readily available in many NWP centers, may be used in calculation of the variations to provide scalability to computations.

4.1 Sensor Placement

We begin by considering the design of an observing network, specifically, given sensors where are the best places for them. For this experiment we consider the case with a limited number of sensors so we look at maximizing the smallest eigenvalue of (16) which is essentially optimizing the worst case scenario. We generate one hundred sets of initial backgrounds by randomly perturbing the truth $u(0)$ such that $u_j^b(0) - u(0)$ has a Gaussian distribution. For each background we generate sensor information by adding white Gaussian noise:

$$y(t_i) = y^{true}(t_i) + R^{1/2} \xi_i$$

where y^{true} is the value of the true state, ξ_i is white Gaussian noise, and R is the sensor variance. For each set of initial conditions and sensor information we perform four dimensional variational (4D-Var) type assimilation over the time interval. We assume the background error covariance, P_b , is unknown but is of the form

$$P_b = b_0 \exp \left(\frac{-|i - i'|^2}{L^2} \right)$$

where b_0 represents the variances of the initial errors, i is the spatial index, and L is the spatial decorrelation length. The parameters we use are as follows:

$N_\lambda = 6$	number of sensors
$L = 20$	length of x interval
$T_A = 0.8$	length of assimilation window
$T_F = 1.6$	forecast time
$\rho = 0.01$	observability tolerance
$R_h = 1e-5$	variance in height measurements
$R_u = 1e-5$	variance in velocity measurements
$b_0 = 1e-4$	used to compute P_b .

We assume each sensor measures the height and velocity. Intuitively, based on the discussion in Sect. 2.1, we expect that increasing the observability may lead to increased observation impact while the converse is not necessarily true. As these concepts measure different aspects of the dynamical and assimilation systems it is hard to quantify their exact effect on each other.

To explore the relationship between these concepts we consider the case where we add sensors to an existing configuration. We start with an existing configuration of three equally spaced sensors and add optimally placed sensors to create a new network. We then compute the observation impact via (11). We optimize the locations of three sensors using the Fourier space as our estimation space W , see Appendix A for more details. Generally, the number of frequencies used in the Fourier space should be based on the desired accuracy of the state space estimation. For our example the higher frequencies suggested by the state space estimation were effectively unobservable so we focus on the observable frequencies. We use two frequencies ensuring all modes in our estimation space are observable. It is possible that if we were optimizing all six sensors at the same time more modes would be observable and we could potentially use more frequencies.

To solve for the optimal sensors we use the empirical gramian which reduces (19) to an eigenvalue maximization problem

$$\begin{aligned} & \max_{\lambda} \sigma_{\min}(G(\lambda)) \\ & \text{subject to } \lambda_{\text{lower}} < \lambda < \lambda_{\text{upper}} \end{aligned} \tag{20}$$

where σ_{\min} is the smallest eigenvalue of the gramian matrix G . The derivation of G for the shallow water equations is in Appendix A. Note that we solve for all λ simultaneously. Optimizing (20) can be numerically challenging due to the nonsmooth nature of the objective function, i.e., it is not necessarily differentiable at its extrema because of the tendency of eigenvalues to coalesce at solution points. To find the conditions necessary for optimality let (σ, λ) be an eigenpair for the gramian G then

$$G(\lambda)x = \sigma x.$$

Taking the derivative with respect to λ leads the expression

$$G \frac{dx}{d\lambda} + \frac{dG}{d\lambda}x = \sigma \frac{dx}{d\lambda} + \frac{d\sigma}{d\lambda}x.$$

Since G is symmetric the left eigenvector of σ is given by x^T so our expression for the derivative reduces to

$$\frac{d\sigma}{d\lambda} = x^T \frac{dG}{d\lambda} x = 0$$

assuming $x^T x = 1$. We elected to solve this eigenvalue problem using the Broyden-Fletcher-Goldfrab-Shanno (BFGS) method. It is by far the most popular quasi-newton method and has proven effective for nonsmooth, nonconvex problems (Lewis and Overton 2012). As this problem is nonsmooth we are only guaranteed a local minimizer. Additionally, care must be taken with the selection of a line search method in order to avoid the situation of vanishing derivatives.

When determining the optimal sensors to be added to the existing sensors we use the entire set of sensors in our observability computations giving us the sensor configuration in Fig. 2. This configuration with all six sensors was then used in one hundred 4D-Var data assimilations and the average observation impacts determined in Table 1. As a group, the equally placed sensors have a less negative average observation impact which implies they do not improve the analysis as much as the optimal sensors which have a larger negative impact implying they improve the analysis more.

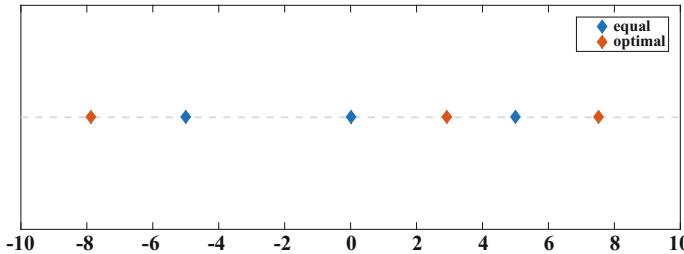


Fig. 2 Sensor configuration after the optimal sensors were added to the equally spaced ones

Table 1 Averaged observation impacts

Configuration	Observation impact
Equal	-0.0298
Optimal	-0.0597

These results indicate that we have a greater performance from the optimal sensors added to the existing framework.

We now consider the case where there is no previous observing network and we optimally select a network configuration. We solve (20) with the parameters

$$\begin{aligned} N_\lambda &= 6 && \text{number of sensors} \\ N_F &= 4 && \text{number of frequencies for each } u, \phi \end{aligned}$$

with the remaining parameters unchanged. As we can now adjust the positions of all six sensors we are able to include higher frequencies with all of them being observable. It is possible to target specific frequencies in the estimation space W giving the flexibility to target specific features. We are interested in the overall estimation error so we choose the first four frequencies.

The optimal sensor locations computed can be seen in Fig. 3. The solution of (20) is not guaranteed to be globally optimal so it is possible that a different global solution exists. Using a different initial guess may produce a different solution entirely. In addition to the optimal solution, we will be comparing our results to a random configuration and an equally spaced configuration. The random configuration was produced by taking ten random configurations and then choosing the configuration with the lowest RMSE to compare our results against. For our experiment we use the same one hundred initial backgrounds and sensor information as previously used. We estimate the overall error of the analysis $u^a(t)$ using the following norm

$$||u^a - u^{truth}|| = \int_0^T ||u^a(t) - u^{truth}(t)||^2 dt$$

where $|| \cdot ||^2$ is the L^2 norm with the RMSE defined as

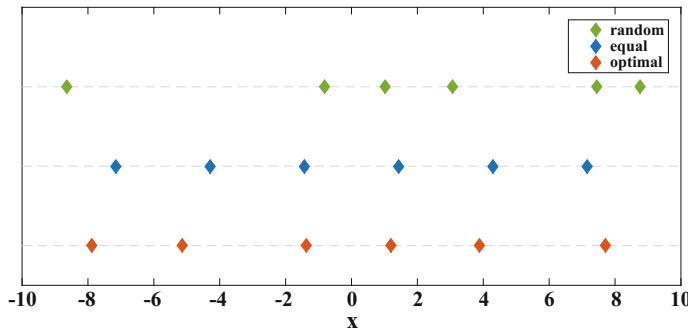


Fig. 3 The sensor locations for the random, equally spaced, and optimal sensor configurations

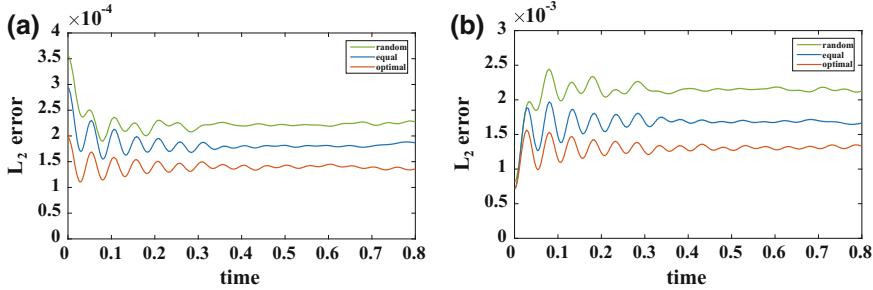


Fig. 4 **a** The average L^2 error over time in the height for the equally spaced and optimal sensors and **b** the average L^2 error for the velocity

Table 2 RMES error, N=100

Configuration	Height	Velocity
Random	0.0797	0.2445
Equal	0.0721	0.2168
Optimal	0.0629	0.1915

$$\sqrt{\frac{\sum_{k=1}^N ||u_k^a - u^{truth}||^2}{N}}.$$

We can see the RMSE of the different configurations as a function of time in Fig. 4 in terms of both wave height and velocity. After $t = 0.3$ the error in all cases stabilizes. Table 2 gives the RMSE over the entire data assimilation window. Compared to the equally placed sensors the optimal sensors saw performance improvements of 13 % for the height and 12 % for the velocity. Compared to the randomly placed sensors there was a performance improvement of 21 and 22 % for the height and velocity, respectively. For this experiment the optimally placed sensors outperformed equally spaced and random configurations. This indicates that increasing the observability improves the quality of the analysis.

The worst case scenario may be significantly improved but the overall quality of your estimation may not improve as significantly due to the nature of optimizing the worst case scenario. Depending on the situation other measures of observability may yield better results. For example, looking at the determinate of the gramian is more appropriate for domains with good coverage as we will see in the next example.

4.2 Data Thinning

Instruments such as satellites, radar, and radiosondes produce high density data sets that are of high value in data assimilation. However, they are computationally expensive and there are underlying theoretical constraints that pose issues. Data thinning schemes are used to reduce the computational cost as well as to eliminate temporal and spatial correlations the observations. Uniform sampling or super ob-ing are typically used operationally but other adaptive schemes based on singular vectors (Bauer et al. 2011) or estimation error (Ochotta et al. 2005) are available. Observability as a metric for data thinning has the advantage of being assimilation and observation independent.

For the previous experiments, we assumed that observations are somewhat sparse and we sought to improve the worst case scenario by looking at the minimum eigenvalue. For data thinning we will be using the determinate of the observability gramian to measure the overall quality of the system via

$$\begin{aligned} & \max \log(\det(G(\lambda))) \\ & \text{subject to } \lambda_{\text{lower}} < \lambda < \lambda_{\text{upper}}. \end{aligned}$$

This formulation looks at the entire spectrum of the gramian G . Care must be taken when using this approach as a large determinate may mask small eigenvalues, i.e., unobservable directions of error. We use the logarithm of the determinate for computational ease. We continue to use BFGS in our computations noting that now our problem is smooth and therefore solved more easily. For this example we are assuming that the observations are dense in time or space. If the observations are not dense, then this method would not be appropriate and a combinatorial optimization technique should be utilized. We use a dense data set of ninety nine sensors with the parameters

$$\begin{aligned} N_\lambda &= 12 \quad \text{number of sensors in the reduced set} \\ N_F &= 9 \quad \text{number of frequencies for each } u, \phi. \end{aligned}$$

The number of observations kept may be determined by a number of factors, for example, available resources or desired estimation error. For our example we selected twelve sensors because given the size of our problem if we choose more sensors the effect of the sensor placement is negligible. This is a considerable decrease of about 88 % of available observations. We compare different methods of observation sampling as in Fig. 5. For the dense set of ninety nine sensors the RMSE is 0.0091 and 0.0277 for the height and velocity, respectively. Using the optimal sensors the error increased about four times the amount in the dense set but decreased the number of observations used by about seven eighths. For comparison we have a randomly sampled set that has the lowest RMSE of ten randomly sampled sets. We also compare our results against an equally sampled set. We compute the RMSE as before over the time interval in Table 3.

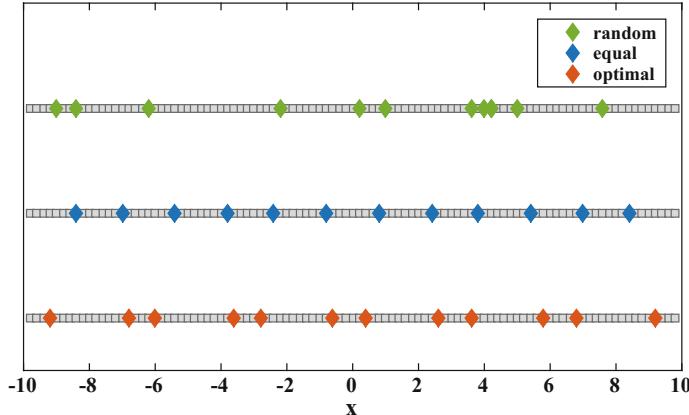


Fig. 5 Sensor locations for the dense (gray) and thinned data

Versus the random sample the optimal sample saw a 21.4 and 20.8 % improvement for the height and velocity respectively. Compared to the equally sampled set, the optimal saw an improvement of 10.1 and 8.7 % for height and velocity, respectively. For larger systems it would be possible to use an observability based approach that maintains a certain level of estimation error.

5 Final Remarks

We have discussed observability as a versatile tool for data assimilation: it can predict the potential effect of losing or gaining a sensor, it can be used to optimize observing networks, and it can be used to optimally reduce dense observations. Observability is a property inherent in a dynamical system which in turn affects data assimilation systems. As was discussed in Sect. 1, observability can be viewed as a measure of well-posedness. In our discussion of applications of observability we have used it as a metric for the optimization of observing networks. Optimization based on observability improves the conditioning of the dynamical system by adjusting H in (5) which may lead to an improvement of the data assimilation system regardless of the assimilation technique used. In our examples by improving the observation locations we were able to improve the estimation error of a 4D-Var system.

We have discussed observability's relationship to other quantification tools in data assimilation, namely, observation impacts. While observability has the potential to influence observation impacts these concepts are measuring different aspects of a data assimilation system. Observability is determined by the NWP model and observation operator and is independent of the information contained in observations,

Table 3 RMES error, N=100

Configuration	Height	Velocity
Random	0.0489	0.1493
Equal	0.0427	0.1294
Optimal	0.0384	0.1182

where as sensitivity and observation impact measure the effect of the information provided by the content of the observations. Observability is more appropriate for information quantification a priori whereas observation impact is an a posteriori measurement.

A Empirical Gramian

To compute the gramian G in (16) we let $w_i \in W$, for $i = 1, \dots, s$, be an orthonormal basis. Consider the variation in the direction of w_i

$$u_0 + \delta u_0 = u_0 \pm \rho w_i.$$

Then, the variation of the output is

$$\Delta y_i(t) = \frac{1}{2\rho} (y(t; u_0 + \rho w_i) - y(t; u_0 - \rho w_i)). \quad (21)$$

To evaluate (21), we perform an integration for $u_0 + \rho w_i$ and $u_0 - \rho w_i$ using the full nonlinear model. Then

$$G_{ij} = \langle \Delta y_i(k), \Delta y_j(k) \rangle_{P_2}$$

where

$$\langle \Delta y_i(k), \Delta y_j(k) \rangle = \int_0^t \Delta y_i(t)^T P_2 \Delta y_j(t) dt.$$

Alternatively we can use the tangent linear model (TLM) which is a linearization of the nonlinear model around a state u . Letting $t_k = k\Delta t$

$$u_k = M_{k-1} u_{k-1}, \quad u_k \in \mathbb{R}^n \quad (22)$$

and

$$y_k = H u_k + V_k, \quad R_k = E(V_k V_k^T), \quad y_k \in \mathbb{R}^p \quad (23)$$

where R_k is the covariance matrix for h_k . Additionally we note that

$$\begin{aligned}
\Delta y_0^i &= H[(u_0 + \rho w_i) - u_0] &= \rho H w_i \\
\Delta y_1^i &= H M_0 (u_0 + \rho w_i) - H M_0 u_0 = \rho H M_0 w_i \\
&\vdots \\
\Delta y_k^i &= \rho H M_{k-1} \cdots M_0 w_i.
\end{aligned} \tag{24}$$

The TLM reduces the number of model evaluations by propagating the perturbations in time and we avoid the centered differences approximation in (21). For the shallow water equations in (17) and (18), the state space is given by

$$[h, v]^T = [h_0 \ h_1 \ \dots \ h_N \ v_0 \ \dots \ v_N]^T \text{ and } X = [x_0 \ \dots \ x_N]^T$$

where $h_i = h(x_i)$, $v_i = v(x_i)$. To compute the partial observability we use the Fourier space as our estimation space W represented by

$$\begin{aligned}
h &= \frac{a_0}{\sqrt{2}} + \sum_{k=1}^K \left(a_k \cos \left(\frac{2k\pi}{L} X \right) + b_k \sin \left(\frac{2k\pi}{L} X \right) \right) \\
v &= \frac{\alpha_0}{\sqrt{2}} + \sum_{k=1}^K \left(\alpha_k \cos \left(\frac{2k\pi}{L} X \right) + \beta_k \sin \left(\frac{2k\pi}{L} X \right) \right)
\end{aligned}$$

where $a, b, \alpha, \beta \in \mathbb{R}^n$. Suppose the tolerance of error for h and v are ρ_h and ρ_v then we define the norm in W as

$$||(a_0, a_1, \dots, a_k, b_0, \dots, b_k, \alpha_1, \dots, \alpha_k, \beta_0, \dots, \beta_k)^T||_W^2 = \sum_{k=0}^K a_k^2 + \sum_{k=1}^K b_k^2 + \frac{\rho_h^2}{\rho_u^2} \left(\sum_{k=1}^K \alpha_k^2 + \sum_{k=1}^K \beta_k^2 \right).$$

The output is denoted by

$$y = s_{output}(h, v, \lambda)$$

where λ is the sensor locations. We assume that the sensors measure both h and v at the location λ . In our examples if λ is not located at a grid point in X , the sensor measurement is modeled as an interpolation operator, s_{output} , using the value of v and h at the nearby grid points. To form the cost function (14) consider the nominal trajectory $(h(t), v(t))$ with nominal coefficients (a, b, α, β) and perturbation given by

$$(a, b, \alpha, \beta)^T + (\Delta a, \Delta b, \Delta \alpha, \Delta \beta)^T \tag{25}$$

with a scaled scalar tolerance

$$||(\Delta a^T, \Delta b^T, \Delta \alpha^T, \Delta \beta^T)^T||_W = \rho.$$

Let $(\hat{h}(t), \hat{v}(t))$ be the trajectory associated with (25) with output $\hat{y}(t)$. To compute the variation define

$$\begin{bmatrix} \Delta h \\ \Delta v \end{bmatrix} = \begin{bmatrix} \hat{h}(t) \\ \hat{v}(t) \end{bmatrix} - \begin{bmatrix} h(t) \\ v(t) \end{bmatrix} \text{ with } \Delta y_i = \hat{y}(t_i) - y(t_i)$$

then the variation (14) for (17) and (18) is given by

$$\begin{aligned} J &= \frac{1}{2(N+1)} \begin{bmatrix} \Delta h(0) \\ \Delta v(0) \end{bmatrix}^T P_b^{-1} \begin{bmatrix} \Delta h(0) \\ \Delta v(0) \end{bmatrix} + \frac{1}{N_t+1} \sum_{i=1}^{N_t} \Delta y_i^T R_i^{-1} \Delta y_i \\ &= [\Delta h(0) \ \Delta v(0) \ \Delta Y]^T \begin{bmatrix} P_b^{-1} & 0 \\ \frac{1}{2(N+1)} & 0 \\ 0 & \frac{R^{-1}}{N_t+1} \end{bmatrix} \begin{bmatrix} \Delta h(0) \\ \Delta v(0) \\ \Delta Y \end{bmatrix} \end{aligned}$$

where $\Delta Y = [\Delta y_0 \dots \Delta y_{N_t}]^T$. To summarize the gramian method we first note that the dimension of $[\Delta a, \Delta b, \Delta \alpha, \Delta \beta]$ is $4K+2$. The coefficients are perturbed by

$$[a, b, \alpha, \beta]^T \pm [0 \dots 0 \ \rho \ 0 \dots 0]^T$$

where ρ is at index i and $1 < i < 2K+1$ for perturbations associated with a, b and $2K+2 < i < 4K+2$ associated with perturbations to α, β . Define

$$\begin{bmatrix} \Delta h^i(0) \\ \Delta v^i(0) \\ \Delta Y^i \end{bmatrix} = \frac{1}{2\rho} \begin{bmatrix} h^{+i}(0) - h^{-i}(0) \\ v^{+i}(0) - v^{-i}(0) \\ \Delta Y^{+i} - \Delta Y^{-i} \end{bmatrix} \quad (26)$$

and let

$$\Delta = \begin{bmatrix} \Delta h^1(0) \dots \Delta h^{4k+2}(0) \\ \Delta v^1(0) \dots \Delta v^{4k+2}(0) \\ \Delta Y^1 \dots \Delta Y^{4k+2} \end{bmatrix}.$$

Then the gramian matrix is

$$G = \Delta^T \begin{bmatrix} P_b^{-1} & 0 \\ \frac{1}{2(N+1)} & 0 \\ 0 & \frac{R^{-1}}{N_t+1} \end{bmatrix} \Delta$$

and $(\epsilon/\rho)^2 \approx \sigma_{\min}$ where σ_{\min} is the smallest eigenvalue of G . We may either approximate the variation in the output Δy_i by (21) using the nonlinear model or approximate the variation by propagating the perturbations via the TLM as in (24).

References

- Baker NL, Daley R (2000) Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Q J R Meteorol Soc* 126:1431–1454
- Bauer P, Buizza R, Cardinali C, Thépaut J (2011) Impact of singular-vector-based satellite data thinning on NWP. *Q J R Meteorol Soc* 137:286–302
- Kang W (2011) The Consistency of partial observability for PDEs. [arXiv:1111.5846](https://arxiv.org/abs/1111.5846)
- Kang W, Krener AJ, Xiao M, Xu L (2013) A survey of Observers for Nonlinear Dynamical Systems. In: data assimilation for Atmospheric, oceanic and hydrologic applications, Vol II. Springer, Heidelberg
- Kang W, Xu L (2009a) A Quantitative measure of observability and controllability. In: Proceedings of IEEE conference on decision and control, Shanghai, China
- Kang W, Xu L (2009b) Computational analysis of control systems using dynamic optimization. [arXiv:0906.0215v2](https://arxiv.org/abs/0906.0215v2)
- Kang W, Xu L (2012) Optimal sensor placement for data assimilations. *Tellus* 64A:17133
- Kang W, Xu L (2014) Partial observability for Some distributed parameter systems. *Int J Dyn Control* 2(4):587–596. doi:[10.1007/s40435-014-0087-4](https://doi.org/10.1007/s40435-014-0087-4)
- King S, Kang W, Xu L (2013) Partial observability for the shallow water equations. In: Proceedings of SIAM conference on control & its applications, San Diego, CA
- King S, Kang W, Xu L (2014) Observability for optimal sensor locations in data assimilation. *Int J Dyn Control*. doi:[10.1007/s40435-014-0120-7](https://doi.org/10.1007/s40435-014-0120-7)
- Krener AJ, Ide K (2009) Measures of unobservability. Proceedings of IEEE conference on decision and control, Shanghai, China, pp 6401–6406
- Langland R, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus* 56A:189–201
- Lewis AS, Overton ML (2012) Nonsmooth optimization via quasi-Newton methods. *Math Program* 141:135–163
- Majumdar SJ, Bishop CH, Etterton BJ, Szunyogh I, Toth Z (2001) Can an ensemble transform Kalman filter predict the reduction in forecast-error variance produced by targeted observations? *Q J R Meteorol Soc* 127:2803–2820
- Ochotta T, Gebhardt C, Saupe D, Wergen W (2005) Adaptive thinning of atmospheric observations in data assimilation and vector quantization and filtering methods. *Q J R Meteorol Soc* 131:3427–3437

Quantification of Forecast Uncertainty and Data Assimilation Using Wiener's Polynomial Chaos Expansion

Junjun Hu, S. Lakshmivarahan and John M. Lewis

Abstract In this chapter we demonstrate the power of the Wiener's polynomial chaos based approach to quantify the uncertainty in the forecast of a dynamical system when the randomness enters through initial conditions, and/or parameters and/or forcing. This method enables an easy generation of forecast ensemble which can then be combined with one of many known methods for ensemble Kalman filtering.

Keywords Wiener chaos • Hermite polynomial • Data assimilation • Uncertainty quantification

1 Introduction

Uncertainty in the forecast based on a dynamical model can arise from the randomness in (1) the initial conditions and/or (2) the forcing term (including both the external forcing and the boundary conditions), and/or (3) randomness in the parameters of the model. In each of these cases, the solution of the model is a stochastic process. Our aim in this tutorial is twofold. First is to demonstrate the power of the Wiener's (1938) polynomial chaos¹ (PC) based expansion of a stochastic process. This is done by expressing the (unknown) solution of the model

¹The word chaos, as used here, refers to the randomness of the underlying stochastic process. Conceptually, it is different from the “deterministic chaos” that was discovered by Ed Lorenz in 1963.

J. Hu · S. Lakshmivarahan (✉)

The School of Computer Science, University of Oklahoma, Norman, OK 73019, USA
e-mail: varahan@ou.edu

J.M. Lewis

Desert Research Institute (DRI), 2215 Raggio Parkway, Reno, NV 89512, USA

J.M. Lewis

National Severe Storm Laboratory (NSSL), Norman, OK 73019, USA

in an orthogonal expansion using a stochastic basis consisting of the set of all Hermite polynomials of the standard Gaussian random variable whose distribution is defined over the real line, where the coefficients (or the strength of the modes) of the expansion are (unknown) deterministic functions of time. By exploiting the orthogonality property of the Hermite polynomial (with respect to the standard Gaussian as the weight function), the given model is reduced to a system of coupled nonlinear dynamics on the deterministic coefficient functions. By solving this reduced spectral dynamics numerically, we can then effectively reconstruct the stochastic solution of the original forecast model, based on which we can provide probabilistic characterization of the model forecast. An overview of the properties of both deterministic and stochastic versions of the Hermite polynomials are given in Appendices A and B.

While this approach is quite similar in principle to the well-known Karhunen–Loëve (K-L) expansion (Loëve 1977), there is a major difference in the choice of the stochastic basis. In K-L expansion, the stochastic basis consists of the eigen functions of the known correlation function of the underlying stochastic process. In our case, since we do not know the (stochastic) solution of the forecast model, let alone its underlying correlation structure, we rely on the more general approach based on Wiener's PC based expansion. However, like everything else in life, there is a price to pay for this lack of knowledge about the solution, namely, the solution based on K-L expansion is inherently optimal but the solution based on the Wiener's PC does not share this inherent optimality property (Loëve 1977; Ghanem and Spanos 1991). Our second goal is to use this PC expansion in an ensemble framework to perform data assimilation.

A succinct account of the role of Wiener's PC based approach in stochastic analysis is given in Kallianpur (1980) and Kuo (2006). Lototsky and Rozovskii (2006) develop a general framework for solving stochastic differential equations (Arnold 1974) using PC approach. Solution to the nonlinear filter (which is a general form of dynamic data assimilation for stochastic models) based on PC is developed in Lototsky (2011). Mathematical generalization of Wiener's PC to include Askey scheme (called gPC) is developed in Xiu and Karniadakis (2002a). The monograph by Xiu (2010) contains an elegant presentation of PC, gPC and their applications.

Earliest application of Wiener's PC based approach to quantify uncertainty in engineering problems—applied mechanics and structural engineering is due to Ghanem and Spanos (1991). Since then there is a virtual explosion of literature in this area. The review paper by Ghanem (1999) provides a very good presentation of the PC methodology and a roadmap for applications. Two recent books by Le Maître and Knio (2010) and Grigoriu (2012) provide excellent presentation of both the theory of PC and its multi-faceted applications.

A note on the other methods for quantifying the forecast uncertainty is in order. If the forecast uncertainty is only due to those in the initial condition, then the well-known partial differential equation known as the Liouville's equation (Saaty 1967) provides the complete solution by describing the evolution of the probability density function of the forecast with time. If the uncertainty in the forecast arises

from two sources—those in the initial condition and in forcing, then the evolution of the probability density of the forecast is given by the celebrated Kolmogorov's forward equation (Jazwinski 1970). Soong (1973) describes several special methods to handle the uncertainty in the parameters in an otherwise deterministic model. But, when the uncertainty arises from all the three sources—initial condition, forcing and parameters, as is considered in this chapter, to our knowledge, the Wiener's polynomial chaos and its generalization are the only known approaches to quantify the model forecast uncertainty.

The related theory of nonlinear and non-Gaussian dynamic data assimilation is embodied in the contemporary theory of nonlinear filtering that deals with combining an uncertain nonlinear model forecast with noisy (nonlinear) observations in a Bayesian framework (Crisan and Rozovskii 2011). In this case, the evolution of the posterior density that describes the evolution of the uncertainty in the analysis is given by the well-known Kushner-Zakai equation, which is a stochastic partial differential equation (Kushner 1962; Zakai 1969). We hasten to add that there is a natural nesting between the three well-known classes of partial differential equations mentioned above in the sense Kushner-Zakai becomes Kolmogorov's forward equation when there is no noisy observation and the latter in turn becomes Liouville's equation when there is no random forcing. Notice that this well-known hierarchy does not handle uncertainty in parameters.

In this chapter we examine the power of the Wiener's PC approach to assimilate noisy data into linear and nonlinear stochastic models. This approach is patterned after the ensemble Kalman filtering approach championed by Evensen (2007) and surveyed in Lakshmivarahan and Stensrud (2009). Li and Xiu (2009) were the first to apply the PC based approach to perform ensemble kalman fitering. Kalman's original paper (Kalman 1960) still continues to be a great inspiration for generations of researchers in this area. Lewis et al. (2006) provides a comprehensive summary of various approaches to dynamic data assimilation.

In Sect. 2 we provide a short summary of basic algorithmic framework of PC based analysis to the forecast and data assimilation problem. In Sect. 3, we consider the forecast analysis. In Sect. 4, PC based data assimilation approach is investigated in linear and nonlinear problems. Section 5 contains conducting remarks.

2 PC Framework for Forecast Analysis

Let $\mathbf{x}(k) \in R^n$ denote the state of a forecast model at time $k \geq 0$ defined by a nonlinear stochastic difference equation

$$\mathbf{x}(k+1) = \mathbf{M}[(\mathbf{x}(k), \boldsymbol{\alpha})] + \mathbf{w}(k). \quad (1)$$

where $\mathbf{M}: R^n \times R^p \rightarrow R^n$ is the one step state transition map or simply model map, $\mathbf{w}(k) \sim N(0, \mathbf{Q})$ is the white Gaussian noise representing the model error with known covariance \mathbf{Q} (mean $0 \in R^n$ is a vector of all zeros), $\boldsymbol{\alpha} \in R^p$ is a random

parameter vector drawn from a known Gaussian distribution, that is, $\alpha \sim N(\bar{\alpha}, \Sigma)$ where $\bar{\alpha}$ is the mean and Σ is its covariance, and $\mathbf{x}(0) \sim N(\mathbf{x}_0^a, \mathbf{P}_0^a)$ is the random initial condition which is again Gaussian with the known mean $\mathbf{x}_0^a = \mathbf{x}^a(0)$ and covariance $\mathbf{P}_0^a = \mathbf{P}^a(0)$. It is further assumed that $\mathbf{x}(0)$, α and $\mathbf{w}(k)$ are stochastically independent.

Remark 2.1 The key idea of the Wiener's PC approach is centered on the orthogonality of Hermite polynomials with respect to the standard Gaussian density as the weighting function. To be consistent with this philosophy and for added simplicity in the analysis that follows, it is tacitly assumed that $\mathbf{x}(0)$, α and $\mathbf{w}(k)$ are all Gaussian. If it turns out these are not Gaussian, we could then use the generalized polynomial chaos (gPC) expansion to deal with these cases. The mathematics is however very similar. Refer to the monograph by Xiu (2010) for an elegant treatment of gPC and their applications.

Let $\mathbf{x}(k) = \mathbf{x}(k, \xi)$ be the discrete time stochastic process defined by the (unknown) solution of (1). The basic premise of PC approach is to approximate the evolution of $\mathbf{x}(k)$ by an orthogonal expansion using Hermite polynomials as

$$\mathbf{x}(k) = \sum_{i=0}^N \mathbf{v}_i(k) \phi_i(\xi) \quad (2)$$

where $\mathbf{v}_i(k) \in R^n$ are the unknown deterministic coefficient vectors to be determined and $\phi_i(\xi)$ is the known i th degree Hermite polynomial in the standard Gaussian random variable $\xi \sim N(0, 1)$. Refer to Appendices A and B for a review of the properties of Hermite polynomials.

It follows from Appendices A and B that, $\xi = \phi_1(\xi)$ and $\{\phi_i(\xi)\}$ constitute an orthogonal basis based on the inner product

$$\langle \phi_i, \phi_j \rangle = \int_R \phi_i(\xi) \phi_j(\xi) p(\xi) d\xi = (i!) \delta_{ij}. \quad (3)$$

where

$$p(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{\xi^2}{2}\right).$$

and $\delta_{ij} = 1$ if $i = j$ and 0 otherwise. To simplify the notation, we now define an extension of the definition of the inner product in (3). Let $\mathbf{f} = (f_1, f_2, \dots, f_n)^T \in R^n$, then define a new vector $\bar{\mathbf{f}}_j \in R^n$, where

$$\bar{\mathbf{f}}_j = \langle \mathbf{f}, \phi_j \rangle = (\bar{f}_{1j}, \bar{f}_{2j}, \dots, \bar{f}_{nj})^T. \quad (4)$$

with $\bar{f}_{ij} = \langle f_i, \phi_j \rangle$.

As the first step in deriving the dynamics of evolution of the unknown coefficient vectors $\mathbf{v}_i(k)$, $0 \leq i \leq N$ and $k \geq 0$ in (2), we start by expanding the initial condition $\mathbf{x}(0)$, parameter $\boldsymbol{\alpha}$ and the model forcing $\mathbf{w}(k)$ using Hermite polynomials.

PC expansion of $\mathbf{x}(0)$: Let

$$\mathbf{x}(0, \xi) = \sum_{j=0}^N \mathbf{I}_j \phi_j(\xi). \quad (5)$$

be the PC expansion of the initial condition $\mathbf{x}(0)$, where $\mathbf{I}_j \in R^n$ are the unknown constant coefficient vectors to be determined. Taking inner product of both sides of (5) with $\phi_i(\xi)$, it follows from (4) that for $0 \leq i \leq N$,

$$\mathbf{I}_i = \frac{\langle \mathbf{x}(0, \xi), \phi_i(\xi) \rangle}{i!}. \quad (6)$$

where $0! = 1$ by definition.

But recall that $\mathbf{x}(0, \xi) \sim N(\mathbf{x}^a(0), \mathbf{P}^a(0))$. Let $\mathbf{P}^a(0) = \mathbf{A}\mathbf{A}^T$ be the Cholesky factorization of $\mathbf{P}^a(0)$, where $\mathbf{A} = [a_{ij}]$ is a lower triangular matrix. Then it can be verified that

$$\mathbf{x}(0, \xi) = \mathbf{x}^a(0) + \mathbf{A}1\phi_1(\xi). \quad (7)$$

where $1 \in R^n$ is a vector of all ones and we have used the fact that $\xi = \phi_1(\xi)$ (Refer to Appendices A and B). Substituting (7) in (6) and simplifying, it follows that the r^{th} component of the numerator on the right hand side of (6) is given by

$$\begin{aligned} \langle x_r(0, \xi), \phi_i(\xi) \rangle &= \langle x_r^a(0), \phi_i(\xi) \rangle + \langle (\sum_{s=1}^r a_{rs})\phi_1(\xi), \phi_i(\xi) \rangle \\ &= x_r^a(0)\delta_{0i} + (\sum_{s=1}^r a_{rs})\delta_{1i}. \end{aligned} \quad (8)$$

Substituting (8) in (6), it follows that, for $0 \leq i \leq N$

$$\mathbf{I}_i = \mathbf{x}^a(0)\delta_{0i} + (\mathbf{A}1)\delta_{1i}. \quad (9)$$

That is, the new initial condition for the amplitudes is given by

$$\mathbf{I}_0 = \mathbf{x}^a(0), \text{ the mean of } \mathbf{x}(0, \xi)$$

$$\mathbf{I}_1 = \mathbf{A}1, \text{ sum of the columns of } \mathbf{A}$$

and

$$\mathbf{I}_j \equiv 0 \text{ for } 2 \leq j \leq N. \quad (10)$$

Hence, the PC expansion for the initial condition is given by

$$\mathbf{x}(0, \xi) = \mathbf{x}^a(0) + (\mathbf{A}1)\phi_1(\xi). \quad (11)$$

Remark 2.2 If the initial condition is deterministically specified, then $\mathbf{P}^a(0)$ is a zero matrix and so is \mathbf{A} . In this case, $\mathbf{I}_0 = \mathbf{x}^a(0)$ and $\mathbf{I}_j \equiv 0$ for $1 \leq j \leq N$.

PC expansion of the parameter α : Let

$$\alpha(\xi) = \sum_{j=0}^N \alpha_j \phi_j(\xi). \quad (12)$$

be the PC expansion of the parameters where $\alpha_j \in R^p$ are the unknown constant coefficients and $\alpha(\xi) \sim N(\bar{\alpha}, \Sigma)$. Let $\Sigma = \mathbf{C}\mathbf{C}^T$ be the Cholesky decomposition of Σ with $\mathbf{C} = [c_{ij}]$ being a lower triangular matrix. Then, by repeating the procedure described above, it can be easily verified that

$$\alpha_0 = \bar{\alpha}, \text{ the mean of } \alpha$$

$$\alpha_1 = \mathbf{C}1, \text{ sum of the columns of } \mathbf{C}$$

$$\alpha_j = 0, \text{ for } 2 \leq j \leq N. \quad (13)$$

Hence, the PC expansion for the parameter α is given by

$$\alpha(\xi) = \bar{\alpha} + (\mathbf{C}1)\phi_1(\xi). \quad (14)$$

Remark 2.3 When α is nonrandom, then $\alpha_0 = \bar{\alpha}$ and $\alpha_j = 0$, for $1 \leq j \leq N$.

PC expansion of the forcing term $\mathbf{w}(k)$:

Since $\mathbf{w}(k)$ is a stationary Gaussian white noise with a common distribution $N(0, \mathbf{Q})$, let

$$\mathbf{w}(k, \xi) = \sum_{j=0}^N \mathbf{F}_j \phi_j(\xi). \quad (15)$$

where $\mathbf{F}_j \in R^n$ are the unknown, time invariant coefficient vectors to be determined. Let $\mathbf{Q} = \mathbf{B}\mathbf{B}^T$ be the Cholesky factorization of \mathbf{Q} with $\mathbf{B} = [b_{ij}]$ being a lower triangular matrix. By proceeding along similar line, it immediately follows that

$$\mathbf{F}_0 = 0, \mathbf{F}_1 = \mathbf{B}1, \text{ the sum of the columns of } \mathbf{B}$$

and

$$\mathbf{F}_j = 0, 2 \leq j \leq N. \quad (16)$$

Hence the PC expansion of $\mathbf{w}(k, \xi)$ is

$$\mathbf{w}(k, \xi) = (\mathbf{B}1)\phi_1(\xi). \quad (17)$$

We now turn to our main task.

PC expansion for the solution $\mathbf{x}(k, \xi)$:

Substituting (2), (14), and (17) in (1), the latter becomes

$$\sum_{j=0}^N \mathbf{v}_j(k+1)\phi_j(\xi) = \mathbf{M} \left[\sum_{j=0}^N \mathbf{v}_j(k)\phi_j(\xi), \bar{\alpha} + (\mathbf{C}1)\phi_1(\xi) \right] + (\mathbf{B}1)\phi_1(\xi). \quad (18)$$

Taking inner products of both sides with $\phi_i(\xi)$ and simplifying, in view of (3) it follows that

$$i! \mathbf{v}_i(k+1) = \langle \mathbf{M} \left[\sum_{j=0}^N \mathbf{v}_j(k)\phi_j(\xi), \bar{\alpha} + (\mathbf{C}1)\phi_1(\xi) \right], \phi_i(\xi) \rangle + \langle (\mathbf{B}1)\phi_1(\xi), \phi_i(\xi) \rangle \quad (19)$$

Define a new vector $\mathbf{V}(k) \in R^{n \times (N+1)}$ by concatenating $(N+1)$ vectors $\mathbf{v}_0(k), \mathbf{v}_1(k), \dots, \mathbf{v}_N(k)$ as

$$\mathbf{V}(k) = [\mathbf{v}_0^T(k), \mathbf{v}_1^T(k), \dots, \mathbf{v}_N^T(k)]^T.$$

Also, let β denotes a new parameter vector that contains the elements of the vector α and the lower triangular matrix C . Performing the inner product on the right side of (19), we obtain a nonlinear dependence of $\mathbf{v}_i(k+1)$ on $\mathbf{V}(k)$ and β , and a linear dependence on \mathbf{B} , denoted by

$$\mathbf{v}_i(k+1) = \bar{\mathbf{M}}_i(\mathbf{V}(k), \beta) + \frac{(\mathbf{B}1)\delta_{1i}}{i!}. \quad (20)$$

with

$$\mathbf{v}_0(0) = \mathbf{I}_0, \mathbf{v}_1(0) = \mathbf{I}_1, \text{ and } \mathbf{v}_i(0) = 0, 2 \leq i \leq N$$

given in (10) as the initial conditions. The actual functional form of $\bar{\mathbf{M}}_i$ depends on the model map \mathbf{M} .

By combining the $(N+1)$ equations, the system (20) can be succinctly represented as

$$\mathbf{V}(k+1) = \bar{\mathbf{M}}[\mathbf{V}(k), \beta] + \mathbf{g}(k). \quad (21)$$

where the forcing $\mathbf{g}(k)$ and the initial condition $\mathbf{V}(0)$ are given in the partitioned form as

$$\mathbf{g}(k) = \begin{bmatrix} 0 \\ \mathbf{B}1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in R^{n \times (N+1)} \text{ and } \mathbf{V}(0) = \begin{bmatrix} \mathbf{I}_0 = \mathbf{x}^a(0) \\ \mathbf{I}_1 = \mathbf{A}1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \in R^{n \times (N+1)}. \quad (22)$$

Solving the first-order, nonlinear deterministic recurrence relation (21) in $R^{n \times (N+1)}$, we can compute the evolution of the time varying coefficients $\mathbf{v}_i(k)$ for $0 \leq i \leq N$ and $k > 0$. Hence, we obtain an approximate representation of the actual solution $\mathbf{x}(k)$ of (1).

Remark 2.4 It can be shown that the accuracy of this approximation improves as N increases. Only for simplicity in this tutorial, we have used Hermite polynomials in a single Gaussian variable. One could readily develop a parallel derivation using multivariate Hermite polynomials. One can also readily undertake the analysis of the impact of the use of multivariate versus univariate Hermite polynomials as well as the number N of terms on the accuracy of the representation such as (2).

Using (2) we can indeed quantify the approximate expressions for the mean and covariance of $\mathbf{x}(k)$. Taking expectations on both sides of (2) and using the properties of the Hermite polynomials (Appendices A and B), it follows that

$$\bar{\mathbf{x}}(k) = E[\mathbf{x}(k)] = \mathbf{v}_0(k).$$

and

$$\bar{\mathbf{P}}(k) = \text{Cov}[\mathbf{x}(k)] = \sum_{i=1}^N i! [\mathbf{v}_i(k) \mathbf{v}_i^T(k)]. \quad (23)$$

By drawing samples of ξ_r ($1 \leq r \leq M$) from the standard normal distribution, we can indeed construct the r th member of the forecast ensembles of the solution (1) as

$$[\mathbf{x}(k)]_r = \sum_{i=0}^N \mathbf{v}_i(k) \phi_i(\xi_r). \quad (24)$$

Using this forecast ensemble, one can in principle build a histogram and analyze its evolution experimentally.

3 Examples

In this section we apply the theory developed in Sect. 2 to the analysis of the ensemble forecast from a stochastic dynamics.

3.1 Experiment 3.1 A Linear Problem

To establish the base, we first consider the deterministic version of linear model given by the scalar ($n = 1$) differential equation

$$\dot{x} = f(x) = ax. \quad (25)$$

The solution $x(t)$ of (25) is given by

$$x(t) = x(0)e^{at}. \quad (26)$$

where $x(0)$ is the initial condition.

The standard Ito type stochastic differential equation (Arnold 1974) corresponding to (25) is

$$dx(t) = f[x(t)]dt + \bar{q}dw(t). \quad (27)$$

where $dw(t) \sim N(0, \Delta t)$ is the (Gaussian distributed) Wiener independent increment process and \bar{q} controls the intensity of the forcing term in (27). Discretizing (27) using the standard Euler scheme, we get

$$x(k+1) = M[x(k), \alpha] + w(k+1). \quad (28)$$

where $\alpha (= a)$ is a real stochastic parameter $\alpha \sim N(\bar{\alpha}, \sigma^2)$ and $w(k) \sim N(0, \bar{q}^2 \Delta t)$ with

$$M[x(k), \alpha] = x(k) + \alpha \Delta t x(k). \quad (29)$$

It is further assumed that the initial condition for (28) $x(0) \sim N(x^a(0), P^a(0))$.

Following the development in Sect. 2, it follows that the initial condition $x(0)$, the parameter α and the forcing $w(k)$ admit the following PC expansion:

$$x(0, \xi) = x^a(0) + \sqrt{P^a(0)} \phi_1(\xi). \quad (30)$$

$$\alpha = \bar{\alpha} + \sigma \phi_1(\xi). \quad (31)$$

and

$$w(k, \xi) = \bar{q} \sqrt{\Delta t} \phi_1(\xi). \quad (32)$$

For definiteness, we seek a PC expansion of the solution $x(k)$ of (28) in the form

$$x(k) = \sum_{j=0}^N v_j(k) \phi_j(\xi). \quad (33)$$

where the spectral amplitudes $v_j(k) \in \mathbb{R}$ for $0 \leq j \leq N$ and $k \geq 0$. Substituting (31) –(33) in (28), we get

$$\sum_{j=0}^N v_j(k+1) \phi_j(\xi) = M \left[\sum_{j=0}^N v_j(k) \phi_j(\xi), \bar{\alpha} + \sigma \phi_1(\xi) \right] + \bar{q} \sqrt{\Delta t} \phi_1(\xi). \quad (34)$$

From (29),

$$M \left[\sum_{j=0}^N v_j(k) \phi_j(\xi), \bar{\alpha} + \sigma \phi_1(\xi) \right] = (1 + \bar{\alpha} \Delta t + \sigma \phi_1(\xi) \Delta t) \sum_{j=0}^N v_j(k) \phi_j(\xi). \quad (35)$$

Computing the inner product of both sides of (34) with $\phi_i(\xi)$ for $0 \leq i \leq N$, while is quite straightforward, involves tedious work. To avoid distraction from the main flow of things, we provide the details of these computations in Appendix C.

Let $\mathbf{V}(k) = [v_0(k), v_1(k), \dots, v_{N-1}(k), v_N(k)]^T \in \mathbb{R}^{N+1}$ be the vector of spectral coefficients in the PC expansion of $x(k)$ in (33), β denotes a new parameter vector that contains $\bar{\alpha}$ and σ . From above, it follows that the evolution of $\mathbf{V}(k)$ is given by the following deterministic vector difference equation

$$\mathbf{V}(k+1) = \bar{\mathbf{M}}[\mathbf{V}(k), \beta] + \mathbf{g}(k). \quad (36)$$

where the forcing term $\mathbf{g}(k) = (0, \bar{q} \sqrt{\Delta t}, 0 \dots 0)^T$ and $\mathbf{V}(0) = [x^a(0), \sqrt{P^a(0)}, 0 \dots 0]^T$.

Solving (36) numerically we obtain the values of $\mathbf{V}(k)$ for $k \geq 0$, using which we get an expression for the stochastic solution (33). The PC based approximate mean and variance of $x(k)$ are given by

$$\mathbb{E}[x(k)] = v_0(k). \quad (37)$$

and

$$\text{Var}[x(k)] = \sum_{j=1}^N j! v_j^2(k). \quad (38)$$

We can create an ensemble of $x(k)$ by creating an ensemble of ξ . That is, the i th ensemble $x_i(k)$ induced by i th ensemble ξ_i is given by

$$x_i(k) = \sum_{j=0}^N v_j(k) \phi_j(\xi_i). \quad (39)$$

This ensemble reflects the randomness in the solution induced by that of the initial condition, the parameter and the forcing.

One thing that needs to be clarified is that when the parameter α is deterministic, then the system operator M in Eq. (28) is not only linear but deterministic. In this case, the solutions will be Gaussian distributed all the time. There is no need to use the PC expansion method (since the expansion reduces to its first two terms). In real

applications such as meteorology, the true linear problem considers random initial conditions, random parameters and random forcing.

In the experiment, we consider random initial conditions, random parameter and random forcing for the linear problem given by

$$x(0) \sim N(0.8, 1), \alpha \sim N(0.1, 0.1), w(k) \sim N(0, 0.5). \quad (40)$$

The results are compared with those obtained from the Monte Carlo (MC) method with 10,000 samples. Figure 1 shows the ensemble mean and variance evolution over the time for $N=4$.

A comparison of the evolution of the histograms of the PC based solution in (39) and a direct MC method using (28) and the distributions in (40) at different times are given in Fig. 2 through 5. As can be seen from these figures, the PC method and MC method produce very similar results (Figs. 2, 3, 4 and 5).

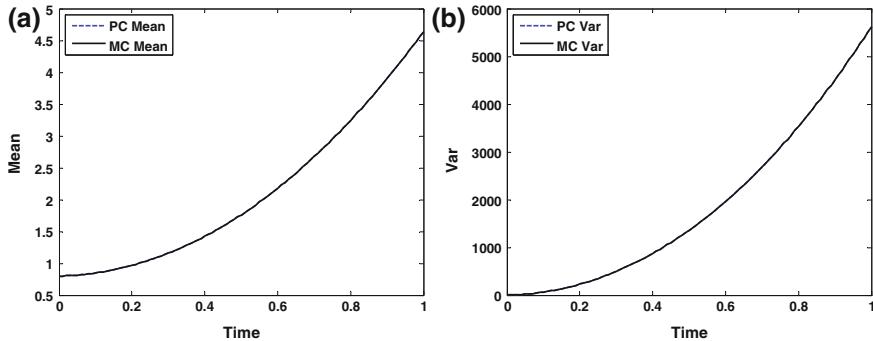


Fig. 1 The evolution of the ensemble mean and variance obtained from Polynomial Chaos (PC) expansion method and Monte Carlo (MC) method for linear problem: **a** mean **b** variance

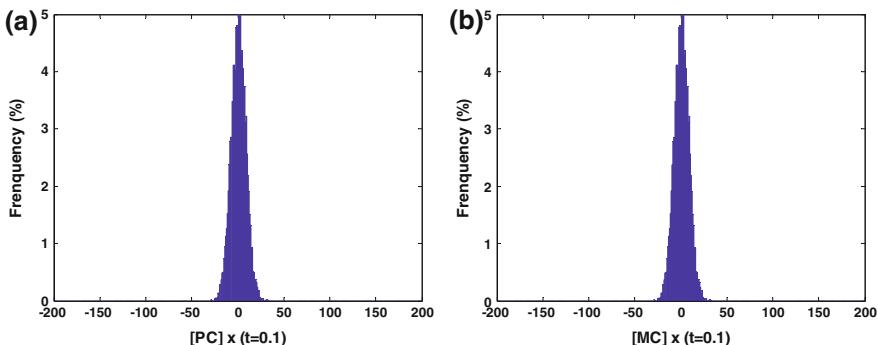


Fig. 2 Histogram of the ensembles for linear problem at time $t = 0.1$. **a** PC, **b** MC

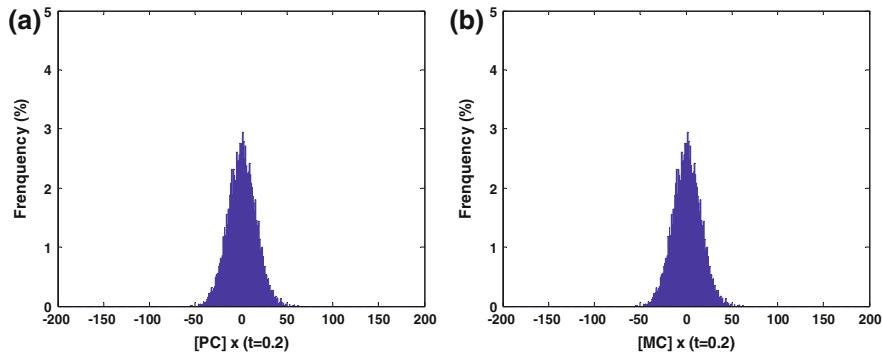


Fig. 3 Histogram of the ensembles for linear problem at time $t = 0.2$. **a** PC, **b** MC

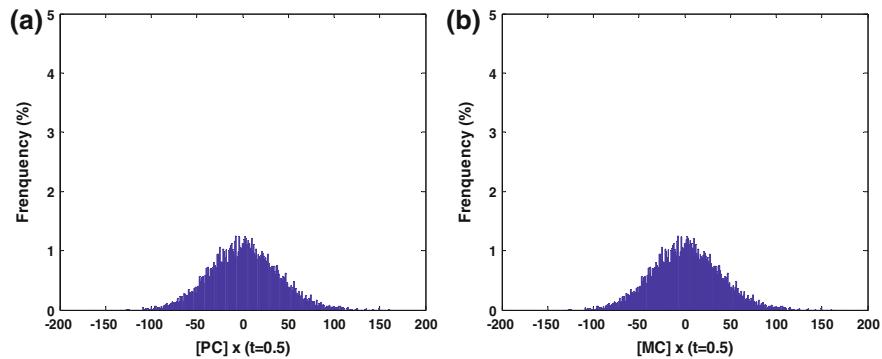


Fig. 4 Histogram of the ensembles for linear problem at time $t = 0.5$. **a** PC, **b** MC

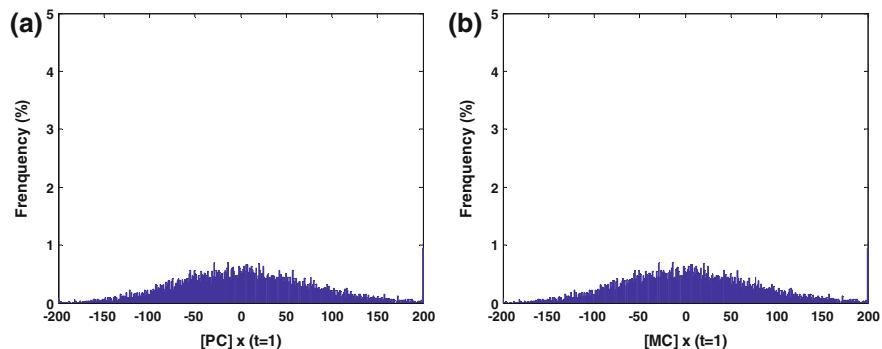


Fig. 5 Histogram of the ensembles for linear problem at time $t = 1$. **a** PC, **b** MC

3.2 Experiment 3.2 Nonlinear Problem

Next, we consider the nonlinear population equation given by Li and Xiu (2009),

$$\frac{dx}{dt} = -r \left(1 - \frac{x}{A}\right)x. \quad (41)$$

The standard Ito type stochastic differential equation corresponding to (41) is represented as

$$dx(t) = -r \left(1 - \frac{x}{A}\right)x(t)dt + \bar{q}dw(t). \quad (42)$$

where $dw(t) \sim N(0, \Delta t)$ is the (Gaussian distributed) Wiener independent increment process and \bar{q} controls the intensity of the forcing term in (42). Discretizing (42) using the standard Euler scheme, we get

$$x(k+1) = M[x(k), \alpha, \beta] + w(k+1). \quad (43)$$

where $\alpha = r$ and $\beta = A$ are stochastic parameters with $\alpha \sim N(\bar{\alpha}, \sigma_1^2)$, $\beta \sim N(\beta, \sigma_2^2)$ and $w(k) \sim N(0, \bar{q}^2 \Delta t)$ with

$$M[x(k), \alpha, \beta] = x(k) - \alpha \Delta t \left(1 - \frac{x(k)}{\beta}\right)x(k). \quad (44)$$

It is further assumed that the initial condition for (41) $x(0) \sim N(x^a(0), P^a(0))$.

Following the development in Sect. 2, it follows that the initial condition $x(0)$, the parameter α, β and the forcing $w(k)$ admit the following PC expansion:

$$x(0, \xi) = x^a(0) + \sqrt{P^a(0)}\phi_1(\xi). \quad (45)$$

$$\alpha = \bar{\alpha} + \sigma_1 \phi_1(\xi). \quad (46)$$

$$\beta = \beta + \sigma_2 \phi_1(\xi). \quad (47)$$

and

$$w(k, \xi) = \bar{q} \sqrt{\Delta t} \phi_1(\xi). \quad (48)$$

For definiteness, we seek a PC expansion of the solution $x(k)$ of (43) in the form

$$x(k) = \sum_{j=0}^N v_j(k) \phi_j(\xi). \quad (49)$$

where the spectral amplitudes $v_j(k) \in R$ for $0 \leq j \leq N$ and $k \geq 0$. Substituting (49) in (43), we get

$$\sum_{j=0}^N v_j(k+1) \phi_j(\xi) = M \left[\sum_{j=0}^N v_j(k) \phi_j(\xi), \alpha, \beta \right] + \bar{q} \sqrt{\Delta t} \phi_1(\xi). \quad (50)$$

From (44) and (46)–(48),

$$\begin{aligned} M \left[\sum_{j=0}^N v_j(k) \phi_j(\xi), \alpha, \beta \right] \\ = (1 - \bar{\alpha} \Delta t - \sigma_1 \phi_1(\xi) \Delta t) \sum_{j=0}^N v_j(k) \phi_j(\xi) + \frac{[\bar{\alpha} + \sigma_1 \phi_1(\xi)] \Delta t}{\beta + \sigma_2 \phi_1(\xi)} \left[\sum_{j=0}^N v_j(k) \phi_j(\xi) \right]^2. \end{aligned} \quad (51)$$

The details of computing the inner product of both sides of (51) with $\phi_i(\xi)$ are given in Appendix D, where we consider A as a deterministic parameter, that is, $\sigma_2 = 0$ in (51).

As in the linear problem, let $\mathbf{V}(k) = [v_0(k), v_1(k), \dots, v_{N-1}(k), v_N(k)]^T \in R^{N+1}$ be the vector of spectral coefficients in the PC expansion of $x(k)$ in (49). The evolution of $\mathbf{V}(k)$ will be given by a deterministic vector difference equation with formula (36). The values of $\mathbf{V}(k)$ for $k \geq 0$ are obtained by solving (36) numerically. Afterwards, the stochastic solution can be constructed by using (49). Same as linear problem, the PC based mean and variance estimates of $x(k)$ are given by (37) and (38). And the creation of an ensemble of $x(k)$ can be done similarly as that for linear problem.

In this experiment, we consider the stochastic dynamic model with random initial condition, random parameters r , deterministic parameter A , but with forcing. The values are

$$x(0) \sim N(0.1, 0.005), r \sim N(1, 0.04), A = 2.0, \bar{q} = 0.02. \quad (52)$$

Again, the experiments results are compared to those obtained from Monte Carlo method with 10,000 samples. The evolution of the ensemble mean and variance when $N = 4$ are given in Fig. 6. As can be seen, the PC method has similar performance with MC method, but the difference between them is not as small as that for linear problem.

Same as for linear problem, the histograms for the ensembles obtained at different times are given in Figs. 7, 8, 9, and 10.

More terms ($N = 8$) are added to solve the nonlinear problem, but the results show us that there is little improvement when we increase the term number from 5 to 9, that is from $N = 4$ to $N = 8$. The experiment results for $N = 8$ are given in Figs. 11, 12, 13, 14, and 15.

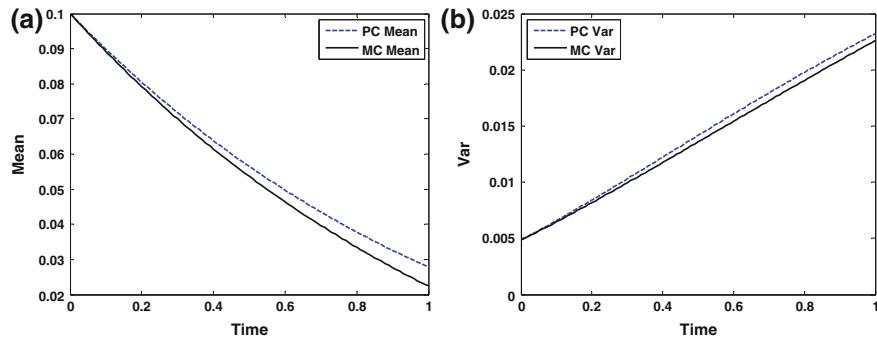


Fig. 6 The evolution of the ensemble mean and variance obtained from PC method ($N = 4$) and MC method for nonlinear problem: **a** mean, **b** variance

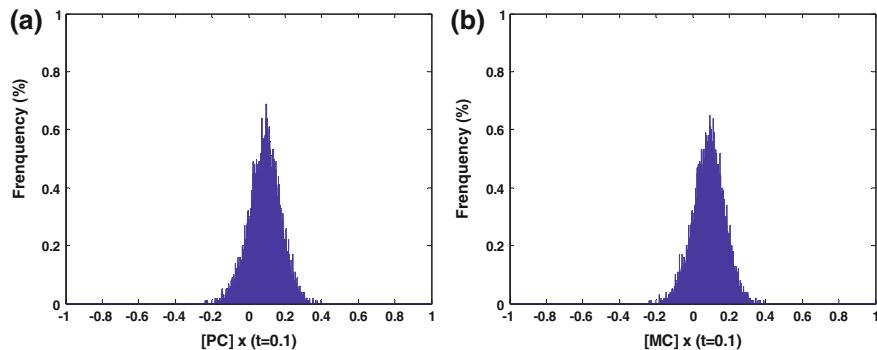


Fig. 7 Histogram of the ensembles for nonlinear problem at time $t = 0.1$. **a** PC ($N = 4$) **b** MC

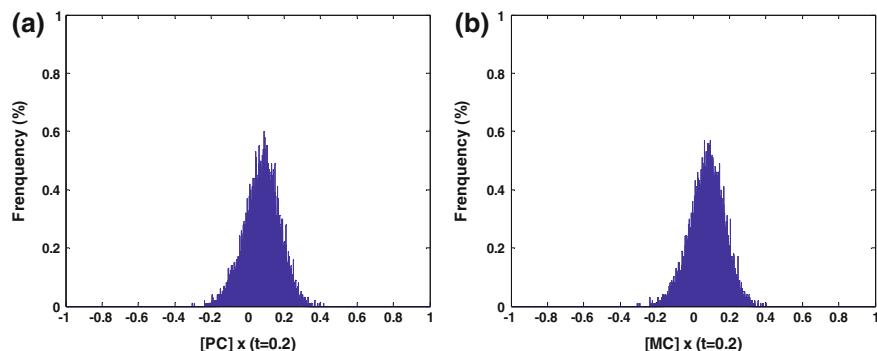


Fig. 8 Histogram of the ensembles for nonlinear problem at time $t = 0.2$. **a** PC ($N = 4$), **b** MC

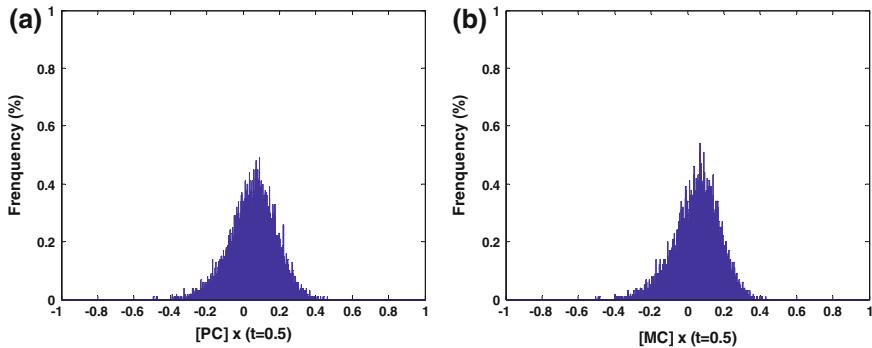


Fig. 9 Histogram of the ensembles for nonlinear problem at time $t = 0.5$. **a** PC ($N = 4$) **b** MC

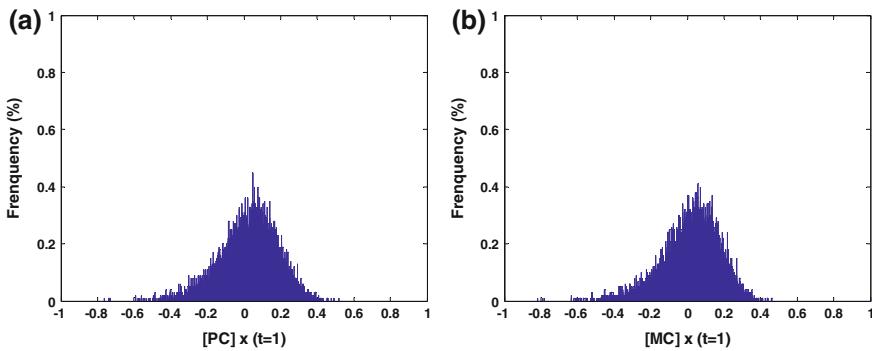


Fig. 10 Histogram of the ensembles for nonlinear problem at time $t = 1$. **a** PC ($N = 4$), **b** MC

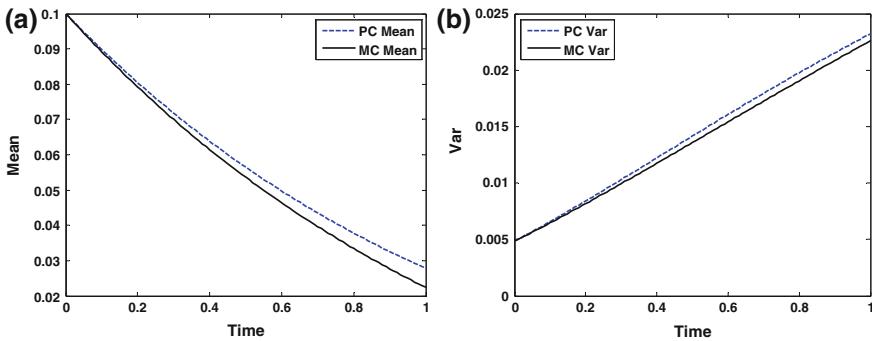


Fig. 11 The evolution of the ensemble mean and variance obtained from PC method ($N = 8$) and MC method for nonlinear problem: **a** mean, **b** variance

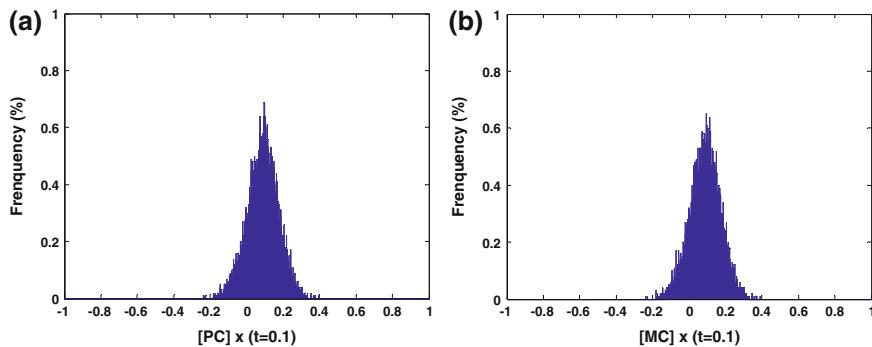


Fig. 12 Histogram of the ensembles for nonlinear problem at time $t = 0.1$. **a** PC ($N = 8$), **b** MC

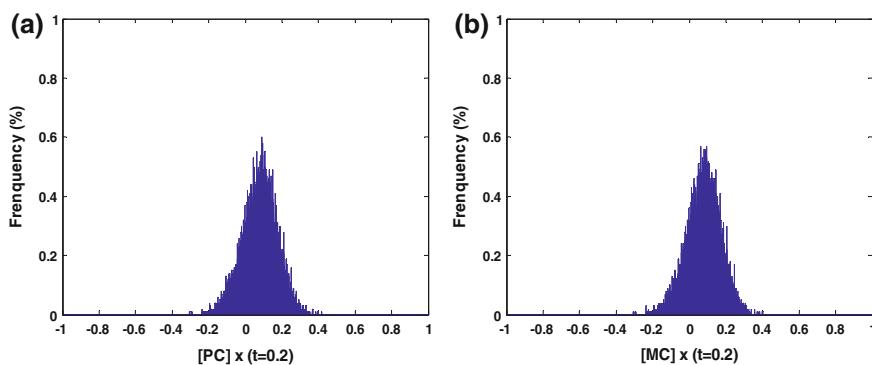


Fig. 13 Histogram of the ensembles for nonlinear problem at time $t = 0.2$. **a** PC ($N = 8$), **b** MC

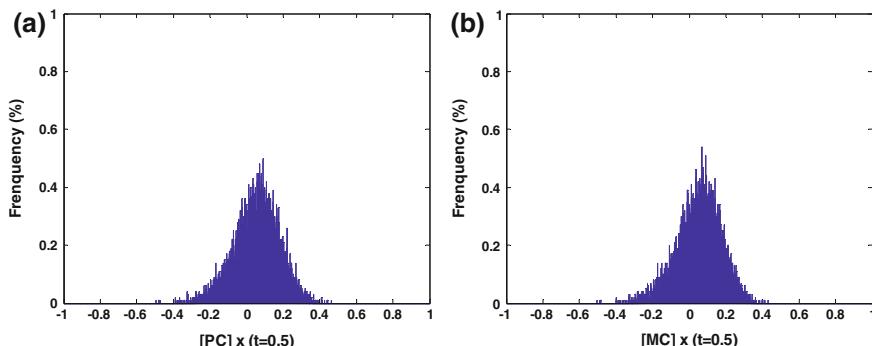


Fig. 14 Histogram of the ensembles for nonlinear problem at time $t = 0.5$. **a** PC ($N = 8$), **b** MC

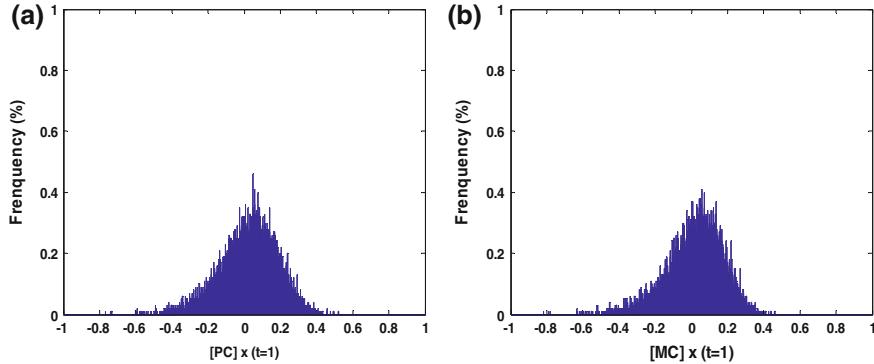


Fig. 15 Histogram of the ensembles for nonlinear problem at time $t = 1$. **a** PC ($N = 8$), **b** MC

4 Data Assimilation Using PC and Ensemble Method

By combining the forecast ensemble that is generated at time k in (24) and the observation $\mathbf{z}_k \in \mathbb{R}^m$ available at time k , we can now obtain analysis ensemble using one of the several known methods (Evensen 2007; Lewis et al. 2006; Lakshmi-varahan and Stensrud 2009).

To this end, first define, using (23)–(24), the forecast anomaly

$$[\mathbf{e}(k)]_r = [\mathbf{x}(k)]_r - \bar{\mathbf{x}}_k, 1 \leq r \leq M. \quad (53)$$

The goal is then to generate the analysis ensemble. This can be done in one of several ways: (a) stochastic method using the virtual observation and (b) deterministic methods that transform the forecast anomaly ensemble into analysis anomaly ensemble.

In the following we use the stochastic method which may be summarized as follows: The analysis mean and anomaly are given by

$$\bar{\mathbf{x}}_k^a = \bar{\mathbf{x}}_k + \mathbf{K}_k [\mathbf{z}_k - \mathbf{H}\bar{\mathbf{x}}_k]. \quad (54)$$

$$[\mathbf{e}_k^a]_r = [\mathbf{e}(k)]_r + \tilde{\mathbf{K}}_k \mathbf{H} [\mathbf{e}(k)]_r, 1 \leq r \leq M. \quad (55)$$

where \mathbf{H} is the observation operator and the gain matrices are given by

$$\mathbf{K}_k = \mathbf{P}_k \mathbf{H}^T [\mathbf{H} \mathbf{P}_k \mathbf{H}^T + \mathbf{R}]^{-1}. \quad (56)$$

$$\tilde{\mathbf{K}}_k = \mathbf{P}_k \mathbf{H}^T [\mathbf{H} \mathbf{P}_k \mathbf{H}^T]^{-1} \left[\sqrt{\mathbf{H} \mathbf{P}_k \mathbf{H}^T + \mathbf{R}} + \sqrt{\mathbf{R}} \right]^{-1}. \quad (57)$$

where $\tilde{\mathbf{K}}_k$ is the solution of

$$[\mathbf{I} - \tilde{\mathbf{K}}_k \mathbf{H}] \mathbf{P}_k [\mathbf{I} - \tilde{\mathbf{K}}_k \mathbf{H}] = [\mathbf{I} - \tilde{\mathbf{K}}_k \mathbf{H}] \mathbf{P}_k. \quad (58)$$

The analysis ensemble then is given by

$$[\mathbf{x}^a(k)]_r = \bar{\mathbf{x}}_k^a + [\mathbf{e}_k^a]_r, \quad 1 \leq r \leq M. \quad (59)$$

Let the PC expansion of the analysis \mathbf{x}_k^a at time k be given by

$$\mathbf{x}_k^a = \sum_{i=0}^N \mathbf{v}_i^a(k) \phi_i(\xi). \quad (60)$$

Then

$$\mathbf{v}_i^a(k) = \frac{1}{i!} \langle \mathbf{x}_k^a, \phi_i(\xi) \rangle \approx \frac{1}{i!M} \sum_{r=1}^M [\mathbf{x}_k^a]_r \phi_i(\xi_r). \quad (61)$$

Now using $\mathbf{v}_i^a(k), 0 \leq i \leq N$ as the initial condition for the deterministic dynamics (21), we can readily obtain the forecast $\mathbf{v}_i(k+1), 0 \leq i \leq N$ at time $(k+1)$, and the assimilation forecast cycle repeats.

Experiment 4.1 Linear problem In this section, we study the linear problem discussed in Sect. 3.1. In our example, Runge–Kutta method is firstly used to solve the deterministic Eq. (25) in which $a = 0.1$, and the initial values $x(0) = 0.8$, to obtain the deterministic solutions in the time interval $t \in [0, 1]$. The time increment is set to be $\Delta t = 0.01$. Afterwards, observations are created every $\Delta t = 0.05$ time unit by adding Gaussian measurement noise u_i following $N(0, R)$ to the deterministic solutions, that is

$$z_i = x_i + u_i. \quad (62)$$

where i is observation time.

In the data assimilation experiment, the stochastic linear model with parameter and uncertainties given in Eq. (40) is considered. The 4th-order PC approximation ($N = 4$) method is adopted to solve the stochastic model in the experiment. In order to study the behavior of the PC based data assimilation, we have selected three different values for the measurement noise variance R , which are 0.2^2 , 0.5^2 and 1 , along with four different values for the number of ensemble members M , which are 10 , 10^2 , 10^3 and 10^4 in the experiment. Figures 16 and 17 present the results when $R = 1$ and $M = 10^4$.

We have two sets of designs in the experiment for each combination of R and M . As shown in Fig. 16, the first design does not take into consideration of data assimilation. Open-Loop solutions at different times are obtained through the evolution of the stochastic model equation using PC based method. Figure 16 presents the Open-Loop ensemble mean at each observation time. The second

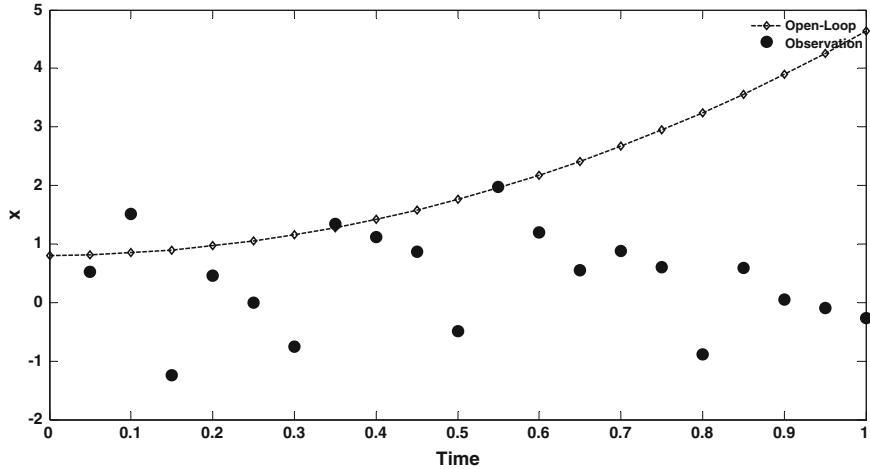


Fig. 16 Open-Loop solution for linear problem

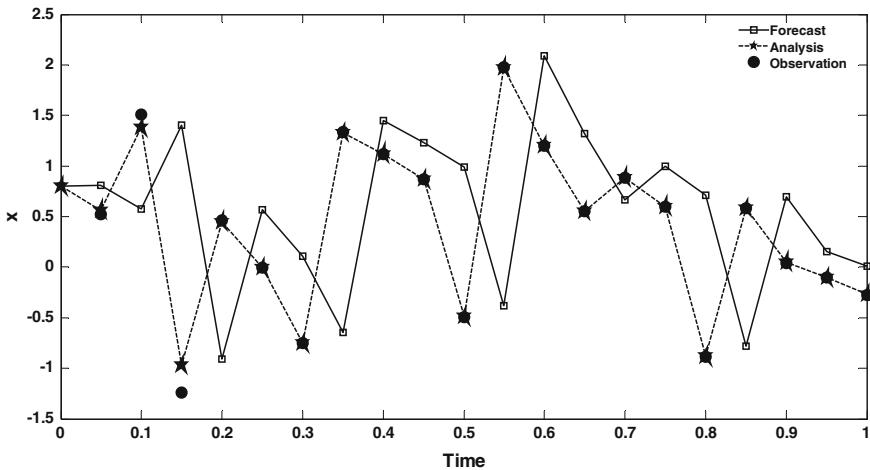


Fig. 17 Data assimilation for linear problem

design studies the combination of PC method and data assimilation scheme. As shown in Fig. 17, at each observation time, the forecast ensembles are obtained by using the analysis at previous time. The analysis ensembles are calculated by assimilating the observations into forecast ensembles. The forecast and analysis shown in Fig. 17 are the forecast ensemble mean and analysis ensemble mean. From Figs. 16 and 17, we can see that the analysis stays much closer to observation than Open-Loop solution.

In order to quantify the impact of assimilation, we compute the Root Mean Square Error (RMSE) as follows:

$$RMSE^o = \sqrt{\frac{\sum_{i=1}^q (op_i - obs_i)^2}{q}}. \quad (63)$$

$$RMSE^a = \sqrt{\frac{\sum_{i=1}^q (a_i - obs_i)^2}{q}}. \quad (64)$$

Here, op_i refers to the Open-Loop ensemble mean at each observation time i , obs_i is the observation value at time i , and a_i is the analysis ensemble mean at time i .

Values of $RMSE^o$ and $RMSE^a$ resulting from different selection of R and M are given in Table 1. The numbers in Table 1 further show us that data assimilation method can make analysis much closer to observation, even when we only use 10 ensemble members in the experiment.

Table 1 Difference to observations for linear problem

RM	10		10 ²		10 ³		10 ⁴	
	RMSE ^a	RMSE ^o						
0.2 ²	0.0113	3.7194	0.0040	0.5180	0.0004	0.2148	0.0009	0.2129
0.5 ²	0.0308	3.6845	0.0204	0.7500	0.0027	0.6595	0.0062	0.6398
1 ²	0.0571	4.2124	0.0498	1.2194	0.0197	0.8976	0.0681	0.9620

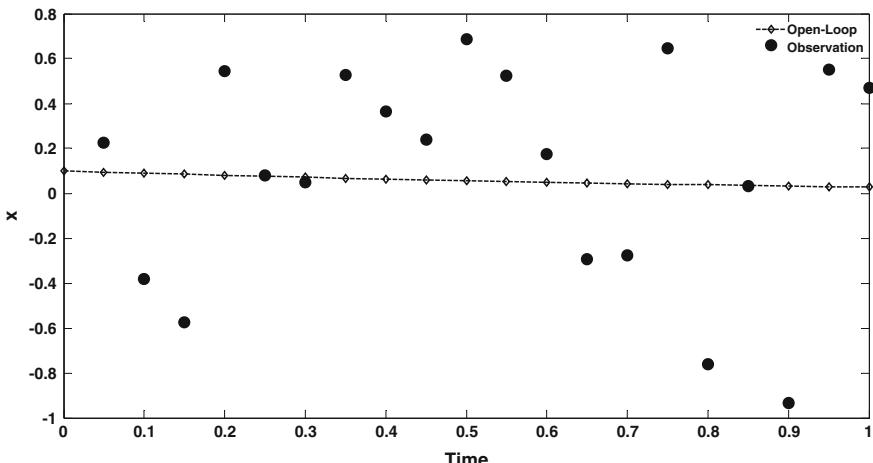


Fig. 18 Open-Loop solution for nonlinear problem

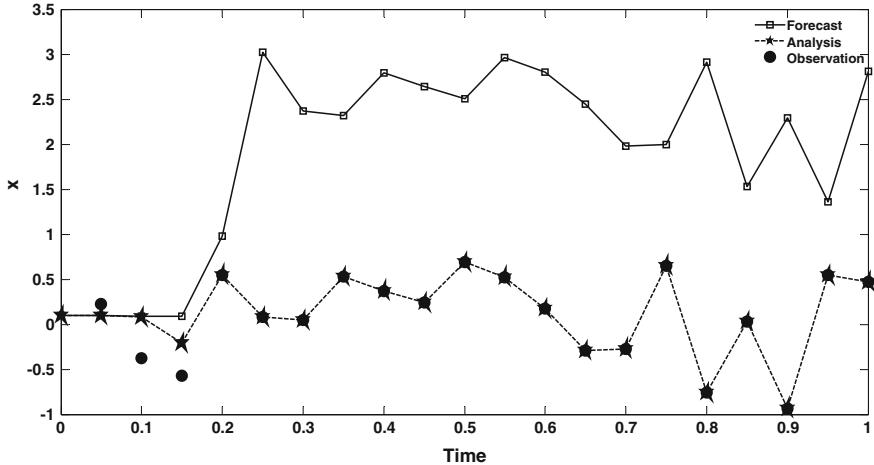


Fig. 19 Data assimilation for nonlinear problem

Table 2 Difference compared to observations for nonlinear problem

RM	10		10 ²		10 ³		10 ⁴	
	RMSE ^a	RMSE ^o						
0.06 ²	0.0064	0.0973	0.0190	0.0788	0.0061	0.0759	0.0192	0.0766
0.2 ²	0.0588	0.2097	0.0583	0.1929	0.0505	0.1899	0.0522	0.1907
0.5 ²	0.1049	0.4720	0.1921	0.4712	0.0977	0.4712	0.1352	0.4712

Experiment 4.2 Nonlinear problem The design of the nonlinear problem experiment is similar as that for the linear problem. The problem is concentrating on the nonlinear problem discussed in Sect. 3.2.

First, a series of deterministic solutions are generated by using Runge–Kutta method to solve the deterministic nonlinear population problem (41) in time interval $[0, 1]$, with parameters $r = 1$, $A = 2.0$ and initial condition $x(0) = 0.1$. The time increment is set to be $\Delta t = 0.01$. Observations are generated at every $\Delta t = 0.05$ time unit by adding the measurement error u_i (i is observation time) following $N(0, R)$ to the deterministic solutions.

Same as linear problem, the 4th-order PC approximation ($N = 4$) method is adopted to solve the stochastic model with parameter and uncertainties in Eq. (52). In the experiment, three different values 0.06^2 , 0.2^2 and 0.5^2 for R , and four different values 10 , 10^2 , 10^3 , 10^4 for the number of ensemble members M are studied.

Figures 18, 19 and Table 2 are given in the same way as those for linear problem. Figures 18 and 19 show the case when $R = 0.5^2$ and $M = 10^4$. Table 2 gives the detailed $RMSE$ values for different selection of R and M . As can be seen in

Figs. 18, 19 and Table 2, PC method can be used in combination with data assimilation scheme (here EnSRF) to produce analysis which has less difference to observation than Open-Loop solution without assimilation does.

5 Conclusions

Using simple scalar linear and nonlinear models we have amply demonstrated the power of the PC based approach. We can readily create forecast ensemble using the PC based spectral expansion as in (24) and can derive histogram of forecast. This histogram can be the basis for generating various types of forecast products. We also combined the above forecast ensemble with EnSRF scheme to perform data assimilation in linear and nonlinear problems. By varying R , the observation variance and M , the number of ensemble members, we have shown the effectiveness of the PC based method for dynamic data assimilation. For simplicity, the polynomials used in this chapter are univariate Hermite (for Gaussian variables) polynomials. In general case, multivariate polynomials from Askey-scheme may be used.

The framework presented in the chapter is actually the stochastic Galerkin (SG) method in which the original dynamic system has been transformed to a set of equations for expansion coefficients. Though Galerkin method is effective and has been adopted in various applications (Ghanem and Spanos 1991; Xiu and Karniadakis 2002a, b, 2003; Babuska et al. 2004; Le Maître et al. 2004; Frauenfelder et al. 2005), there are some limitations to SG method. From the implementation perspective, the process of deriving PC equations is sometimes tedious and challenging. When the governing equations take complicated forms, e.g. highly complex and nonlinear equations, it is difficult to derive the explicit equations for the PC coefficients. To overcome the disadvantage of SG method, the stochastic Collocation (SC) method has been investigated (Xiu 2007). Refer to (Xiu 2009) for the comparison of SG and SC in the context of complex dynamic system. The future work will be applying PC approach in more complex dynamic systems and the study on its efficiency and effectiveness.

Acknowledgements We are grateful to Dongbin Xiu and an anonymous reviewer for their comments on the earlier version of this chapter.

Appendix A

Hermite Polynomials

In this appendix we provide a succinct characterization of the deterministic Hermite polynomials in single and multiple variables.

1. Hermite polynomial—Scalar Case:

The Hermite polynomial $H_m(x)$ of degree m in a scalar variable x is defined by (Kuo 2006).

$$H_m(x) = (-1)^m e^{\frac{x^2}{2}} \frac{d^m}{dx^m} \left[e^{-\frac{x^2}{2}} \right]. \quad (65)$$

There are number of equivalent characterizations (Kuo 2006) of $H_m(x)$. In particular, the generating function for $\{H_m(x)\}_{m \geq 0}$ is given by

$$e^{tx - \frac{t^2}{2}} = \sum_{m=0}^{\infty} \frac{t^m}{m!} H_m(x). \quad (66)$$

In generating a polynomial for a specific degree m , the following formula is useful:

$$H_m(x) = \sum_{k=0}^{\lfloor \frac{m}{2} \rfloor} (-1)^k \binom{m}{2k} (2k-1)!! x^{m-2k}. \quad (67)$$

where $[x]$ denotes the integer part of x ,

$$\binom{m}{2k} = \frac{m!}{2k!(m-2k)!},$$

and $m!$ is the usual factorial of m and $(2k-1)!! = 1 \cdot 3 \cdot 5 \cdot \dots \cdot (2k-1)$.

2. Orthogonality property:

Let

$$w(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}. \quad (68)$$

denotes the standard Gaussian density function.

For integers m and k , define the inner product

$$\langle H_m, H_k \rangle_w = \int_{-\infty}^{\infty} H_m(x) H_k(x) w(x) dx. \quad (69)$$

This inner product induces a norm $\|H_m(x)\|_w$ of $H_m(x)$ defined by

$$\|H_m(x)\|_w^2 = \int_{-\infty}^{\infty} H_m^2(x) w(x) dx. \quad (70)$$

It can be verified that

$$\langle H_m, H_k \rangle_w = \|H_m\|_w^2 \delta_{mk}. \quad (71)$$

where $\delta_{mk} = 0$, if $m \neq k$; and $= 1$, if $m = k$. That is, H_m and H_k are orthogonal for $m \neq k$.

$$\|H_m\|_w^2 = m!. \quad (72)$$

Consequently, $\{H_m(x)\}_{m \geq 0}$ constitutes an orthogonal system of polynomials and $\left\{\frac{H_m(x)}{\sqrt{m!}}\right\}_{m \geq 0}$ constitutes an orthonormal system.

Examples of $H_m(x)$ for $0 \leq m \leq 4$ are given in Table 3.

3. Hermite polynomials—Multivariate Case:

Let $\mathbf{x} = (x_1, x_2, \dots, x_n)^T \in R^n$ and define the n -variate weight function

$$W(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{\mathbf{x}^T \mathbf{x}}{2}} = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-\frac{x_i^2}{2}} = \prod_{i=1}^n w(x_i). \quad (73)$$

where $w(x_i)$ is defined in (68). Let

$$m = p_1 + p_2 + \dots + p_n \text{ with } 0 \leq p_i \leq m. \quad (74)$$

be an additive partition of the integer $m \geq 0$. We define such a partition of m as a n -tuple (p_1, p_2, \dots, p_n) . For example, when $n = 2$, there are 3 partitions — $\{(2, 0), (1, 1), (0, 2)\}$ of $m = 2$.

Given m and one of its partitions (p_1, p_2, \dots, p_n) , define an n -variate homogeneous Hermite polynomial of degree m ,

$$H_{p_1 p_2 \dots p_n}(\mathbf{x}) = (-1)^m e^{\frac{\mathbf{x}^T \mathbf{x}}{2}} \frac{\partial^m}{\partial x_1^{p_1} \partial x_2^{p_2} \dots \partial x_n^{p_n}} e^{-\frac{\mathbf{x}^T \mathbf{x}}{2}}. \quad (75)$$

Using (73) in (75), it can be verified that

$$H_{p_1 p_2 \dots p_n}(\mathbf{x}) = \prod_{i=1}^n (-1)^{p_i} e^{\frac{x_i^2}{2}} \left(\frac{\partial^{p_i}}{\partial x_i^{p_i}} e^{-\frac{x_i^2}{2}} \right) = \prod_{i=1}^n H_{p_i}(x_i). \quad (76)$$

Table 3 A list of $H_m(x)$, $0 \leq m \leq 4$

Degree m	$H_m(x)$	$\ H_m\ _w^2$
0	1	1
1	x	1
2	$x^2 - 1$	2
3	$x^3 - 3x$	$3! = 6$
4	$x^4 - 6x^2 + 3$	$4! = 24$

By combining the multiplicative decomposition of the multi-variate Hermite polynomials in terms of the univariate Hermite polynomials given in (76) and the orthogonality of the latter expressed in (69)–(71), we can readily infer the orthogonality of the multivariate Hermite polynomials.

Hence, if $m = p_1 + p_2 + \dots + p_n$ and $k = q_1 + q_2 + \dots + q_n$, then

$$\begin{aligned} \langle H_m(\mathbf{x}), H_k(\mathbf{x}) \rangle_W &= 0, \text{ if } m \neq k \\ &= \prod_{i=1}^n H_{p_i}(x_i)_W^2 \delta_{p_i q_i}, \text{ if } m = k. \end{aligned} \quad (77)$$

Clearly,

$$\|H_{p_1 p_2 \dots p_n}(\mathbf{x})\|_W^2 = \prod_{i=1}^n p_i!. \quad (78)$$

While there is a unique total ordering of singly indexed scalar Hermite polynomials $H_k(x)$ as shown in Table 3, there are many ways of ordering the multi-indexed Hermite polynomials in (76).

A useful ordering of this latter class of polynomials is called **graded lexicographic ordering**. In this ordering scheme, polynomials of lower total degree precede those of higher degree. Among polynomials of same degree, the members are ordered according to the lexicographic order induced by the natural ordering of the indeterminates, that is $x_1 > x_2 > \dots > x_n$. Thus, polynomials degree one precede those of degree two. For $m = 3, n = 2$, the lexicographic ordering of the two tuples is given by $(3, 0), (2, 1), (1, 2), (0, 3)$.

It can be verified that there are exactly $\binom{m+n-1}{n}$ members in the lexicographic ordering of n tuples (p_1, p_2, \dots, p_n) , such that $\sum_{i=1}^n p_i = m$. Hence, there are this many linearly independent n -variate Hermite polynomials of degree m . Further, it can be verified that the total number of linearly independent n -variate Hermite polynomials of degree less than or equal to m is given by $\binom{m+n}{n}$.

Table 4 provides a list of the set of all 15 two variate ($n = 2$) Hermite polynomials of degree less than or equal to 4. The last column in Table 4 gives the value of $\|H_{p_1 p_2}(x_1, x_2)\|_W^2$.

4. Hilbert Space:

Let $L_2 = L_2(\mathbb{R}^n, W)$ denote the set of all square integrable functions on \mathbb{R}^n , that is

$$L_2 = \left\{ f: \mathbb{R}^n \rightarrow \mathbb{R}: \int_{\mathbb{R}^n} f^2(x) W(x) dx < \infty \right\}. \quad (79)$$

where W is defined in (73). If $f, g \in L_2$, then a natural inner product on L_2 is defined by

Table 4 Two variate ($n = 2$) Hermite polynomials, degree less than or equal to $m = 4$

Degree m	Multi index	$H_{p_1 p_2}(x_1, x_2)$	$H_{p_1}(x_1) H_{p_2}(x_2)$	
0	0 0	1	1	1
1	1 0	x_1	$H_1(x_1) H_0(x_2)$	1
	0 1	x_2	$H_0(x_1) H_1(x_2)$	1
2	2 0	$x_1^2 - 1$	$H_2(x_1) H_0(x_2)$	2
	1 1	$x_1 x_2$	$H_1(x_1) H_1(x_2)$	1
	0 2	$x_2^2 - 1$	$H_0(x_1) H_2(x_2)$	2
3	3 0	$x_1^3 - 3x_1$	$H_3(x_1) H_0(x_2)$	6
	2 1	$x_1^2 x_2 - x_2$	$H_2(x_1) H_1(x_2)$	2
	1 2	$x_1 x_2^2 - x_1$	$H_1(x_1) H_2(x_2)$	2
	0 3	$x_2^3 - 3x_2$	$H_0(x_1) H_3(x_2)$	6
4	4 0	$x_1^4 - 6x_1^2 + 3$	$H_4(x_1) H_0(x_2)$	24
	3 1	$x_1^3 x_2 - 3x_1 x_2$	$H_3(x_1) H_1(x_2)$	6
	2 2	$x_1^2 x_2^2 - x_1^2 - x_2^2 + 1$	$H_2(x_1) H_2(x_2)$	4
	1 3	$x_1 x_2^3 - 3x_1 x_2$	$H_1(x_1) H_3(x_2)$	6
	0 4	$x_2^4 - 6x_2^2 + 3$	$H_0(x_1) H_4(x_2)$	24

$$\langle f, g \rangle_W = \int_n f(x) g(x) W(x) dx. \quad (80)$$

Hence, the norm $\|f_W\|$ is defined by

$$\|f\|_W^2 = \int_{R^n} f^2(x) W(x) dx. \quad (81)$$

It is well known that L_2 is a Hilbert space which is an infinite dimensional, complete, normal linear space where the norm is induced by the inner product.

5. Basis for L_2 .

Let P_m denote the linear span of set of all the n -variate Hermite polynomials $H_k(\mathbf{x})$ of degree $k \leq m$. That is,

$$P_m = \left\{ P(\mathbf{x}) \mid \sum_{k=0}^m \sum_{p_1+p_2+\dots+p_n=k} a_{p_1, p_2, \dots, p_n} H_{p_1, p_2, \dots, p_n}(\mathbf{x}) \right\} \text{ and } p_1, p_2, \dots, p_n \in R \quad (82)$$

Since there are $\binom{m+n}{n}$ linearly independent n -variate Hermite polynomials of degree less than or equal to m , P_m is a linear vector space of finite dimension. Let P_{m-1}^\perp denote the orthogonal complement of P_{m-1} . That is, members of P_m and

P_{m-1}^\perp are mutually orthogonal. Now define the set of all homogeneous polynomials HP_m of degree exactly equal to m as

$$HP_m = P_m \cap P_{m-1}^\perp. \quad (83)$$

It can be verified that members of HP_m are mutually orthogonal to those in P_{m-1} .

Define

$$HP = \bigoplus_{m=0}^{\infty} HP_m. \quad (84)$$

The direct sum of homogeneous polynomials of degree $m \geq 0$. Clearly, for a fixed n ,

$$|HP| = \lim_{N \rightarrow \infty} \sum_{m=0}^N \binom{m+n-1}{n} = \infty. \quad (85)$$

We now state a number of properties without proof.

P1. Basis for L_2 :

$HP \subset L_2$ and HP constitutes a basis for L_2 . Thus, any $f \in L_2$ can be uniquely expressed as

$$f(\mathbf{x}) = \sum_{m=0}^{\infty} \sum_{p_1+p_2+\dots+p_n=m} a_{p_1, p_2, \dots, p_n} H_{p_1, p_2, \dots, p_n}(\mathbf{x}). \quad (86)$$

where

$$a_{p_1, p_2, \dots, p_n} = \frac{\langle f(\mathbf{x}), H_{p_1, p_2, \dots, p_n}(\mathbf{x}) \rangle}{\|H_{p_1, p_2, \dots, p_n}(\mathbf{x})\|^2}. \quad (87)$$

P2. Orthogonal Projection:

Let, for any N finite, define

$$\prod_N(f) = f_N(\mathbf{x}) = \sum_{m=0}^N \sum_{p_1+p_2+\dots+p_n=m} a_{p_1, p_2, \dots, p_n} H_{p_1, p_2, \dots, p_n}(\mathbf{x}). \quad (88)$$

Then it can be verified that $\prod_N(f)$ is orthogonal projection of $f(\mathbf{x})$ onto the subspace P_N .

Define the error in the projection as

$$\varepsilon_N(\mathbf{x}) = f(\mathbf{x}) - f_N(\mathbf{x}). \quad (89)$$

Then, it can be verified that

$$\langle f_N(\mathbf{x}), \varepsilon_N(\mathbf{x}) \rangle = 0 \text{ for each } N > 0. \quad (90)$$

P3. Mean Square Convergence:

It can be verified that

$$\lim_{N \rightarrow \infty} \|f(\mathbf{x}) - f_N(\mathbf{x})\|^2 = 0.$$

i.e. the quality of the projection $f_N(\mathbf{x})$ improves as N grows large.

Example A.1 From (66), it follows that

$$e^{tx} = e^{\frac{t^2}{2}} \sum_{m=0}^{\infty} \frac{t^m}{m!} H_m(x). \quad (91)$$

and

$$e^{-tx} = e^{\frac{t^2}{2}} \sum_{m=0}^{\infty} \frac{t^m}{m!} (-1)^m H_m(x). \quad (92)$$

Since $H_m(x) = H_m(-x)$ for m even and $H_m(-x) = -H_m(x)$ for m odd. By truncating, the infinite sum in (91) and (92), we can get a good family of approximation to e^{tx} and e^{-tx} . Using these we can readily obtain approximation to $\sin(tx)$, $\cos(tx)$, etc.

It would be a good exercise to compute the quality of these approximations for varying degree of truncation and ranges of values for x .

Appendix B

Hermite Polynomial Chaos

Let (Ω, \mathcal{F}, P) be a standard probability space where Ω represents the set of all elementary events, \mathcal{F} denotes the σ -algebra of subsets of simple events and P is the probability measure defined on the members of \mathcal{F} . Let $x: \Omega \rightarrow \mathbb{R}$ be a real valued random variable. Let $P_x(x)$ be the distribution induced by x and let $p_x(x)$ be the corresponding probability density function. Then, the properties of x can be equivalently described using the triplet $(\mathbb{R}, \mathcal{B}, p_x(x))$ where \mathcal{B} denotes the Borel σ -algebra over \mathbb{R} , and $p_x(x)$ is the density of x .

Let x and y be two real valued random variables with joint density $p_{x,y}(x,y)$. Define an inner product

$$\langle x, y \rangle = E(xy) = \int_{\Omega} x(w)y(w)dp(w) = \int_{\mathbb{R}} \int_{\mathbb{R}} xy p_{x,y}(x,y) dx dy. \quad (93)$$

and the corresponding norm (second moment)

$$\|x\|^2 = E(x^2) = \int_{\Omega} x^2(w) dp(w) = \int_{\mathbb{R}} x^2 p_x(x) dx. \quad (94)$$

Let $L_2(\Omega)$ denotes the set of all random variables with finite second moment, that is,

$$L_2(\Omega) = \left\{ x: \Omega \rightarrow \mathbb{R} \mid \int_{\Omega} x^2(w) dp(w) < \infty \right\}. \quad (95)$$

It can be verified $L_2(\Omega)$ is a Hilbert space.

Let x be a Gaussian random variable with mean m and variance σ^2 . Then

$$p_x(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(x-m)^2}{2\sigma^2}\right]. \quad (96)$$

A random variable is said to be centered if its expectation is zero. Let x be a centered Gaussian, that is, $x \sim N(0, \sigma^2)$. Then it can be verified

$$\begin{aligned} E(x^m) &= 0 \quad \text{if } m \text{ is odd.} \\ &= 1 \cdot 3 \cdot 5 \cdots (m-1) \sigma^m \text{ if } m \text{ is even.} \end{aligned} \quad (97)$$

Let G denotes the set of all centered Gaussian random variables (differing only in their variance). Clearly, G is an infinite set. Let \overline{G} denotes the closure of the linear span of G . It can be verified $\overline{G} \subset L_2(\Omega)$ and is itself a Hilbert space, called the Gaussian Hilbert space. If $\xi_1, \xi_2, \dots, \xi_n \in \overline{G}$, then it can be well known that

$$E(\xi_1 \xi_2 \dots \xi_n) = \sum \prod_{(i_1, i_2)} E(\xi_{i_1} \xi_{i_2}). \quad (98)$$

where (i_1, i_2) runs through the pairwise distinct partition of $\{1, 2, \dots, n\}$ and the sum is over all such partitions. For example,

$$E(\xi_1 \xi_2 \xi_3 \xi_4) = E(\xi_1 \xi_2) E(\xi_3 \xi_4) + E(\xi_1 \xi_3) E(\xi_2 \xi_4) + E(\xi_1 \xi_4) E(\xi_2 \xi_3). \quad (99)$$

Recall, the probability density $p(\xi)$ of a standard Gaussian random variable ξ is given by

$$p(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{\xi^2}{2}\right]. \quad (100)$$

Since $p(\xi)$ in (100) is the same as $w(x)$ in (68), it is immediate that the properties of the Hermite polynomials $H_m(x)$ given in Appendix A directly carry over to $H_m(\xi)$ where ξ is a centered standard Gaussian random variable.

We now state the following properties of $H_m(\xi)$ which can be easily verified.

- (1) Examples of $H_m(\xi)$ are obtained by replacing x by ξ in Table 3.
- (2) $E[H_m(\xi)] = 0$, for $m > 0$.
- (3) $\{H_m(\xi)\}_{m \geq 0}$ are orthogonal, that is

$$\langle H_m(\xi), H_k(\xi) \rangle = E[H_m(\xi)H_k(\xi)] = 0 \text{ if } m \neq k.$$

- (4) The norm of $H_m(\xi)$ is

$$\langle H_m(\xi), H_m(\xi) \rangle = \|H_m(\xi)\|^2 = E[H_m^2(\xi)] = m!.$$

- (5) $\left\{ \frac{H_m(\xi)}{\sqrt{m!}} \right\}_{m \geq 0}$ form an orthonormal system of Hermite polynomials.

The above properties can be readily extended to multivariate Hermite polynomials over a set of n centered Gaussian random variables $\xi_1, \xi_2, \dots, \xi_n$.

- (6) Let $p_1 + p_2 + \dots + p_n = m$, where $0 \leq p_i \leq m$ for $1 \leq i \leq n$. Then

$$H_{p_1, p_2, \dots, p_n}(\xi_1, \xi_2, \dots, \xi_n) = H_{p_1}(\xi_1)H_{p_2}(\xi_2) \dots H_{p_n}(\xi_n).$$

- (7)

$$\|H_{p_1, p_2, \dots, p_n}(\xi_1, \xi_2, \dots, \xi_n)\|^2 = \text{var}[H_{p_1, p_2, \dots, p_n}(\xi_1, \xi_2, \dots, \xi_n)] = p_1!p_2! \dots p_n!.$$

- (8) The definitions of the sets P_m , $H P_m$ directly carry over to the case of Hermite polynomial over a finite set of centered Gaussian random variables.
- (9) The linear space HP_m are called m^{th} order polynomial chaos in n centered Gaussian random variables.
- (10) P_m is called homogeneous chaos in n centered Gaussian random variables.

Appendix C

In this Appendix we derive the functional form of the mapping $\bar{M}: \mathbb{R}^{N+1} \rightarrow \mathbb{R}^{N+1}$ in (36) by taking inner product of both sides of (34):

$$x(k+1) = M[x(k), \alpha] + \bar{q}\sqrt{\Delta t}\phi_1(\xi). \quad (101)$$

with $\phi_i(\xi)$ for $0 \leq i \leq N$. Using the PC expansion for $x(k)$ in (33), it follows that

$$i!v_i(k+1) = \langle x(k+1), \phi_i(\xi) \rangle \quad (102)$$

Similarly,

$$\langle \bar{q} \sqrt{\Delta t} \phi_1(\xi), \phi_i(\xi) \rangle = \bar{q} \sqrt{\Delta t} \delta_{1i}. \quad (103)$$

Then we readily get the expression for the forcing term $\mathbf{g}(k)$ in (36). According to (35),

$$M[x(k), \alpha] = [1 + \Delta t(\bar{\alpha} + \sigma \phi_1(\xi))] \sum_{j=0}^N v_j(k) \phi_j(\xi). \quad (104)$$

Let

$$\langle M[x(k), \alpha], \phi_i(\xi) \rangle = \text{TermI} + \text{TermII}. \quad (105)$$

where

$$TermI = (1 + \bar{\alpha} \Delta t) \langle x(k), \phi_i(\xi) \rangle = i! (1 + \bar{\alpha} \Delta t) v_i(k). \quad (106)$$

and

$$TermII = \sigma \Delta t \langle \phi_1(\xi) \sum_{j=0}^N v_j(k) \phi_j(\xi), \phi_i(\xi) \rangle. \quad (107)$$

Here, $N = 4$ and $N = 8$ are considered. The values of these $(N+1) \times (N+1)$ inner products in (107) are listed in Table 5.

Combining (101)–(107) with the entries in Table 5, we get the explicit form of the spectral dynamics in (36) as follows:

For $N = 4$,

$$v_0(k+1) = (1 + \bar{\alpha} \Delta t) v_0(k) + \sigma \Delta t v_1(k)$$

Table 5 Values of the 5×5 inner product $\langle x, y \rangle$, where the 5 values of x are given in the row and values of y in the column

$$\begin{aligned}
v_1(k+1) &= (1 + \bar{\alpha} \Delta t)v_1(k) + \sigma \Delta t[v_0(k) + 2v_2(k)] + \bar{q}\sqrt{\Delta t}. \\
v_2(k+1) &= (1 + \bar{\alpha} \Delta t)v_2(k) + \frac{\sigma \Delta t}{2!}[2v_1(k) + 6v_3(k)]. \\
v_3(k+1) &= (1 + \bar{\alpha} \Delta t)v_3(k) + \frac{\sigma \Delta t}{3!}[6v_2(k) + 24v_4(k)] \\
v_4(k+1) &= (1 + \bar{\alpha} \Delta t)v_4(k) + \frac{\sigma \Delta t}{4!}[24v_3(k)]. \tag{108}
\end{aligned}$$

For $N = 8$,

$$\begin{aligned}
v_0(k+1) &= (1 + \bar{\alpha} \Delta t)v_0(k) + \sigma \Delta t v_1(k). \\
v_1(k+1) &= (1 + \bar{\alpha} \Delta t)v_1(k) + \sigma \Delta t[v_0(k) + 2v_2(k)] + \bar{q}\sqrt{\Delta t}. \\
v_2(k+1) &= (1 + \bar{\alpha} \Delta t)v_2(k) + \frac{\sigma \Delta t}{2!}[2v_1(k) + 6v_3(k)]. \\
v_3(k+1) &= (1 + \bar{\alpha} \Delta t)v_3(k) + \frac{\sigma \Delta t}{3!}[6v_2(k) + 24v_4(k)] \\
v_4(k+1) &= (1 + \bar{\alpha} \Delta t)v_4(k) + \frac{\sigma \Delta t}{4!}[24v_3(k) + 120v_5(k)]. \tag{109} \\
v_5(k+1) &= (1 + \bar{\alpha} \Delta t)v_5(k) + \frac{\sigma \Delta t}{5!}[120v_4(k) + 720v_6(k)]. \\
v_6(k+1) &= (1 + \bar{\alpha} \Delta t)v_6(k) + \frac{\sigma \Delta t}{6!}[720v_5(k) + 5040v_7(k)]. \\
v_7(k+1) &= (1 + \bar{\alpha} \Delta t)v_7(k) + \frac{\sigma \Delta t}{7!}[5040v_6(k) + 40320v_8(k)] \\
v_8(k+1) &= (1 + \bar{\alpha} \Delta t)v_8(k) + \frac{\sigma \Delta t}{8!}[40320v_7(k)].
\end{aligned}$$

Appendix D

In this Appendix we derive the functional form of the mapping $\bar{M}: R^{N+1} \rightarrow R^{N+1}$ in (36) for nonlinear problem by taking inner product of both sides of (43):

$$x(k+1) = M[x(k), \alpha, \beta] + \bar{q}\sqrt{\Delta t}\phi_1(\xi). \tag{110}$$

with $\phi_i(\xi)$ for $0 \leq i \leq N$. Using the PC expansion for $x(k)$ in (49), it follows that

$$i!v_i(k+1) = \langle x(k+1), \phi_i(\xi) \rangle. \tag{111}$$

Similarly,

$$\langle \bar{q}\sqrt{\Delta t}\phi_1(\xi), \phi_i(\xi) \rangle = \bar{q}\sqrt{\Delta t}\delta_{1i}. \quad (112)$$

According to (51), let

$$\langle M[x(k), \alpha, \beta], \phi_i(\xi) \rangle = \text{TermI} + \text{TermII}. \quad (113)$$

Where

$$\begin{aligned} \text{TermI} &= (1 - \bar{\alpha} \Delta t - \sigma_1 \phi_1(\xi) \Delta t) \langle x(k), \phi_i(\xi) \rangle \\ &= i!(1 - \bar{\alpha} \Delta t) v_i(k) - \sigma_1 \Delta t \langle \phi_1(\xi) x(k), \phi_i(\xi) \rangle. \end{aligned} \quad (114)$$

and

$$\text{TermII} = \Delta t \langle \frac{\bar{\alpha} + \sigma_1 \phi_1(\xi)}{\beta + \sigma_2 \phi_1(\xi)} \left[\sum_{j=0}^N v_j(k) \phi_j(\xi) \right]^2, \phi_i(\xi) \rangle. \quad (115)$$

In our experiment, we pay attention to the deterministic parameter A , i.e. $\sigma_2 = 0$. TermII becomes

$$\text{TermII} = \frac{\bar{\alpha} + \sigma_1 \phi_1(\xi)}{\beta} \Delta t \langle \left[\sum_{j=0}^N v_j(k) \phi_j(\xi) \right]^2, \phi_i(\xi) \rangle. \quad (116)$$

Here, the explicit form of the spectral dynamics in (36) when $N = 4$ is as follows:

$$\begin{aligned} v_0(k+1) &= (1 - \bar{\alpha} \Delta t) v_0(k) - \sigma_1 \Delta t v_1(k) \\ &+ \frac{\bar{\alpha}}{\beta} \Delta t \left([v_0(k)]^2 + [v_1(k)]^2 + 2[v_2(k)]^2 + 6[v_3(k)]^2 + 24[v_4(k)]^2 \right) \\ &+ 2 \frac{\sigma_1 \Delta t}{\beta} [v_0(k)v_1(k) + 2v_1(k)v_2(k) + 6v_2(k)v_3(k) + 24v_3(k)v_4(k)]. \end{aligned}$$

$$\begin{aligned} v_1(k+1) &= (1 - \bar{\alpha} \Delta t) v_1(k) - \sigma_1 \Delta t (v_0(k) + 2v_2(k)) \\ &+ \frac{\bar{\alpha}}{\beta} \Delta t [2v_0(k)v_1(k) + 4v_1(k)v_2(k) + 12v_2(k)v_3(k) + 48v_3(k)v_4(k)] \\ &+ \frac{\sigma_1 \Delta t}{\beta} \left([v_0(k)]^2 + 4v_0(k)v_2(k) + 3[v_1(k)]^2 + 12v_1(k)v_3(k) + 10[v_2(k)]^2 \right. \\ &\quad \left. + 48v_2(k)v_4(k) + 42[v_3(k)]^2 + 216[v_4(k)]^2 \right) + \bar{q}\sqrt{\Delta t}. \end{aligned}$$

$$\begin{aligned}
v_2(k+1) = & (1 - \bar{\alpha} \Delta t) v_2(k) - \sigma_1 \Delta t (v_1(k) + 3v_3(k)) \\
& + \frac{\bar{\alpha}}{2!\beta} \Delta t \left(4v_0(k)v_2(k) + 2[v_1(k)]^2 + 12v_1(k)v_3(k) + 8[v_2(k)]^2 \right. \\
& \left. + 48v_2(k)v_4(k) + 36[v_3(k)]^2 + 192[v_4(k)]^2 \right) \\
& + \frac{4\sigma_1 \Delta t}{2!\beta} (v_0(k)v_1(k) + 3v_0(k)v_3(k) + 5v_1(k)v_2(k) \\
& + 12v_1(k)v_4(k) + 24v_2(k)v_3(k) + 132v_3(k)v_4(k)) \\
v_3(k+1) = & (1 - \bar{\alpha} \Delta t) v_3(k) - \sigma_1 \Delta t (v_2(k) + 4v_4(k)) \\
& + \frac{\bar{\alpha}}{3!\beta} \Delta t (12v_0(k)v_3(k) + 12v_1(k)v_2(k) \\
& + 48v_1(k)v_4(k) + 72v_2(k)v_3(k) + 432v_3(k)v_4(k)) \\
& + \frac{6\sigma_1 \Delta t}{3!\beta} \left([v_1(k)]^2 + 14v_1(k)v_3(k) + 8[v_2(k)]^2 + 88v_2(k)v_4(k) \right. \\
& \left. + 2v_0(k)v_2(k) + 54[v_3(k)]^2 + 384[v_4(k)]^2 + 8v_0(k)v_4(k) \right) \\
v_4(k+1) = & (1 - \bar{\alpha} \Delta t) v_4(k) - \sigma_1 \Delta t v_3(k) \\
& + \frac{\bar{\alpha}}{4!\beta} \Delta t \left(48v_0(k)v_4(k) + 48v_1(k)v_3(k) + 24[v_2(k)]^2 \right. \\
& \left. + 384v_2(k)v_4(k) + 216[v_3(k)]^2 + 1728[v_4(k)]^2 \right) \\
& + \frac{48\sigma_1 \Delta t}{4!\beta} (v_0(k)v_3(k) + v_1(k)v_2(k) + 9v_1(k)v_4(k) \\
& + 11v_2(k)v_3(k) + 96v_3(k)v_4(k))
\end{aligned}$$

References

- Arnold L (1974) Stochastic Differential Equations - Theory & Applications. Wiley, New York
- Babuska I, Tempone R, Zouraris GE (2004) Galerkin finite element approximations of stochastic elliptic partial differential equations. *SIAM J Numer Anal* 42(2):800–825
- Crisan D, Rozovskii B (2011) The oxford handbook of nonlinear filtering. Oxford Handbook in Mathematics, Oxford, England
- Evensen G (2007) Data assimilation: the ensemble Kalman filter. Springer, New York
- Frauenfelder P, Schwab C, Todor RA (2005) Finite elements for elliptic problems with stochastic coefficients. *Comput Methods Appl Mech Eng* 194(2):205–228
- Ghanem R, Spanos PD (1991) Stochastic finite elements: a spectral approach. Springer, New York
- Ghanem R (1999) Ingredients for a general purpose stochastic finite elements implementation. *Comput Methods Appl Mech Eng* 168(1):19–34
- Grigoriu M (2012) Stochastic systems: uncertainty quantification and propagation. Springer
- Jazwinski AH (1970) Stochastic processes and filtering theory. Academic Press, New York
- Kallianpur G (1980) Stochastic filtering theory. Springer

- Kalman RE (1960) A new approach to linear filtering and prediction problems. *J Basic Eng* 82 (1):35–45
- Kuo HH (2006) *Introduction to Stochastic Integration*. Springer
- Kushner HJ (1962) On the differential equations satisfied by conditional probability densities of markov processes, with applications. *SIAM J Control* 2:106–119
- Lakshmivarahan S, Stensrud D (2009) Ensemble Kalman Filter: A innovative approach for meteorological data assimilation. *IEEE Control Syst Soc Special Issue* 29:34–46
- Le Maître OOP, Knio OM, Najm HN, Ghanem RG (2004) Uncertainty propagation using Wiener-Haar expansions. *J Comput Phys* 197(1):28–57
- Le Maître OOP, Knio OM (2010) *Spectral methods for uncertainty quantification: with applications to computational fluid dynamics*. Springer
- Lewis JM, Lakshmivarahan S, Dhall S (2006) *Dynamic data assimilation: a least squares approach*, vol 104. Cambridge University Press
- Li J, Xiu D (2009) A generalized polynomial chaos based ensemble Kalman filter with high accuracy. *J Comput Phys* 228(15):5454–5469
- Loève M (1977) *Probability theory*. Graduate Texts in Mathematics, vol 45
- Lototsky S, Rozovskii B (2006) *Stochastic differential equations: a Wiener chaos approach. From stochastic calculus to mathematical finance*. Springer, pp 433–506
- Lototsky S (2011) *Chaos approach to nonlinear filtering*. Oxford University Press, pp 231–264
- Saaty TL (1967) *Modern nonlinear equations*. McGraw-Hill, New York, Chapter 8
- Soong TT (1973) *Random differential equations in science and engineering*. Academic Press, New York
- Wiener N (1938) The homogeneous chaos. *Am J Math* 60(4):897–936
- Xiu D, Karniadakis GE (2002a) The wiener-askey polynomial chaos for stochastic differential equations. *SIAM J Sci Comput* 24(2):619–644
- Xiu D, Karniadakis GE (2002b) Modeling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Comput Methods Appl Mech Eng* 191(43):4927–4948
- Xiu D, Karniadakis GE (2003) Modeling uncertainty in flow simulations via generalized polynomial chaos. *J Comput Phys* 187(1):137–167
- Xiu D (2007) Efficient collocation approach for parametric uncertainty analysis. *Commun Comput Phys* 2(2):293–309
- Xiu D (2009) Fast numerical methods for stochastic computations: a review. *Commun Comput Phys* 5(2–4):242–272
- Xiu D (2010) *Numerical methods for stochastic computations: a spectral method approach*. Princeton University Press
- Zakai M (1969) On the optimal filtering of diffusion processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* 11:230–243

The Treatment, Estimation, and Issues with Representation Error Modelling

Daniel Hodyss and Elizabeth Satterfield

Abstract This chapter links the theory of data assimilation to representation error and then illustrates several issues with applying observation covariance estimation algorithms to contemporary data assimilation systems. This work will show that contemporary data assimilation systems cannot precisely identify representation error using standard estimation procedures, and this issue will be traced back to the definition of what is the “optimal” data assimilation system for forecast models that represent the true system using a truncated state space.

1 Introduction

Data assimilation schemes blend observational data, with limited coverage, with a short term forecast to produce an analysis, which is meant to be the best estimate of the atmospheric state. Appropriately specifying error statistics is necessary to obtain an optimal analysis. Errors associated with sub-grid scale features, or errors of representation, are often poorly accounted for or overlooked entirely. Representation error is a key, if not dominant, contributor to correlated observation errors (Stewart et al. 2010; Weston et al. 2011; Waller et al. 2014, 2015), which are often neglected. However, properly accounting for such errors of representation is more complicated than a more sophisticated model of the observation error covariance, or compensating for neglected, potentially correlated, errors through inflation. More generally, properly accounting for errors of representation changes the Kalman

Submitted as a chapter to.

Data Assimilation for Atmospheric, Oceanic, and Hydrologic Applications, Vol. III.

January 5, 2016.

D. Hodyss (✉) · E. Satterfield

Naval Research Laboratory, Marine Meteorology Division,
7 Grace Hopper Ave., Stop 2, Monterey, CA 93943, USA
e-mail: daniel.hodyss@nrlmry.navy.mil

Filter equations. The Kalman gain must then incorporate terms associated with unresolved scales and the covariance between resolved and unresolved scales. Unfortunately, simplifying assumptions must be made in an operational setting, leading to the typical practice of neglecting such terms.

1.1 *Definition of a “True State”*

In order to discuss error statistics, one must have in mind a verification state about which such errors could be calculated. Lorenc (1986) discussed such a verification state as being obtained by projecting the true state of the atmosphere onto the model basis. (Please see Sect. 2, Eq. (2.5) for further discussion.) Mitchell and Daley (1997b) further explained that model variables at grid points are usually thought of as representing actual point values; however, these values are combined with physical parameterizations that are considered to be a box averaged quantity. Frehlich (2011) discussed the truth for error statistics as the convolution for the continuous atmospheric variables by the effective spatial filter of a model.

Since models are only capable of representing such a filtered true state, the background error does not contain errors associated with the inability of relatively coarse grids to resolve small-scale features, or errors of representation. Such terms must be included in the observation error covariance matrix. However, these terms are often poorly modeled or neglected. Further complicating the issue, appropriate treatment in data assimilation requires the knowledge of the covariance between resolved and unresolved scales as well as the unresolved error covariance.

1.2 *Modifications of the Kalman Filter Equations*

Mitchell and Daley (1997a, b) explored discretization errors both in the numerical models used in atmospheric data assimilation and the forward interpolation from the analysis mesh to the observation. Forward interpolation errors were defined as the difference between the true state in observation space and the model representation of the truth mapped into observation space by an approximate observation operator. Their study extended to the case where errors in the forward interpolation were due to the impacts of unresolved scales as well as an approximate observation operator. To clarify their discussion, they stated that grid point values should represent only those scales that can be represented by a coarse mesh and therefore defined unresolved scales through the resulting though implicit spectral truncation. They considered a generalization of the standard Kalman Filter gain matrix, which would require knowledge not only of the complete forward interpolation error, but also of the covariance between the forecast error and the truth as well as the covariance between the forecast error and the forward interpolation error. Since such a gain matrix would be computationally infeasible in an operational setting, they

consequently defined a more conventional gain matrix, formed by neglecting terms involving the covariance between the forecast errors on the grid and the truth, and examined the behavior of both forms.

Liu and Rabier (2002) defined representation error following Mitchell and Daley (1997a, b) by specifically invoking a spectral truncation of the true state. A simplified one-dimensional framework allowed them to consider under which conditions the best balance between correlation and thinning could be reached and whether the inclusion of observation-error correlation in a cost function could further improve the analysis quality.

Janjić and Cohn (2006) mathematically formalized the issue of representation error through the definition of a projection of a continuum state onto a finite dimensional spectral-space. They also decomposed the forward interpolation error, as defined by Mitchell and Daley (1997a, b), into the sum of two components, the error due to approximating the observation operator and the error of representation, defined as the difference between a perfect observation of the complete state and a perfect observation of the resolved component of the state. Similar to Daley (1997a, b) they compared the performance of two filters: the Schmidt-Kalman filter, similar to the generalized filter formulation of Mitchell and Daley (1997a, b), and the traditional discrete Kalman filter. Numerical experiments demonstrated that approximate filters worked well for the model problem given that the exact covariance of the unresolved scales is known. Unfortunately, the covariance matrix of the unresolved scales is rarely known in practice.

Therefore, Hodyss and Nichols (2015) generalized the work of Mitchell and Daley (1997a, b) by reframing the problem through an arbitrary linear operator which acts on the true state to generate the forecast state. The Hodyss and Nichols (2015) formulation of this problem removes the requirement of explicitly producing an unresolved-scale covariance matrix and places the emphasis on estimating a matrix operator that can map from high-resolution to low-resolution. This generalized framework properly accounts for both unresolved-scale errors of representation as well as resolved-scale errors through a matrix operator that acts to map between the true attracting manifold, which is being observed by the observational instruments, and the forecast attracting manifold that is available to make state estimates.

1.3 *Estimating Observation Error Covariance Matrices in an Operational Setting*

Recently, increased emphasis has been placed on methods to statistically estimate observation error covariance matrices. Methods based on using (observation-minus-background) innovation statistics and separating statistics into observation and background error covariances were introduced by Rutherford (1972) and Höllingsworth and Lonnberg (1986, hereafter HL). The HL method constructs a histogram of innovation covariances binned by separation distance. The histogram can be fit to an

isotropic correlation function and extrapolated to zero distance, providing a partition of innovation variance into correlated and uncorrelated components. The assumption of this method is that the observation error should be uncorrelated, so we can identify the correlated error component as background (forecast) error variance. The uncorrelated observation error variance is then computed by taking the difference between the innovation variance and the background error variance. Although, the HL method is not dependent on prescribed covariance models, this method requires a dense observing network to bin observations by distance, is dependent on the choice of correlation function, and the presence of any correlated observation error can lead to errors in the estimation.

More recently, the Desroziers method (Desroziers et al. 2005), which relies on the assumptions of an optimal data assimilation algorithm, has become increasingly popular due to the simplicity of the required calculations. This method involves taking the expected value of the outer product of the analysis residual and the innovation to obtain the observation error covariance matrix. Although, simple to implement, this method is usually referred to as a consistency check due to its dependence on prescribed covariance models and an iterative method is suggested. Waller et al. (2015) showed that, although this method is subject to prescribed background and observation covariance matrices, a useful solution can often be obtained in a single iteration even when iterative techniques cannot be expected to converge. Further, Ménard (2015) examined the theoretical foundation of the Desroziers' method and carried out a mathematical analysis of convergence, proposing a combination of the Desroziers' scheme and maximum likelihood estimation be used to estimate the spatial correlation length of observation errors.

Both the Desroziers method and the HL method have been used extensively to obtain correlation terms for certain observation types (Stewart et al. 2009, 2014; Borman et al. 2002, 2010; Borman and Bauer 2010; Weston et al. 2014; Waller et al. 2015). Stewart et al. (2012) showed benefit to the analysis even when approximate correlation structures were used. Such correlation terms are, at least in part if not dominantly, due to representation error (Stewart et al. 2010; Weston et al. 2014; Waller et al. 2014). Waller et al. (2014) further showed that the errors of representation are correlated and more significant for humidity than for temperature and vary with altitude.

While the Desroziers method and the HL method can be used to estimate the total observation error covariance matrix, other methods focus on estimating the structure of representation error from observational data and model analyses. Oke and Sakov (2008) and Forgot and Wunch (2007) used observations which were averaged to model resolution and interpolated to represent the model resolved state from which raw observations were subtracted to produce maps of representation error. Richman et al. (2005) produced maps of representation error by taking differences between sequence of innovations and their orthogonal projections on the space spanned by the leading model empirical orthogonal functions (EOFs).

Such observation error covariance estimation schemes produce static estimates from a predefined training period. Frehlich (2006) addressed spatial variations of the observation sampling error, which is dependent on the statistics of small scales,

by calculating observation sampling error covariances based on estimates of local turbulence. In this way, the total observation error, consisting of both instrument error and observation sampling error, was defined to be dependent on the location of the observation with respect to the analysis coordinate. Frelich (2008) further showed that temporal and location dependence of total observation error derived from innovation statistics can be dominated by observation sampling error and therefore determined by atmospheric turbulence.

In this manuscript we will thoroughly link the theory of data assimilation to representation error and then illustrate several issues with applying observation covariance estimation algorithms to contemporary data assimilation systems. In Sect. 2 we review the theory of Hodyss and Nichols (2015). The goal there is to remind the reader that the Kalman gain must be different from the common form when one accounts for representation error. This difference implies that, contrary to typical operational practice, one must not only change the observation error covariance matrix to account for representation error, but also the numerator of the Kalman gain must also be modified. In Sect. 3 we will show that typical observation covariance matrix estimation algorithms only deliver the correct observation error statistics if the correct form of the Kalman gain is used or if there is no model error on the resolved scales. If the typical form of the Kalman gain is used, and even if the diagonal of the observation error covariance matrix has been tuned to account for representation error, the result of common observation error covariance matrix estimation procedures will result in an estimate that includes a portion of the background covariance. In Sect. 4, we build on the work of Waller et al. 2015 and show that mis-specifying the background covariance matrix in the data assimilation algorithm will result in an estimated observation error covariance matrix that also appears to depend on the background covariance matrix. Section 5 closes the manuscript with a summary of the results and their most important conclusions.

2 Representation Error in Data Assimilation

In this section we review the general theory for multi-resolution data assimilation presented in Hodyss and Nichols (2015). Our review will only cover the aspects necessary to build the tools required to analyze common estimation procedures used to find representation error. Here, we will build a Gaussian covariance model and from it deduce the correct state-estimation procedure for the case where the observations are viewing a state with a higher dimension than the available forecast model is capable of reproducing.

A simple way to construct a Gaussian problem that is amenable to analysis is through the use of a discrete Fourier series representation. To this end we assume a Gaussian covariance model for the high-resolution states of the form

$$\mathbf{x}_H = \bar{\mathbf{x}}_H + \mathbf{Z}\boldsymbol{\eta} \quad (2.1)$$

where $\bar{\mathbf{x}}_H$ is an N -vector, \mathbf{Z} is the square-root of the true forecast error covariance matrix,

$$\mathbf{P}_H = \mathbf{Z}\mathbf{Z}^T \quad (2.2)$$

and $\boldsymbol{\eta}$ is an N -vector of random numbers drawn from $N(\mathbf{0}, \mathbf{I})$. We construct (2.2) using a sinusoidal basis in which the columns of \mathbf{E}_H ($N \times N$) contain the sinusoids such that

$$\mathbf{P}_H = \mathbf{E}_H \boldsymbol{\Gamma} \mathbf{E}_H^T \quad (2.3)$$

$\boldsymbol{\Gamma}$ is a diagonal matrix whose i th element of the diagonal defines the weight given to that basis function.

We connect the high-resolution states (of length N) to the low-resolution states (of length M)

$$\mathbf{x}_L = \mathbf{S}\mathbf{x}_H \quad (2.4)$$

through a “smoother” that operates as:

$$\mathbf{S} = \mathbf{E}_L [\mathbf{D}^{1/2} \mathbf{T} \quad \mathbf{0}] \mathbf{E}_H^T \quad (2.5)$$

where \mathbf{D} ($M \times M$) is a diagonal matrix whose diagonal elements are $d_i = e^{-\beta^2 k_i^2}$, \mathbf{E}_L ($M \times M$) is the low-resolution basis whose columns are also the sinusoids, \mathbf{T} ($M \times M$) is a diagonal matrix with the value $\sqrt{M/N}$ along the diagonal, and k_i is the wavenumber associated with the i th basis function of \mathbf{E}_H . If we assume that the columns of \mathbf{E}_L are simply the subsampled columns of \mathbf{E}_H then the interpretation of (2.5) becomes straightforward. The matrix \mathbf{D} represents the climatological “model error” on the resolved scales and would be equal to the identity matrix if the forecast model’s climate at the resolved scales was identical to the true model’s climate at those same scales. The matrix implied by the bracket in (2.5) performs a truncation of the high-resolution basis to the M -dimensional subspace while the matrix \mathbf{T} assures that the Fourier coefficients calculated from the high-resolution basis are reweighted consistently with respect to the low-resolution basis.

Equation (2.5) allows for the creation of the low-resolution forecast states from the high-resolution true states in (2.1). This implies that the low-resolution error covariance matrix may be written as

$$\mathbf{P}_L = \mathbf{S}\mathbf{P}_H\mathbf{S}^T = \mathbf{E}_L [\mathbf{D}^{1/2} \mathbf{T} \quad \mathbf{0}] \boldsymbol{\Gamma} [\mathbf{D}^{1/2} \mathbf{T} \quad \mathbf{0}]^T \mathbf{E}_L^T \quad (2.6)$$

Because $\boldsymbol{\Gamma}$ are the true, high-resolution eigenvalues, Eq. (2.6) shows that the forecast covariance matrix would be correct up to its M eigenvalues if the

climatological model error \mathbf{D} could be removed. We show next how to remove this climatological model error from the state estimate by accounting for the error of representation.

It is shown in Hodyss and Nichols (2015) that the best linear unbiased estimate of the low-resolution forecast state given observations of the high-resolution true state for the problem setup here is

$$\bar{\mathbf{x}}_L^a = \bar{\mathbf{x}}_L + \mathbf{G}[\mathbf{v}_L - \langle \mathbf{v}_L \rangle] \quad (2.7)$$

where

$$\mathbf{G} = [\mathbf{P}_L \mathbf{H}_L^T + \mathbf{P}_{LH}] \left[\mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T + \bar{\mathbf{R}}_L^* \right]^{-1} \quad (2.8)$$

$$\mathbf{P}_{LH} = \mathbf{S} \mathbf{P}_H (\mathbf{H}_H - \mathbf{H}_L \mathbf{S})^T \quad (2.9)$$

$$\bar{\mathbf{R}}_L^* = \mathbf{R}_{ins} + \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T \quad (2.10)$$

$$\mathbf{v}_L = \mathbf{y} - \mathbf{H}_L \bar{\mathbf{x}}_L \quad (2.11)$$

$$\langle \mathbf{v}_L \rangle = \mathbf{H}_H \bar{\mathbf{x}}_H - \mathbf{H}_L \bar{\mathbf{x}}_L \quad (2.12)$$

In (2.7) through (2.12), \mathbf{y} is a p -vector of observations, \mathbf{H}_H is a $p \times N$ observation operator, $\bar{\mathbf{x}}_H$ is the high-resolution prior mean, $\bar{\mathbf{x}}_L$ is low-resolution prior mean, and \mathbf{R}_{ins} is the instrument observation covariance matrix whose diagonal contains the instrument error variances.

It is shown in Hodyss and Nichols (2015) that a portion of $\mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T$ in (2.10) is the representation error and takes the following form for this problem:

$$\bar{\mathbf{R}}_{rep} = \mathbf{H}_H \mathbf{E}_H \Theta \mathbf{E}_H^T \mathbf{H}_H^T \quad (2.13)$$

where Θ is a diagonal matrix with the value of the diagonal of Θ satisfying:

$$\Theta_i = \begin{cases} 0, & i = 1, \dots, M \\ \Gamma_i, & i = M + 1, \dots, N \end{cases} \quad (2.14)$$

Equation (2.13) clearly shows that the representation error is simply the portion of the high resolution spectrum that is missing from the low-resolution states. Note that for $M = N$, and therefore no truncation, the elements of Θ vanish and there is no representation error. This point is important because it shows that the climatological model error on the resolved scales (\mathbf{D}) is irrelevant to both the existence of representation error and to the structure of the representation error.

An important quantity in the theory of Hodyss and Nichols (2015) was the state-dependent bias of the forecast model's estimate of the observation, viz.:

$$\mathbf{y}_L(\mathbf{x}_L) = \int_{-\infty}^{\infty} \mathbf{y} \rho(\mathbf{y}|\mathbf{x}_L) d\mathbf{y} = \mathbf{H}_L \mathbf{x}_L + \mathbf{b} \quad (2.15)$$

where $\rho(\mathbf{y}|\mathbf{x}_L)$ is the observation likelihood but conditioned on the low-resolution forecast states,

$$\mathbf{b} = \left[\mathbf{H}_H \mathbf{S}^\dagger - \mathbf{H}_L \right] \mathbf{x}_L + \mathbf{H}_H \left[\bar{\mathbf{x}}_H - \mathbf{S}^\dagger \bar{\mathbf{x}}_L \right] \quad (2.16)$$

and the superscript \dagger in (2.16) denotes the Moore-Penrose pseudo-inverse (Golub and Van Loan 1996). The first term in (2.16) is interesting because it corresponds to the mismatch between the forecast model's estimate of the true observation operator $\mathbf{H}_H \mathbf{S}^\dagger$ and the observation operator we are actually using \mathbf{H}_L . We emphasize here that this mismatch is *not* one in which we are implying that \mathbf{H}_L is incorrect in the sense that if, for example, we had a point measurement that there would be some form of inaccuracy in the interpolation to the observation location. Indeed, even in this case of a point measurement, in which we are assuming we have a perfect interpolation, the mismatch implied by (2.16) is between the statistically derived observation operator $\mathbf{H}_H \mathbf{S}^\dagger$, which now corresponds to more than just an interpolation, and the operator, \mathbf{H}_L .

The Eq. (2.16) suggests that if we define \mathbf{H}_L such that,

$$\mathbf{H}_L \equiv \mathbf{H}_H \mathbf{S}^\dagger \quad (2.17)$$

we may remove this state-dependent portion of the bias term, which subsequently renders $\mathbf{P}_{LH} = \mathbf{0}$. This implies that in this data assimilation system the observation operator does *not* simply map the truth to the observation, but rather it maps the forecast to the observation and, because the forecast is in a different portion of state space than the truth, this requires the matrix operator, $\mathbf{H}_H \mathbf{S}^\dagger$ rather than \mathbf{H}_H .

By employing (2.17) Hodyss and Nichols (2015) showed that the gain matrix, (2.8), is now:

$$\mathbf{G} = \mathbf{P}_L \mathbf{S}^{\dagger T} \mathbf{H}_H^T \left[\mathbf{H}_H \mathbf{S}^\dagger \mathbf{P}_L \mathbf{S}^{\dagger T} \mathbf{H}_H^T + \bar{\mathbf{R}}_L^* \right]^{-1} \quad (2.18)$$

The denominator of the gain, (2.18), has a modified covariance matrix within it. Note that this covariance matrix is

$$\mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T = \mathbf{H}_H \mathbf{S}^\dagger \mathbf{P}_L \mathbf{S}^{\dagger T} \mathbf{H}_H^T = \mathbf{H}_H \mathbf{E}_H \begin{bmatrix} \mathbf{\Gamma}_M & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{E}_H^T \mathbf{H}_H^T \quad (2.19)$$

where Γ_M denotes the first M eigenvalues of Γ and $\mathbf{0}$ is the zero matrix. Equation (2.19) shows that the choice (2.17) results in $\mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T$ being correct up to the resolution of the forecast model in the sense that the climatological model error denoted by \mathbf{D} has been removed.

3 Estimation of Representation Error

The estimation of observation error covariance matrices is commonly performed using two different methods. One of the methods is that of Desroziers et al. (2005) and the other is that of HL. To our knowledge their behavior when applied to forecast models with a reduced state-vector as compared to the true state has not been examined.

3.1 Desroziers' Method

To focus our discussion of representation error, here we will explicitly extend the method of Desroziers' et al. (2005) to the case where the model state vector is shorter than the true system state vector. The mathematical setup will closely parallel Sect. 2. We will assume that the true state-vector has a length N and the low-resolution vector has a length M , such that $M < N$. We also assume that the observations observe this high-resolution true state. In principle, there exists a data assimilation algorithm for the high-resolution state:

$$\bar{\mathbf{x}}_H^a = \bar{\mathbf{x}}_H + \mathbf{G}_H \mathbf{v}_H \quad (3.1)$$

where

$$\mathbf{v}_H = \mathbf{y} - \mathbf{H}_H \bar{\mathbf{x}}_H \quad (3.2)$$

$$\mathbf{G}_H = \mathbf{P}_H \mathbf{H}_H^T [\mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T + \mathbf{R}_{ins}]^{-1} \quad (3.3)$$

The data assimilation algorithm (3.1) is not possible in numerical weather prediction because of computational constraints. Hence, the data assimilation is performed at a lower resolution:

$$\bar{\mathbf{x}}_L^a = \bar{\mathbf{x}}_L + \mathbf{G}_L [\mathbf{v}_L - \langle \mathbf{v}_L \rangle] \quad (3.4)$$

with

$$\mathbf{G}_L = \mathbf{P}_L \mathbf{H}_L^T [\mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T + \mathbf{R}_L]^{-1} \quad (3.5)$$

Note that in (3.4) and (3.5) we have not assumed the optimal data assimilation algorithm in (2.8) or (2.18); we have assumed a data assimilation algorithm of standard form in order to show that standard observation error estimation techniques do not work unless the data assimilation algorithm they are using is optimal for the reduced resolution data assimilation problem defined in Sect. 2.

In any event, we know that an accurate variance in the denominator of the Kalman gain \mathbf{G}_L must be equal to the innovation variance,

$$\langle [\mathbf{v}_L - \langle \mathbf{v}_L \rangle][\mathbf{v}_L - \langle \mathbf{v}_L \rangle]^T \rangle = \langle \mathbf{v}_H \mathbf{v}_H^T \rangle = \mathbf{R}_{ins} + \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T \quad (3.6)$$

Therefore, we must allow the matrix \mathbf{R}_L to be defined by,

$$\mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T + \mathbf{R}_L = \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T + \mathbf{R}_{ins} \quad (3.7)$$

We show next that the Desroziers' technique is constrained to deliver a result consistent with (3.7).

The Desroziers' estimate for the observation error covariance matrix, \mathbf{R}_D , is:

$$\mathbf{R}_D = \langle [\mathbf{y} - \mathbf{H}_L \bar{\mathbf{x}}_L^a - \langle \mathbf{v}_L^a \rangle][\mathbf{v}_L - \langle \mathbf{v}_L \rangle]^T \rangle \quad (3.8)$$

Note that we must de-bias both the innovation as well as the difference of the analysis from the observations. The analysis innovation bias is

$$\langle \mathbf{v}_L^a \rangle = \langle \mathbf{y} - \mathbf{H}_L \bar{\mathbf{x}}_L^a \rangle = \langle \mathbf{y} \rangle - \mathbf{H}_L \bar{\mathbf{x}}_L = \langle \mathbf{v}_L \rangle \quad (3.9)$$

Now we can rewrite (3.8) as,

$$\mathbf{R}_D = [\mathbf{I} - \mathbf{H}_L \mathbf{G}_L] \langle [\mathbf{v}_L - \langle \mathbf{v}_L \rangle][\mathbf{v}_L - \langle \mathbf{v}_L \rangle]^T \rangle \quad (3.10)$$

By substituting (3.6) we obtain,

$$\mathbf{R}_D = [\mathbf{I} - \mathbf{H}_L \mathbf{G}_L] [\mathbf{R}_{ins} + \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T] \quad (3.11)$$

which may be written as

$$\mathbf{R}_D = \mathbf{R}_{ins} + \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T \quad (3.12)$$

Equation (3.12) is in agreement with (3.7). Furthermore, Eq. (3.12) shows that the Desroziers' estimate calculates the observation covariance matrix as the instrument error variance plus the difference between the high-resolution and low-resolution covariance matrices as mapped into observation space.

Therefore, we need to understand what $\mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T$ means. To this end, we will re-write the relationship (2.4) between the high and low-resolution states in the following way:

$$\mathbf{x}_L = \mathbf{S}_{sc} \mathbf{x}_H + (\mathbf{S} - \mathbf{S}_{sc}) \mathbf{x}_H \quad (3.13)$$

where \mathbf{S}_{sc} is the smoother from a “spectral-cut”, viz.

$$\mathbf{S}_{sc} = \mathbf{E}_L [\mathbf{T} \quad \mathbf{0}] \mathbf{E}_H^T \quad (3.14)$$

We employ a “spectral-cut” operator in (3.13) because we desire to develop a term below that is solely induced by the implicit spectral cut from our truncated forecast model. Equation (3.13) can be thought of as having in its first term a spectral cut of the high-resolution state to the low-resolution state and this term corresponds to any missing small-scale phenomena. In the second term of (3.13) any resolved scale model error between the high and low-resolution models is represented. If there is no resolved scale model error, and therefore the second term is missing, then the results below will reduce to those of Liu and Rabier (2002).

Equation (3.13) implies that the difference in the high and low resolution covariance matrices is

$$\begin{aligned} \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T &= \mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{S}_{sc} \mathbf{P}_H \mathbf{S}_{sc}^T \mathbf{H}_L^T - \mathbf{H}_L \mathbf{S}_{sc} \mathbf{P}_H (\mathbf{S} - \mathbf{S}_{sc})^T \mathbf{H}_L^T \\ &\quad - \mathbf{H}_L (\mathbf{S} - \mathbf{S}_{sc}) \mathbf{P}_H \mathbf{S}_{sc}^T \mathbf{H}_L^T - \mathbf{H}_L (\mathbf{S} - \mathbf{S}_{sc}) \mathbf{P}_H (\mathbf{S} - \mathbf{S}_{sc})^T \mathbf{H}_L^T \end{aligned} \quad (3.15)$$

We will assume that the low-resolution observation operator is:

$$\mathbf{H}_L = \mathbf{H}_H \mathbf{S}_{sc}^\dagger \quad (3.16)$$

Therefore, we have not assumed (2.17). The assumption (3.16) is the optimal observation operator in the case where the low-resolution forecast model is exact at the resolved scales. This assumption is being used to create a specific term that is the covariance matrix of the unresolved scales to be associated with representation error.

Note the following properties:

$$\mathbf{S}_{sc}^\dagger \mathbf{S}_{sc} = \mathbf{E}_H [\mathbf{T}^{-1} \quad \mathbf{0}]^T \mathbf{E}_L^T \mathbf{E}_L [\mathbf{T} \quad \mathbf{0}] \mathbf{E}_H^T = \mathbf{E}_H \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{E}_H^T \quad (3.17)$$

$$\mathbf{S}_{sc}^\dagger \mathbf{S} = \mathbf{E}_H [\mathbf{T}^{-1} \quad \mathbf{0}]^T \mathbf{E}_L^T \mathbf{E}_L [\mathbf{D}^{1/2} \mathbf{T} \quad \mathbf{0}] \mathbf{E}_H^T = \mathbf{E}_H \begin{bmatrix} \mathbf{D}^{1/2} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{E}_H^T \quad (3.18)$$

Use (3.16) through (3.18) in (3.15) to obtain

$$\mathbf{H}_H \mathbf{P}_H \mathbf{H}_H^T - \mathbf{H}_L \mathbf{P}_L \mathbf{H}_L^T = \underbrace{\mathbf{H}_H \mathbf{E}_H \mathbf{\Gamma}_{N-M} \mathbf{E}_H^T \mathbf{H}_H^T}_{\text{Representation Error}} + \underbrace{\mathbf{H}_H \mathbf{E}_H [\mathbf{D} - \mathbf{I}] \mathbf{\Gamma}_M \mathbf{E}_H^T \mathbf{H}_H^T}_{\text{Model Error on Resolved Scales}} \quad (3.19)$$

where

$$\boldsymbol{\Gamma}_{N-M} = \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \boldsymbol{\Gamma} \begin{bmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} = \boldsymbol{\Theta} \quad (3.20)$$

which are simply the eigenvalues of \mathbf{P}_H that are not in \mathbf{P}_L and is precisely equivalent to (2.13). Note in (2.5) that if there is no model error in the low-resolution forecast model at the resolved scales then $\mathbf{D} = \mathbf{I}$, and therefore the only process in (3.19) is the representation error. Similarly, if we replace (3.5) in (3.11) with (2.18) we will obtain,

$$\mathbf{R}_D = \mathbf{R}_{ins} + \underbrace{\mathbf{H}_H \mathbf{E}_H \boldsymbol{\Gamma}_{N-M} \mathbf{E}_H^T \mathbf{H}_H^T}_{\text{Representation Error}} \quad (3.21)$$

which is only a function of the instrument error and the representation error. Therefore, we have shown that the Desroziers' method is only a self-consistent estimator of the representation error when (1) the forecast model is perfect at the resolved scales or (2) the data assimilation system is of the form (2.18).

3.2 HL Method

The central assumption of the HL method is that the correlated portion of the innovations as a function of distance between observations is a property of the prior covariance matrix; all other contributions to the innovation variance is due to the observations. Hence, we need to co-vary innovations at different distances. To this end we denote the i th innovation as

$$v_i = \mathbf{y}_i - \mathbf{h}_L^i \bar{\mathbf{x}}_L \quad (3.22)$$

where \mathbf{h}_L^i is the row of \mathbf{H}_L that extracts the low-resolution forecast model estimate of the i th observation.

Therefore, the covariance of the innovation for all observations separated a distance, d , apart is

$$\begin{aligned} \langle v_i v_j \rangle_d &= \langle (\mathbf{y}_i - \mathbf{h}_L^i \bar{\mathbf{x}}_L - \langle \mathbf{v}_L \rangle_i) (\mathbf{y}_j - \mathbf{h}_L^j \bar{\mathbf{x}}_L - \langle \mathbf{v}_L \rangle_j) \rangle_d \\ &= \langle (\mathbf{y}_i - \mathbf{h}_H^i \bar{\mathbf{x}}_H) (\mathbf{y}_j - \mathbf{h}_H^j \bar{\mathbf{x}}_H) \rangle_d \end{aligned} \quad (3.23)$$

where the different subscripts i and j denote different innovations, but all separated a distance, d , apart and where \mathbf{h}_H^i is the row of \mathbf{H}_H that extracts the high-resolution forecast model estimate of the i th observation.

Note that under the assumptions of (1) all innovations are from a single observational instrument and (2) uncorrelated instrument errors, (3.23) reduces simply to

$$\langle v_i v_j \rangle_d = \begin{cases} r_{ins} + (\mathbf{h}_H^i \mathbf{P}_H \mathbf{h}_H^i)_d, & i=j, d=0 \\ (\mathbf{h}_H^i \mathbf{P}_H \mathbf{h}_H^j)_d, & i \neq j, d > 0 \end{cases} \quad (3.24)$$

where r_{ins} is the instrument error variance for the particular instrument in the innovations (3.22) and the parenthesis around $\mathbf{h}_H^i \mathbf{P}_H \mathbf{h}_H^j$ is meant to refer to the prior covariance of the observed variable at a separation distance, d .

In contrast to the Desroziers' technique, the HL method obtains an estimate of the *high-resolution* covariance matrix for d values greater than 0. Note however that this is true if one can actually calculate the expectation in (3.23) at precisely the separation distance d . In practice, one must always use bins of a non-zero width to define the samples that are used to evaluate the expectation in (3.23). The width of these bins effect a smoothing upon the resulting prior covariance estimate. Bin widths used in global numerical weather prediction typically range from 100 to 200 km wide. Such a bin width acts as a spatial filter, smoothing the resulting covariance structure and reduces the value of the prior covariance estimate from that of the high-resolution covariance matrix to that more similar to the low-resolution covariance matrix.

The impact of this smoothing is critical because if the bin widths effected a spatial smoothing at approximately the same scale as the low-resolution model resolution then one would obtain a smoothing of (3.24) for $d > 0$:

$$\langle v_i v_j \rangle_d = \begin{cases} r_{ins} + (\mathbf{h}_H^i \mathbf{P}_H \mathbf{h}_H^i)_d, & i=j, d=0 \\ (\mathbf{h}_L^i \mathbf{P}_L \mathbf{h}_L^j)_d, & i \neq j, d > 0 \end{cases} \quad (3.25)$$

To back out the HL estimate of the observation error one would subtract the zero separation innovation variance from the $d > 0$ prior covariance structure extrapolated to zero separation, viz.

$$R_{HL} = \langle v_i v_j \rangle_{d=0} - (\mathbf{h}_L^i \mathbf{P}_L \mathbf{h}_L^j)_{d \rightarrow 0} = r_{ins} + (\mathbf{h}_H^i \mathbf{P}_H \mathbf{h}_H^i)_{d=0} - (\mathbf{h}_L^i \mathbf{P}_L \mathbf{h}_L^i)_{d=0} \quad (3.26)$$

which is clearly the same as the Desroziers' estimate (3.12). Therefore, the comparison of the HL method and Desroziers' method can be considered to be a self-consistency check. However, as shown in Eq. (3.19) they both have the property that they contain a portion of the resolved scale error variance in their estimates.

4 Issues with Incorrectly Specified Prior Error Variances

This section will illustrate another way in which observation error estimation procedures can erroneously place a portion of the background error variance into the observation error estimate. It is well known (Ménard 2015; Waller 2015) that the Desroziers diagnostic will overestimate (underestimate) the observation error variance if the background error variances are underestimated (overestimated). In the following, we turn our attention to the case where the prior error variances are defined (or partly defined) by flow dependent ensemble variances. As in the previous section, we will use a data assimilation algorithm here that is of standard form, and therefore does not include the modifications suggested in (2.8) and (2.18).

It has been well established that the Desroziers method is dependent on the prescribed error covariance matrices. Here, we address how such dependence impacts the results when the method is binned by ensemble variance. For this purpose, we implement a simplified framework as follows:

1. The true state is drawn from a Gaussian distribution with prescribed mean and variance $\mathcal{N}(\bar{x}^t, \sigma_t)$. In what follows, we define $\bar{x}^t = 2.22$ and $\sigma_t = 0.62$.
2. The forecast is modeled as the true state plus a forecast error, $x^f = x^t + \varepsilon^f$, where $\varepsilon^f \sim \mathcal{N}(0, \sigma_f^2)$ and σ_f^2 is defined by an inverse gamma distribution. This allows for a simple model that qualitatively represents spatial and temporal changes to the forecast error variances. In what follows, σ_f^2 is a random draw from an inverse gamma distributions with mean $\bar{\sigma}_f^2 = 2.22$ and variance 0.62; the corresponding inverse-gamma shape and scale parameters are given by $\alpha = 10$ and $\beta = 20$.
3. Observations are created following $y^o = x^t + \varepsilon^o$, where $\varepsilon^o \sim \mathcal{N}(0, \sigma_o^2)$. In what follows, we set $\sigma_o^2 = 2.5$.
4. Ensemble variances are modeled following $s^2 = (a\sigma_f^2)\eta$, where a specifies the over or under dispersion and $\eta \sim \frac{\chi^2(K)}{(K-1)}$ is drawn from a chi-squared distribution to account for sampling errors due to an ensemble size K .
5. Analyses are defined following $x^a = \frac{\sigma_o^2}{\sigma_o^2 + \sigma_s^2} x^f + \frac{\sigma_s^2}{\sigma_o^2 + \sigma_s^2} y^o$.

We set our experiment to run for 10,000 independent trials and then calculate the Desroziers diagnostic in 5 equally populated bins based on variance. In the first experiment we use $K = 1000$ ensemble members and set $a = 1$. The result is shown in Fig. 1a, the Desroziers result (blue) line is a random fluctuation about the true observation error variance (red line). In Fig. 1b we show results based on an under dispersive ensemble with $a = 0.2$, here we do see a positive slope, which is a case in which dispersion errors could mimic the appearance of representation error. For the over dispersive case with $a = 2$ shown in Fig. 1c we see the opposite result, a negative slope. Figure 1d–f repeat a–c with variance of σ_f^2 set at 0.006 and the mean

$\overline{\sigma_f^2} = 0.22$ ($\alpha = 10$, $\beta = 2$). As the distribution from which the forecast error variance is drawn tightens, the day to day variations in ensemble variance lessen and our result converges to what is typically seen in the literature, a flat over (under) estimation when the background error variance is under (over) estimated, as in Fig. 1e (Fig. 1f). Now we repeat (a–c) with only 10 ensemble members. The dispersion errors due to too few members have created a negative slope in Fig. 1g. For the under dispersive ($\alpha = 0.2$) case (Fig. 1h) we now see a higher magnitude over estimation but the slope has decreased. For the over dispersive ($\alpha = 2$) case (Fig. 1i) we now see an increased slope.

This analysis indicates that errors in ensemble dispersion, as well as sampling errors, can lead to an estimate of observation error that is a function of ensemble variance, even in cases where the observation error variance has no flow dependence.

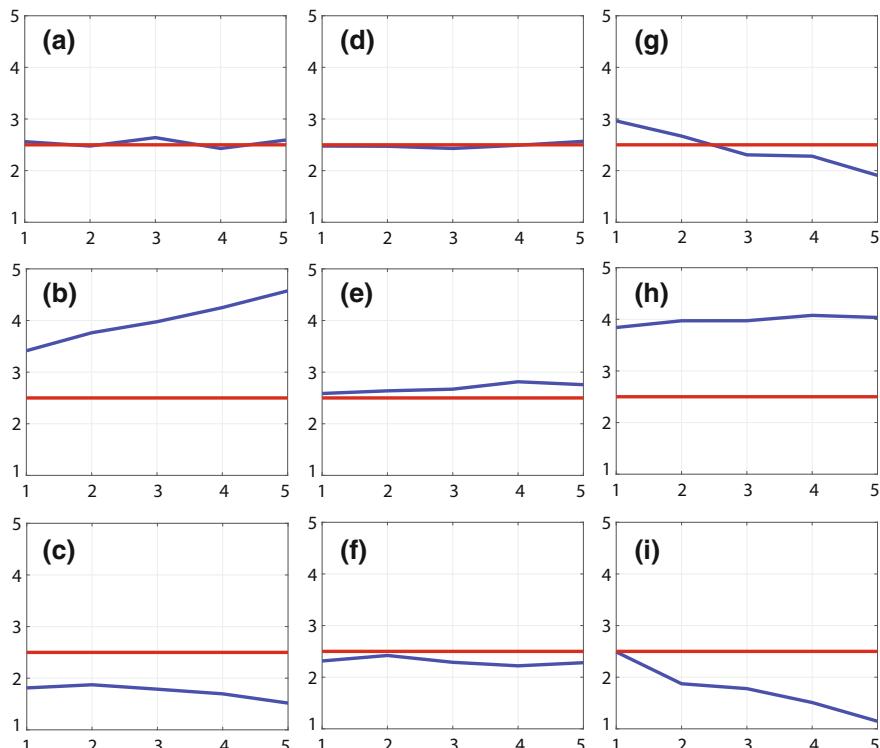


Fig. 1 The Desrozier diagnostic (blue) and the prescribed observation error variances (red) versus ensemble variance (x-axis) for **a** a perfectly dispersive ensemble ($\alpha = 1$) **b** an under dispersive ensemble $\alpha = 0.2$ and **c** an over dispersive ensemble ($\alpha = 2$). **d–f** repeat (a–e) with reduced day to day variation of forecast error variance. **g–i** repeat (a–c) with reduced ensemble size

5 Summary and Conclusions

In this chapter we have reviewed the theory of data assimilation when it includes the effect of misrepresentation of the number of degrees of freedom in the true system. When the forecast model has fewer degrees of freedom, or equivalently, when the data assimilation is chosen to be performed at a reduced resolution than the true system that is being observed, the data assimilation algorithm must be modified to address this deficiency. This modification includes a change to the observation error covariance matrix to account for this error in representation of the true state, but as discussed here it also must include a change to the “background” portion of the covariance matrix.

We have shown that observation error estimation procedures must be modified to account for this change in the definition of the “optimal” data assimilation system. We have shown that the method of Desroziers’ et al. (2005) does not deliver the correct estimate unless (1) the data assimilation method is modified to the form described in Sect. 2, which includes modification to the observation covariance matrix as well as to the numerator of the gain or (2) the truncated forecast model is perfect at all of the resolved scales. Note however that the method of HL will not deliver the correct estimate of the observation covariance matrix unless (1) the bin widths in its estimation procedure have infinitesimal width or (2) the truncated forecast model is perfect at all of the resolved scales. When these conditions are not met then standard estimation techniques will include background error covariance structure in the observation error covariance estimate. Lastly, we showed that the misspecification of the background error variance in a traditional data assimilation method will also lead to an estimate of the observation error variance that includes the background variance.

These results show that fruitful performance improvements in the quality of the analysis are likely obtainable by carefully constructing the data assimilation system to properly account for the misrepresentation of the number of degrees of freedom in the true system. This will require research into applying the data assimilation algorithm of Sect. 2 to the computationally intensive real-world, as well as accounting for the issues in Sects. 3 and 4. Research in these directions is already underway.

References

- Bauer P, Bormann N (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: Methods and application to ATOVS data. *Q J R Meteorol Soc* 136:1036–1050
- Bormann N, Saariene S, Kelly G, Thepaut J (2002) The spatial structure of observation errors in atmospheric motion vectors from geostationary satellite data. *Mon Weather Rev* 131:706–718

- Bormann N, Collard A, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: Application to AIRS and IASI data. *Q J R Meteorol Soc* 136:1051–1063
- Desroziers G, Berre L, Chapnik B, Poli P (2005) Diagnosis of observation, background and analysis-error statistics in observation space. *Q J R Meteorol Soc* 126:3385–3396
- Forget G, Wunsch C (2007) Estimated global hydrographic variability. *J Phys Oceanogr* 563 (37):1997–2008
- Frehlich R (2006) Adaptive data assimilation including the effect of spatial variations in observation error. *Q J R Meteorol Soc* 132:1225–1257
- Frehlich R (2008) Atmospheric turbulence component of the innovation covariance. *Q. J. Roy. Meteorol. Soc.* 134:931–940
- Frehlich R (2011) The definition of ‘truth’ for numerical weather prediction error statistics. *Q J R Meteorol Soc* 137:84–98
- Golub, G. and C. F. Van Loan, 1996: Matrix Computations. The Johns Hopkins University, Press, 687 pgs
- Hollingsworth A, Lönnberg P (1986) The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus A* 38:111–136
- Hodyss D, Nichols N (2015) The error of representation: Basic understanding. *Tellus A* 67:24 822–24 839
- Janjic T, Cohn SE (2006) Treatment of observation error due to unresolved scales in atmospheric data assimilation. *Mon Weather Rev* 134:2900–2915
- Lorenz AC (1986) Analysis methods for numerical weather prediction. *Q J R Meteorol Soc* 112:1177–1194
- Liu Z-Q, Rabier F (2002) The interaction between model resolution, observation resolution and observation density in data assimilation: a one-dimensional study. *Q J R Meteorol Soc* 128:1367–1386
- Ménard R (2015) Error covariance estimation methods based on analysis residuals: theoretical foundation and convergence properties derived from simplified observation networks. *Q J R Meteorol Soc.* doi:[10.1002/qj.2650](https://doi.org/10.1002/qj.2650)
- Mitchell HL, Daley R (1997a) Discretization error and signal/error correlation in atmospheric data assimilation (I). All scales resolved. *Tellus A* 49:32–53
- Mitchell HL, Daley R (1997b) Discretization error and signal/error correlation in atmospheric data assimilation (II). The effect of unresolved scales. *Tellus A* 49:54–73
- Oke PR, Sakov P (2008) Representation error of oceanic observations for data assimilation. *J Atmos Oceanic Technol* 25:1004–1017
- Richman JG, Miller RN, Spitz YH (2005) Error estimates for assimilation of satellite sea surface temperature data I ocean climate models. *J Geophys Res* 32:L18608. doi:[10.1029/2005GL023591](https://doi.org/10.1029/2005GL023591)
- Rutherford I (1972) Data assimilation by statistical interpolation of forecast error fields. *J Atmos Sci* 29:809–815
- Stewart L, Dance S, Nichols N (2012) Data assimilation with correlated observation errors: analysis accuracy with approximate error covariance matrices. Technical report, University of Reading. Department of Mathematics and Statistics Preprint MPS-2012-17. <http://www.reading.ac.uk/maths-and-stats/research/maths-preprints.aspx>
- Stewart L, Dance SL, Nichols NK, Eyre JR, Cameron J (2014) Estimating interchannel observation-error correlations for IASI radiance data in the Met Office system. *Q J R Meteorol Soc* 140:1236–1244. doi:[10.1002/qj.2211](https://doi.org/10.1002/qj.2211)
- Stewart L (2010) Correlated observation errors in data assimilation. Ph.D. thesis, University of Reading
- Stewart L, Cameron J, Dance S, English S, Eyre JR, Nichols N (2009) Observation error correlations in IASI radiance data. Technical report, University of Reading. Mathematics reports series. www.reading.ac.uk/web/FILES/maths/obserrorIASIradiance.pdf

- Waller JA, Dance SL, Lawless AS, Nichols NK, Eyre JR (2014) Representivity error for temperature and humidity using the Met Office high-resolution model. *Q J R Meteorol Soc* 140:1189–1197
- Waller JA, Dance SL, Nichols NK (2015) Theoretical insight into diagnosing observation error correlations using observation-minus-background and observation-minus-analysis statistics. *Q J R Meteorol Soc*. doi:[10.1002/qj.2661](https://doi.org/10.1002/qj.2661)
- Waller JA, Dance SL, Nichols NK, Simonin D, Ballard SP (2015) Diagnosing observation error correlations for Doppler radar radial winds in the Met Office UKV model observation-minus-background and observation-minus-analysis statistics. http://www.reading.ac.uk/web/FILES/math/Preprint_MPS_15-18_Waller.pdf
- Weston P (2011) Progress towards the implementation of correlated observation errors in 4D-VAR. Technical report, Met Office, UK. Forecasting Research Technical Report 560
- Weston P, Bell W, Eyre JR (2014) Accounting for correlated error in the assimilation of high-resolution sounder data. *Q J R Meteorol Soc*. doi:[10.1002/qj.2306](https://doi.org/10.1002/qj.2306)

Soil Moisture Data Assimilation

Viviana Maggioni and Paul R. Houser

Abstract Soil moisture plays an important role in the global to regional water and energy cycle, as it controls the partitioning of water and radiation into runoff, evaporation and infiltration at the land-atmosphere interface. Soil moisture information can be obtained through in situ observations, land surface models and remote-sensing retrievals. This chapter reviews the capability of land data assimilation systems to merge observations (either in situ or remotely sensed) with the spatially and temporally complete information from land surface models, in order to provide an improved dynamic representation of surface and root-zone soil moisture. Among the different data assimilation techniques, the ensemble Kalman Filter and variational methods are becoming the methods of choice for large-scale soil moisture data assimilation. The improvement in soil moisture estimation via data assimilation largely depends on the quality of the land surface model meteorological forcings. Since precipitation is the major driver for soil moisture, uncertainty in precipitation affects the efficiency of assimilating soil moisture observations most profoundly. Data assimilation systems have also been demonstrated to be extremely valuable for downscaling coarser resolution satellite brightness temperature observations in order to produce higher resolution soil moisture estimates. Skill metrics for evaluating the improvement in soil moisture estimates via land data assimilation is also maturing. Besides biases and root mean square errors, common metrics to evaluate data assimilation are the anomaly correlation coefficient, the exceedance and uncertainty ratios and rank histograms.

1 Background

Soil moisture is the quantity of water contained in the unsaturated soil on a volumetric or gravimetric basis (Hillel 1998). Surface and root zone soil moisture are key variables of the water and energy cycle, as they represent the land storage for

V. Maggioni · P.R. Houser (✉)
George Mason University, Fairfax, VA 22030-4444, USA
e-mail: phouser@gmu.edu

water and energy, effectively controlling the balance between sensible and latent heat flux at the land-atmosphere interface (Fig. 1). Thus, soil water content may impact atmospheric processes, such as cloud coverage and rainfall and hydrological processes, such as runoff and plant transpiration (Betts and Ball 1998). Soil moisture-temperature and soil moisture-precipitation feedbacks have the potential to significantly impact weather dynamics and climate-change projections from the local to the regional and global scale (Seneviratne et al. 2010). By constraining plant transpiration and photosynthesis, soil moisture also plays an important role in biogeochemical cycles (e.g. carbon and nitrogen cycles) and, therefore, in fields like agriculture and ecology.

A realistic characterization of surface and root-zone soil moisture and its associated uncertainty can lead to improvements in several areas: from weather and climate forecast to the mitigation of extreme events, like floods and droughts, water budgeting in agriculture, and water resources management. For instance, soil moisture is commonly used for deriving flood-warning schemes that are based on precipitation thresholds (Martina et al. 2006; Carpenter et al. 1999). In order to issue flood warnings, quantitative information about the soil water content is adopted for selecting the most appropriate rainfall-runoff threshold curve to be used together with the estimated rainfall volume. Information on soil moisture can be assessed through three main approaches: (i) in situ measurements (Walker et al. 2004); (ii) remotely sensed observations from low-frequency active and passive microwave sensors (Schmugge et al. 2002); and (iii) integration of a land surface model forced with meteorological data (Peters-Lidard et al. 2007).

There are direct and indirect methods to measure soil moisture. In direct approaches, soil samples are collected and their moisture content is evaluated in the laboratory. Making a spatially-representative measurement may require collecting a large number of samples sufficient to include soil, terrain, micro-climate, and vegetation variability. The most common direct approach is the *gravimetric*

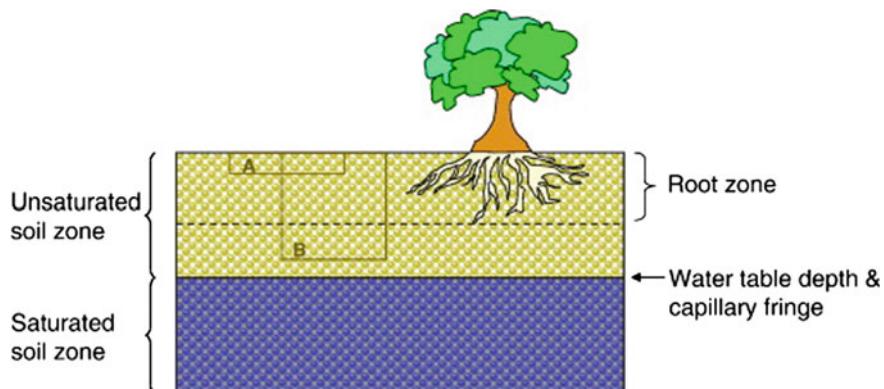


Fig. 1 The saturated and unsaturated soil zones. A and B denote two distinct soil moisture volumes. From Seneviratne et al. (2010)

method. Soil samples of about 50 g are removed from the field with tools like shovels and augers and placed in leak-proof, seamless, pre-weighed containers. The samples and container are then weighed before and after drying and the difference is the soil water content estimate. Volumetric soil moisture measurements require a known volume of soil to be collected. Indirect approaches measure soil properties that can be related back to soil moisture (e.g., through dielectric constant, variations of gravity field, or soil suction), as the capacity of a soil to retain water is a function of its texture and structure. These methods are valuable especially because information on soil moisture can be collected at the same location several times without disturbing the soil water system. However, *in situ* measurements, both direct and indirect, are very localized and limited in spatial and temporal coverage (Robock et al. 2000; Robinson et al. 2008; Dorigo et al. 2011).

Surface soil moisture can be estimated remotely using microwave radiation measurements (Jackson 1993; Njoku et al. 2003). Active and passive L-band microwave remote sensing has been well established through many ground and aircraft studies. These paved the way for the ESA Soil Moisture and Ocean Salinity (SMOS) and the NASA Soil Moisture Active Passive (SMAP) mission, launched in 2009 and 2015 respectively. However, L-band microwave remote sensing provides mostly surface soil moisture information (top 5 cm of the soil column), where most agricultural, hydrological and meteorological applications require root zone information. Remote sensing soil moisture retrievals are also affected by errors due to (i) limitations in the sensor sampling (i.e., the coverage is not spatially and temporally continuous), (ii) effects of land cover heterogeneity within the pixel, (iii) difficulties in defining the physical processes that relate brightness temperature to soil moisture, and (iv) uncertainty in the retrieval algorithm parameters. Retrievals are also limited in areas where the fraction of open water is significant, and where the soil is frozen or densely vegetated. Therefore, the usefulness of the remotely-sensed soil moisture products alone is limited. However, when combined with land surface models, soil moisture and its variations over time, space and with depth in the soil column can be estimated.

Land surface models (LSMs) integrate atmospheric forcings to produce continuous and spatially distributed soil moisture fields. Current LSMs have mostly had their roots in coupled weather and climate models, with the intent of partitioning energy and water at the land surface to provide a lower boundary condition to the atmosphere. Accordingly, the spatial resolution of current LSMs has largely been dictated by the spatial resolutions to which global weather and climate models are constrained by computational limitations: currently, at best, \sim 100 km for climate models and \sim 20 km for weather models (somewhat higher resolutions are used by regional models). Much higher resolutions, which are referred to as hyper-resolution (100 m to 1 km globally), will soon be feasible and will provide more detailed information about the storage, movement, and quality of carbon and water at and near the land surface (Wood et al. 2011). LSM predictions are affected by uncertainties in the meteorological forcing variables, model parameters, and model formulations (Reichle et al. 2004).

In order to remedy for the scarcity of in situ measurements and the uncertainties in both remotely-sensed retrievals and LSM estimates, land data assimilation systems combine the complementary information from observations and the spatially and temporally complete information given by LSMs into a superior estimate of soil moisture (e.g., Reichle and Koster 2005; Li et al. 2010).

2 Land Data Assimilation Systems for Soil Moisture Estimation

Land data assimilation systems (LDASs) merge remotely-sensed and/or ground-based observations with the spatially and temporally complete information provided by land surface models to generate a superior product of soil moisture (Houser et al. 1998; Hoeben and Troch 2000; Walker and Houser 2001; Reichle et al. 2002a, b; Margulis et al. 2002; Reichle and Koster 2003; Crow and Wood 2003; Seuffert et al. 2003; Crow and van Loon 2006; De Lannoy et al. 2007; Dunne and Entekhabi 2006; Pan and Wood 2006; Zhou et al. 2006; Parajka et al. 2006; Reichle et al. 2008b; Drusch et al. 2009). The LDAS captures the key land surface processes, such as the vertical transfer of water between the surface and root zone reservoirs, and interpolates the observations in time and in space. Specifically, model predictions of soil moisture are corrected with a stochastic filtering technique towards the observations, by accounting for the relative observation and prediction uncertainties. The corrected model state is then used to make improved forecasts, but can also be used to diagnose and correct model deficiencies.

The first LDASs assimilating soil moisture were mainly tested using measurements collected during field campaigns (and therefore limited in time and space) and synthetic satellite retrievals, as global observations were not available yet. However, in the recent past, a number of satellite-based soil moisture products have become available: the Advanced Microwave Scanning Radiometer for the Earth Observing System (AMSR-E; Njoku 2011; Owe et al. 2008), the Tropical Rainfall Measuring Mission (TRMM) Microwave Imager (TMI; Gao et al. 2006; Owe et al. 2008), Windsat (Li et al. 2007), the historic Scanning Multichannel Microwave Radiometer (SMMR; Owe et al. 2008), the European Remote Sensing satellites (Wagner et al. 2007), the Advanced Scatterometer (ASCAT; Bartalis et al. 2008), the ESA SMOS (Kerr et al. 2010), the NASA Aquarius mission (Lagerloef et al. 2008), and the NASA SMAP mission (Entekhabi et al. 2010a). Satellite soil moisture products are based on C- and X-band microwave observations with an effective sensing depth of roughly 1 cm, and L-band microwave sensors that measure moisture in the top 5 cm of the soil.

Constraining LSMs with satellite soil moisture retrievals using data assimilation techniques has been demonstrated to be an effective way to estimate soil moisture dynamics (Reichle and Koster 2005; Drusch 2007; Reichle et al. 2007). The ensemble Kalman Filter (EnKF) is becoming a method of choice for large-scale soil

moisture data assimilation systems (Evensen 1994; Garcia-Pintado et al. 2013). The EnKF technique is a Monte Carlo-based filter that requires a number of model runs to represent the forecast uncertainty. Perturbations are applied to the model forcing and state variables to obtain an ensemble of state fields that reflects prediction uncertainty. The major advantages of the EnKF technique are its flexibility in treating errors in model equations and parameters, its independence from the model code, and its suitability for nonlinear problems, such as soil dynamics (Andreadis and Lettenmaier 2006; Durand and Margulis 2008; Kumar et al. 2008; Pan and Wood 2006).

Reichle et al. (2002a) tested the EnKF by assimilating L-band (1.4 GHz) microwave brightness temperature observations into a LSM and demonstrated its superiority with respect to a dynamic variational assimilation method. They also showed how wetting and drying processes dominate the dynamic evolution of error variances. During dry-down variances are large and the soil moisture ensemble distribution is symmetric. But when the soil is either very wet or very dry, variances are smaller and the model nonlinearities are significant. The actual errors are therefore larger than the ones derived by the ensemble forecast, which means that the update is suboptimal. However, the degree of sub-optimality is relatively small and their conclusion is that the EnKF is still a robust data assimilation option even for modest ensemble sizes.

Reichle et al. (2002b) performed a comparison between the extended Kalman filter (EKF) and the EnKF for soil moisture estimation in a twin experiment over the southeastern United States. EKF is a generalization of the traditional Kalman Filter for nonlinear applications, but the computational demand due to the error covariance integration limits the size of the problem (Gelb 1974). For this reason EKF has been used only for estimating the vertical profile of soil moisture (Katul et al. 1993; Entekhabi et al. 1994). For instance, Walker and Houser (2001) have used an EKF technique to estimate soil moisture across North America by neglecting the horizontal error correlations and treating each surface hydrological catchment independently. The way EKF and EnKF treat nonlinearities is different: EKF linearizes the equation of the error covariance propagation while the EnKF propagates a set of ensemble of model trajectories. Both EKF and EnKF filters have been proven to provide satisfactory estimates of soil moisture at comparable computational cost. However, the EnKF was shown to be more flexible, as it avoids integrating the state error covariance matrix by propagating an ensemble of equally probable states from which to obtain the covariance information at each update step.

Gruber et al. (2015) investigated the potential for introducing a two-dimensional LDAS that uses spatial error auto-correlations of active and passive microwave soil moisture observations and LSM predictions. They showed that including information regarding the spatial error auto-correlation of these three products does not significantly improve the performance of the LDAS when compared to the skills of a simpler one-dimensional system.

Variational assimilation methods have the capability of achieving optimal performance, like Kalman filters, but with enhanced computational efficiency, as they do not explicitly estimate the large error covariance matrices that are propagated by

KFs (Thepaut and Courtier 1991; Courtier et al. 1993; McLaughlin 1995). Variational methods process all data concurrently during the assimilation interval and implicitly account for the dynamic error by propagating an adjoint. Specifically, the adjoint equation describes how measurement information obtained at a given step propagates backward in time.

Reichle et al. (2001) assessed the feasibility of estimating large-scale soil moisture from L-band passive microwave measurements, using a four-dimensional variational method that considers model and observation uncertainties. Their synthetic experiments showed that adequate soil moisture estimates could be obtained using the proposed approach. They also demonstrated that reducing the length of the assimilation interval resulted in poorer performances but in substantial improvements in computational efficiency.

In terms of the available software platforms for soil moisture data assimilation, the Land Information System (LIS) represents one of the most mature and complete land data assimilation systems currently available (Reichle et al. 2009; Kumar et al. 2008). LIS provides a common software framework capable of ensemble land surface modeling on points, regions, or the entire globe at spatial resolutions from 2×2.5 degrees down to meter scales. It is an interagency test bed for land surface modeling and data assimilation that allows customized land data assimilation systems to be built, assembled, and reconfigured easily, using shared plugins and standard interfaces. The NASA Goddard Earth Observing System Model, Version 5 (GEOS-5) has recently been integrated with LIS.

3 Data Assimilation Skill Evaluation

Fluctuations between estimated and reference soil moisture, expressed in terms of biases or root mean square errors (RMSEs), may vary considerably. Data assimilation systems commonly apply a cumulative density function transformation to the observations that are merged with the model predictions to remove the systematic error. Therefore, a statistical metric that is independent of any bias in the mean and standard deviation of the observations is needed to evaluate soil moisture predicted by LDASs (Entekhabi et al. 2010b). A common metric that measures the correspondence in phase between estimates and the benchmark, regardless of seasonal mean biases or differences in the standard deviation, is the anomaly correlation coefficient (ACC):

$$ACC = \frac{E[(\theta_{est} - E[\theta_{est}])(\theta_{true} - E[\theta_{true}])]}{\sigma_{est}\sigma_{true}} \quad (1)$$

where θ represents the soil moisture anomalies and σ^2 is the time variance at each location. Soil moisture anomalies are commonly defined as differences between the current values and the monthly climatological average values.

In order to evaluate the skills of the data assimilation system in producing soil moisture ensembles, the combination of exceedance ratio and uncertainty ratio provides a useful tool (Hossain et al. 2004; Hossain and Anagnostou 2005; Maggioni et al. 2012a). The exceedance ratio (ER) is defined as follows

$$ER = \frac{N_{exceedance}}{N_t} \quad (2)$$

where $N_{exceedance}$ is the number of times the true soil moisture falls outside the ensemble bounds and N_t represents the sample size. ER captures the ability of the estimated ensemble to encapsulate the reference soil moisture. A value of ER equal to 1 corresponds to a probability of 0 % that the reference falls within the ensemble bounds, while a value of ER close to 0 shows a good predictability of the model that most of the time encapsulates the reference inside the estimated envelope.

The uncertainty ratio (UR) is the ratio of the simulated uncertainty, defined as the average difference between the upper and lower limits of the ensemble, and the average true soil moisture:

$$UR = \frac{\sum_{i=1}^{N_t} (\theta_{upper}^i - \theta_{lower}^i)}{\sum_{i=1}^{N_t} (\theta^i)} \quad (3)$$

UR measures the variability in the estimated soil moisture ensemble relative to the typical value that reference soil moisture assumes. A perfect ensemble spread would have a UR equal to 1. A UR value lower than 1 indicates an underestimation of the model prediction error which means there are insufficient weights given to observations in an ensemble-based data assimilation system. On the other hand, UR greater than 1 corresponds to an overestimation of the error spread, which can translate into excessive weights given to the observations to be merged by the LDAS. The combination of these metrics account for two contrasting issues: if the uncertainty limits are too narrow (high UR values and low ER values) the model errors are underestimated, and if the ensemble width is too large (low UR values and high ER values) the LDAS has poor predictive capability.

Another valuable tool to evaluate soil moisture ensemble predictions is the rank histogram, introduced by Hamill (2001). Rank histograms are based on the assumption that all ensemble members would be equally probable realizations of the reference dataset, if the land data assimilation were able to reproduce the forecast distribution. In other words, if the simulated ensemble is statistically consistent, the number of times that the reference falls within any two adjacent members does not depend on the position of those members within the ordered ensemble (Siegert et al. 2012). In this case, the rank histogram will have a flat shape (Talagrand et al. 1997; Hamill and Colucci 1997; Hamill 2001). A U-shaped rank

histogram indicates a lack of variability in the ensemble, while a sloped histogram suggests consistent biases in the model (Fig. 2).

Some tests are necessary to verify whether the LDAS filtering technique operates in accordance with its assumptions and whether model and observational uncertainties are appropriately chosen. For instance, if the EnKF functions properly, the mean of the innovations should be statistically indistinguishable from zero. Innovations are defined as the difference between the observations and the soil moisture ensemble mean prior to the assimilation update. Moreover, the normalized innovations, i.e., innovations divided by their expected standard deviation, should be normally distributed with zero mean and standard deviation equal to 1 (Reichle et al. 2007). Therefore, hypothesis testing should be performed to verify whether the normalized innovations could be modeled with a standard normal distribution.

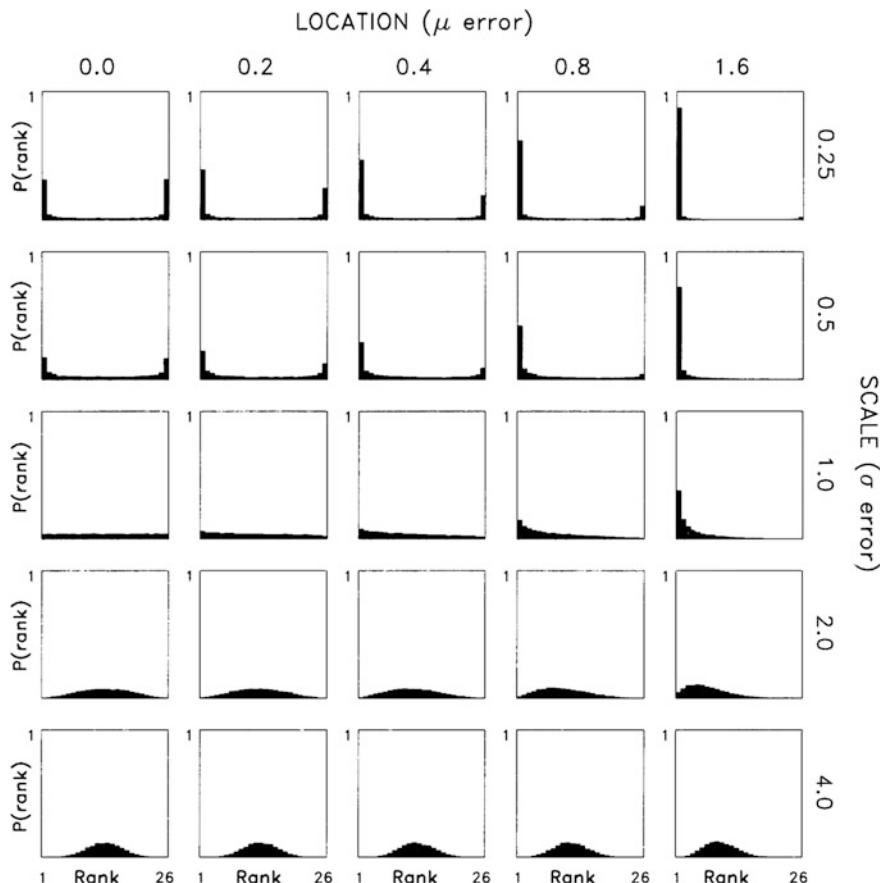


Fig. 2 Rank histograms where verification is sampled from a standard normal distribution and the 25-member ensemble is sampled from a normal distribution with mean μ and standard deviation σ . (Adopted from Hamill 2001)

4 Root-Zone Soil Moisture Estimation

Root-zone soil moisture plays a prime role in the regulation of water and energy budgets at the soil–vegetation–atmosphere interface by regulating evaporation and transpiration processes (Shukla and Mintz 1982). If the initialization of root-zone soil moisture in numerical weather and climate prediction models is not accurate, it may cause drifts of the temporal evolution of the surface state variables and degrade the forecast (Beljaars et al. 1996; Dirmeyer 2000; Koster and Suarez 2003). Root-zone soil moisture cannot be sensed by L-band microwave remote sensing. However, near-surface soil moisture (0–5 cm) is physically related to root-zone soil moisture through diffusion processes. Hence, assimilating satellite retrievals of near-surface soil moisture can yield improvements not only in surface soil moisture estimates, but also in root zone soil moisture estimation.

Sabater et al. (2007) investigated several assimilation techniques derived from Kalman filters and variational methods to correct the model forecasts of root-zone soil moisture from the Interaction between Soil, Biosphere, and Atmosphere (ISBA) LSM. Surface soil moisture ground measurements from the Surface Monitoring of the Soil Reservoir Experiment (SMOSREX) in France over a 4-year period were used as observations. The EnKF and a simplified one-dimensional variational data assimilation (1DVAR) were shown to perform best and both showed an improvement in the updated root-zone soil moisture products.

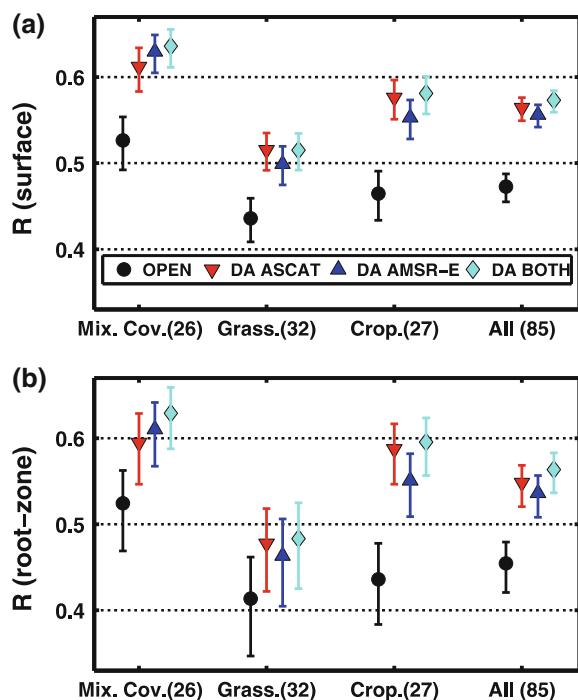
Das et al. (2008) showed that the assimilation of airborne remotely sensed surface soil moisture had limited influence on the profile soil moisture, as root zone soil moisture depended mostly on the soil type. A caveat of this study is that it focused on a limited area, i.e., the Walnut Gulch Experimental Watershed, Arizona, and on a limited period of time, i.e. August 2–27 2004 (Soil Moisture Experiment 2004—SMEX04). Surface soil moisture observed from aircraft was assimilated into a distributed Soil–Water–Atmosphere–Plant (SWAP) model, using an ensemble square root filter, based on a Kalman filter approach, at a spatial resolution of 800 m × 800 m. Assimilated soil moisture profile estimates were compared to in situ measurements collected during the experiment. These comparisons showed a reasonable agreement between the ground observations at various depths and the modeled profile of soil moisture. However, the open-loop (i.e., no-assimilation runs) and the assimilation experiments performed equally well in terms of root zone soil moisture at various depths.

Crow et al. (2008) substantiated the potential of improving root-zone soil moisture through assimilation of near-surface soil moisture showed by Sabater et al. (2007). Specifically, they proved the advantage of assimilating thermal remote sensing (RS) observations into soil-vegetation-atmosphere transfer (SVAT) models for root-zone soil water predictions. RS-SVATs use surface radiometric temperature from thermal remote sensing on cloud-free days and combine them with vegetation information obtained in the visible and near-infrared spectra for solving the surface energy balance (Norman et al. 1995). This methodology was shown to be superior to Water and Energy Balance (WEB) SVATmodels (Noilhan and Planton 1989; Montaldo et al. 2003) in the characterization of root-zone soil moisture.

The USDA Soil Climate Analysis Network (SCAN, <http://www.wcc.nrcs.usda.gov>; Schaefer et al. 2007) represents a valuable validation dataset for soil moisture profile studies, providing hourly soil moisture measurements at 123 sites across the continental United States, taken at depths of 5, 10, 20, 50, and 100 cm with Stevens Water Hydra Probe sensors. Liu et al. (2011) assessed the performance of assimilating surface soil moisture retrievals from AMSR-E using 35 SCAN sites. Assimilating AMSR-E retrievals was shown to increase root zone soil moisture skill in terms of the anomaly time series correlation coefficient.

Maggioni et al. (2012a, b) showed improved root zone soil moisture performance metrics, exhibiting higher ACCs and lower RMSEs when either synthetic or actual satellite surface soil moisture was assimilated in comparison with open-loop experiments. Similarly, the work by Draper et al. (2012) merged near-surface soil moisture data from the active microwave ASCAT and the passive microwave AMSR-E sensors with the NASA Catchment LSM predictions using an EnKF. They analyzed 85 sites in the US and Australia where ground observations were available to be used as reference. They found that the assimilation of ASCAT and AMSR-E held very similar improvements when compared to the open-loop experiments, as the mean root-zone soil moisture anomaly correlation coefficient increased from 0.45 to 0.55, 0.54, and 0.56 by assimilating ASCAT, AMSR-E, and both (Fig. 3).

Fig. 3 Mean correlation coefficient for **a** surface and **b** root zone soil moisture averaged across each land cover class (the number of sites in each class is given in the axis label). 95 % confidence levels are also shown. (Adopted from Draper et al. 2012)



5 Uncertainty Characterization in the Precipitation Forcing

The improvement due to the assimilation of soil moisture observations highly depends on the quality of the LSM meteorological forcings. Since precipitation is the major driver for soil moisture, uncertainty in the precipitation forcing plays a fundamental role in the efficiency of assimilating soil moisture observations. Liu et al. (2011) showed that assimilating satellite retrievals of near-surface soil moisture and correcting the precipitation forcing with rain gauge measurements similarly improve the skill of a LDAS in estimating surface and root zone soil moisture (Fig. 4). Specifically, an increase of ~ 0.08 in the surface soil moisture anomaly correlation coefficient was observed when AMSR-E retrievals were assimilated and an increase of ~ 0.06 was obtained by adding precipitation information. By combining information from both sources, soil moisture anomaly correlation coefficients improved by ~ 0.13 .

The quality of EnKF assimilation techniques largely depends on the representation of model and observational uncertainties (Reichle et al. 2008a). A poor characterization of these uncertainties is likely transferred to the land surface outputs (Crow and Van Loon 2006). The way model and observational errors are currently treated in LDASs is very basic. For instance, in the NASA Global

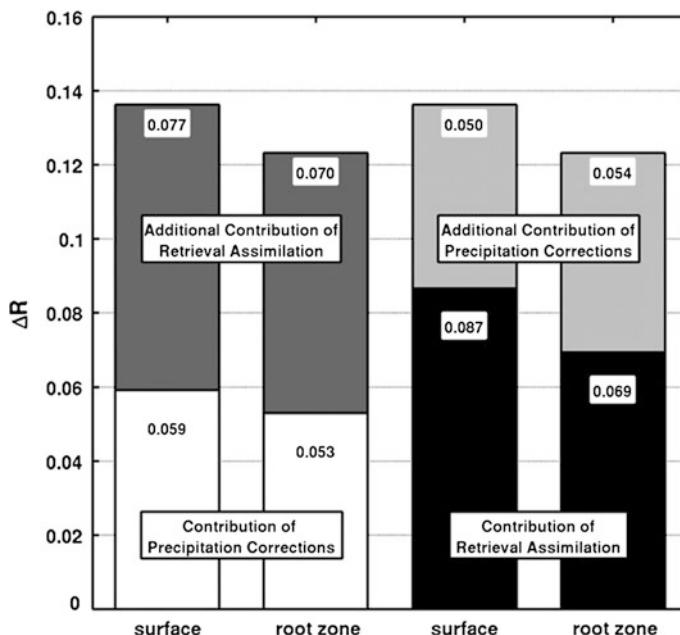


Fig. 4 Improvements in anomaly correlation coefficients due to precipitation corrections and retrieval assimilation. (Adopted from Liu et al. 2011)

Modeling and Assimilation Office Land Data Assimilation System (GMAO-LDAS) the model error scheme is a simple multiplicative perturbation applied to the precipitation forcing and to the state variables (Reichle et al. 2007). Although these precipitation perturbations are spatially and temporally correlated and log-normally distributed, a purely multiplicative error model implies that all ensemble members will be set to zero every time the input precipitation is zero. This kind of approaches is numerically convenient but not appropriate for precipitation errors, characterized not only by biases, but also by multi-dimensional correlations especially at fine space and time scales, detection uncertainties and false alarms. A more complex and complete precipitation error model has the impact of improving the skill of surface and root zone soil moisture estimated by a LDAS.

Past research studies have developed more sophisticated satellite rainfall error models for generating error ensembles of satellite rainfall fields (Hossain and Anagnostou 2006a; Gebremichael et al. 2011; Maggioni et al. 2014). A series of articles by Maggioni et al. (2011, 2012a, b) investigated the potential of implementing one of these more complex precipitation error approaches in a LDAS for soil moisture estimation. Specifically, they adopted the model proposed by Hossain and Anagnostou (2006a), the multi-dimensional Satellite Rainfall Error Model (SREM2D). SREM2D employs stochastic formulations to characterize the error structure of satellite retrievals based on high-accuracy reference rain fields and generates equally probable ensemble members of satellite rain fields. SREM2D accounts for the joint probability of successful delineation of rainy and non-rainy areas, temporal and spatial correlations of the error, missed rain events, and false alarms. Moreover, it was proven capable of conserving the rainfall error structure across scales, unlike simpler error models that commonly show significant scale-dependent biases (Hossain and Anagnostou 2006b).

Maggioni et al. (2011) expanded the study by Hossain and Anagnostou (2006b) to investigate the impact of satellite-rainfall error complexity on soil moisture uncertainty simulated by the NASA Catchment land surface model (CLSM, Koster et al. 2000). SREM2D was compared to the standard model (hereinafter CTRL) used to generate rainfall ensembles in the GMAO-LDAS. Both rainfall error models were shown to reproduce the satellite rainfall error characteristics across different spatial scales. Though, SREM2D generated an ensemble with higher variability than the traditional error scheme and was able to better envelop the rain gauge-calibrated radar rainfall fields, considered as the reference rain dataset. In terms of soil moisture modeled by CLSM, forced with perturbed rain fields, the SREM2D-based soil moisture ensemble also presented larger spread than the one generated using CTRL and showed lower probability of exceedance, i.e., the probability that the reference soil moisture falls outside the ensemble envelope (Fig. 5). Nevertheless, soil moisture was found to be less sensitive than precipitation to the complexity of the rainfall error scheme, due to the characteristics of soil moisture dynamics, which are highly dissipative and nonlinear.

When rain error modeling techniques are compared in the assimilation system, in which synthetic soil moisture observations are merged with CLSM predictions, the more complex rainfall error model was shown to only slightly improve soil

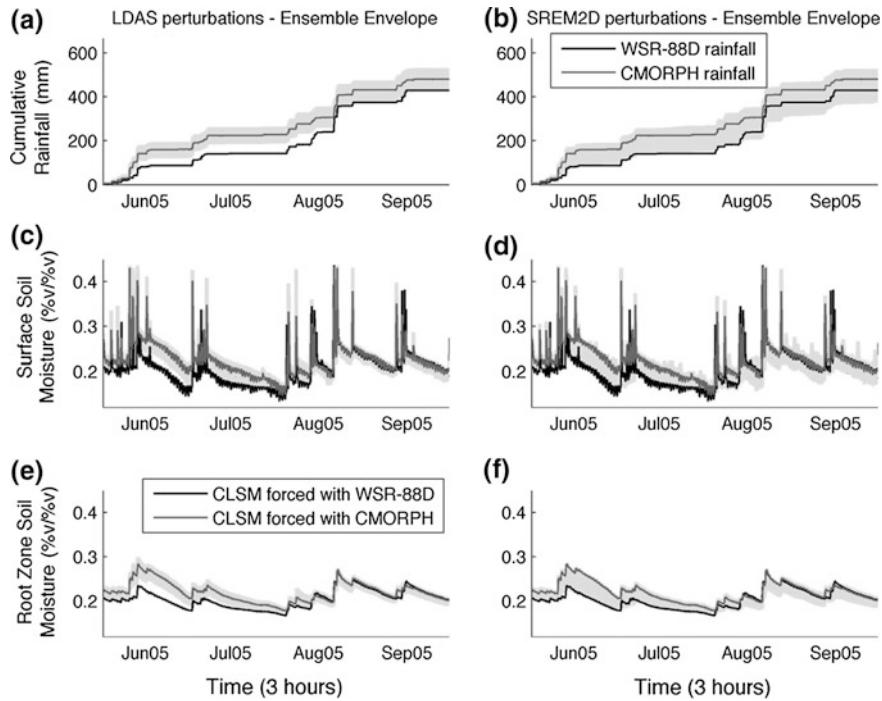


Fig. 5 Time series of cumulative rainfall (a and b), surface soil moisture (c and d), and root zone soil moisture (e and f). Results are shown for the CTRL error model from (a, c, and e) and for SREM2D (b, d, f). (Adopted from Maggioni et al. 2011)

moisture ACCs and the RMSEs (Maggioni et al. 2012a). The relative improvement due to SREM2D over the standard error scheme was observed to be larger for root zone soil moisture, which carries the memory of previous rainfall events (Fig. 6). Maggioni et al. (2012a) also studied the impact of a more complex rain error model on soil moisture data assimilation as a function of a climatological wetness indicator. A positive (negative) value of this indicator indicates an area that is generally wetter (drier) than the domain climatology. The spatial variability in the RMSE reduction amplifies with increasing the climatological wetness indicator, whereas the maximum relative increase in anomaly correlation coefficients is near neutral regimes (i.e., non-extreme rainfall conditions). The fact that normal to wetter climatological conditions are more affected by the way rainfall error is characterized in the LDAS demonstrates the high dependence of soil moisture on precipitation variability.

Similar conclusions were drawn when actual near-surface soil moisture retrievals from AMSR-E were assimilated in a LDAS, adopting the two different rainfall error models to perturb satellite precipitation fields (Maggioni et al. 2012b). First off, the LDAS soil moisture estimates performed as well as the benchmark model simulation forced with high-quality radar precipitation. Both open-loop and data

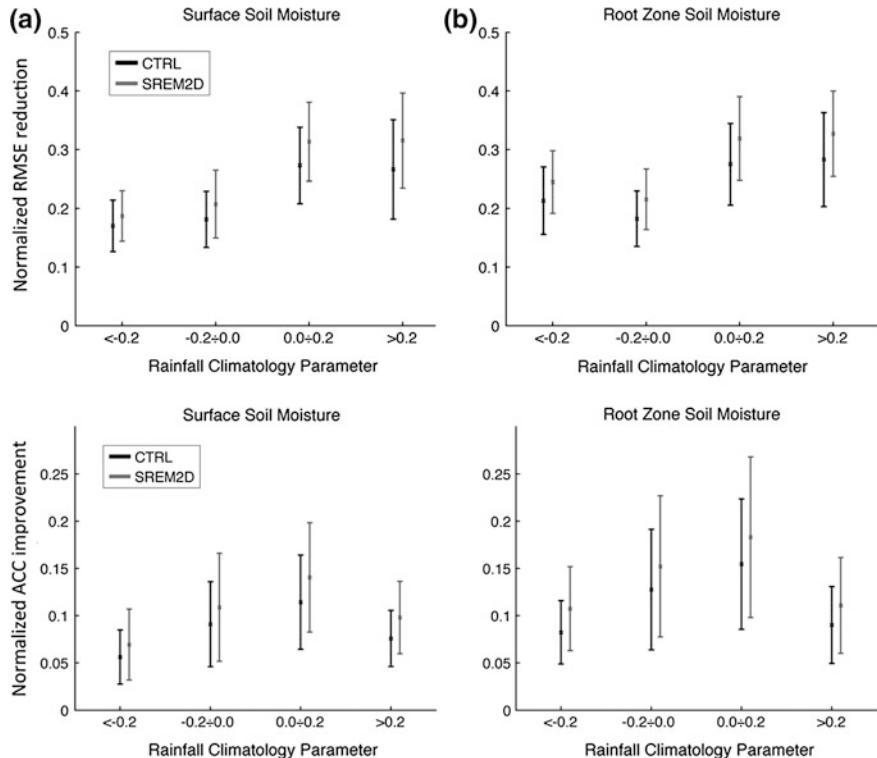


Fig. 6 Normalized RMSE reduction (top) and anomaly correlation coefficient improvements (bottom) as function of rainfall climatology. (Adopted from Maggioni et al. 2012a)

assimilation simulations performed better in the wetter conditions, but the relative improvement in surface and root zone soil moisture skills due to data assimilation was larger in drier conditions (Fig. 7). SREM2D exhibited slight improvements in soil moisture estimates in terms of anomaly correlation coefficients with respect to the CTRL model. Satellite rainfall retrievals are largely affected by false alarms. This is particularly evident in arid and semiarid areas, where microwave techniques are limited in detecting low rain rates, because of potential effects of soil wetness and below-cloud evaporation. It is therefore promising that SREM2D adds more value in the drier regimes.

Errors in the precipitation forcing are not the only source of uncertainty in a LDAS and the contribution of rainfall forcing errors relative to model structure and model parameter uncertainty should be further investigated. A first attempt is offered by Maggioni et al. (2012c), who studied this problem for soil moisture estimates obtained by integrating the NASA CLSM. SREM2D was used to generate an ensemble of rainfall fields from satellite rainfall retrievals and model errors were represented (i) by perturbations in the model parameters only and (ii) by perturbations in the model prognostics to represent the combination of model structure

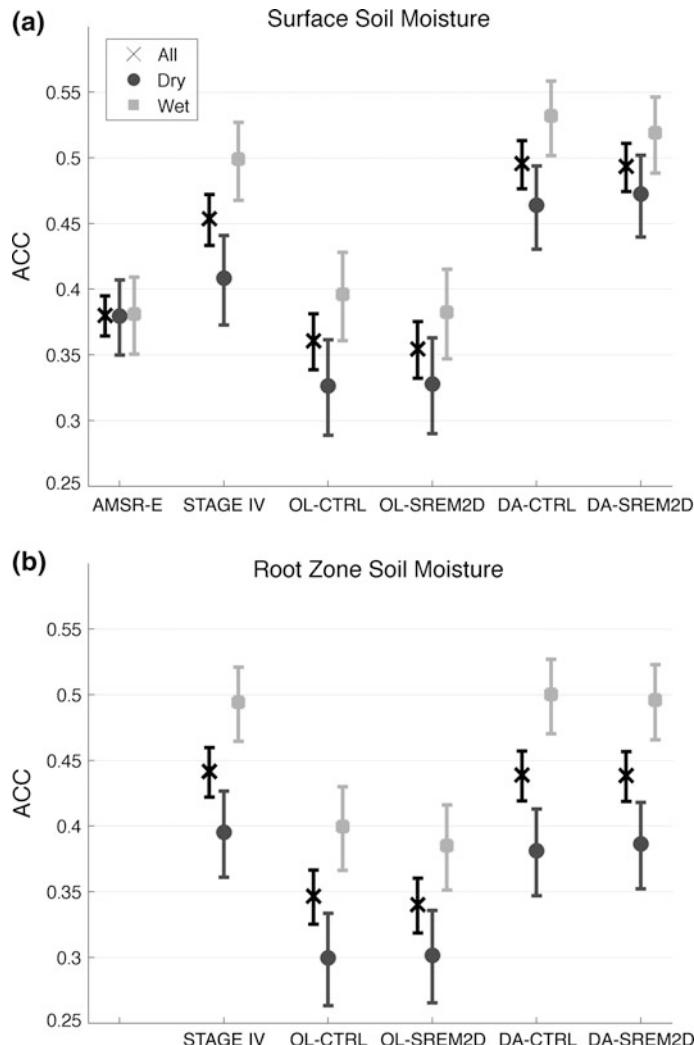


Fig. 7 Daily anomaly correlation coefficients for **a** surface and **b** root zone soil moisture with 95 % confidence intervals for dry and wet conditions. (Adopted from Maggioni et al. 2012b)

and parameter uncertainty. Their findings showed that the method currently used in the NASA LDAS to perturb model variables does not fully describe the modeled soil moisture uncertainty, even when combined with rain forcing perturbations. A better characterization of soil moisture uncertainty is possible by adding model parameter perturbations to rainfall forcing perturbations. Future research work should focus on a more accurate characterization of uncertainty in a LDAS that combines rainfall and model parameter errors to improve soil moisture estimation.

6 Multi-scale Soil Moisture Data Assimilation

Data assimilation systems have been commonly used for downscaling coarser resolution satellite retrievals, which update land surface model prognostics to produce higher resolution soil moisture estimates. For instance, Reichle et al. (2001) conducted several synthetic experiments to evaluate the possibility of assimilating L-band passive microwave measurements, using a four-dimensional variational assimilation method. They observed that adequate soil moisture estimates could be obtained at resolutions of a few kilometers, much finer than the resolution of the satellite brightness temperature retrievals (i.e., tens of kilometers). Nevertheless, this is only possible if high quality meteorological, soil texture, and land cover inputs are available at that finer scale.

De Lannoy et al. (2009) tested different assimilation techniques to merge coarse-scale snow water equivalent observations into finescale land model simulations (Fig. 8). Specifically, they directly assimilated coarse-scale observations

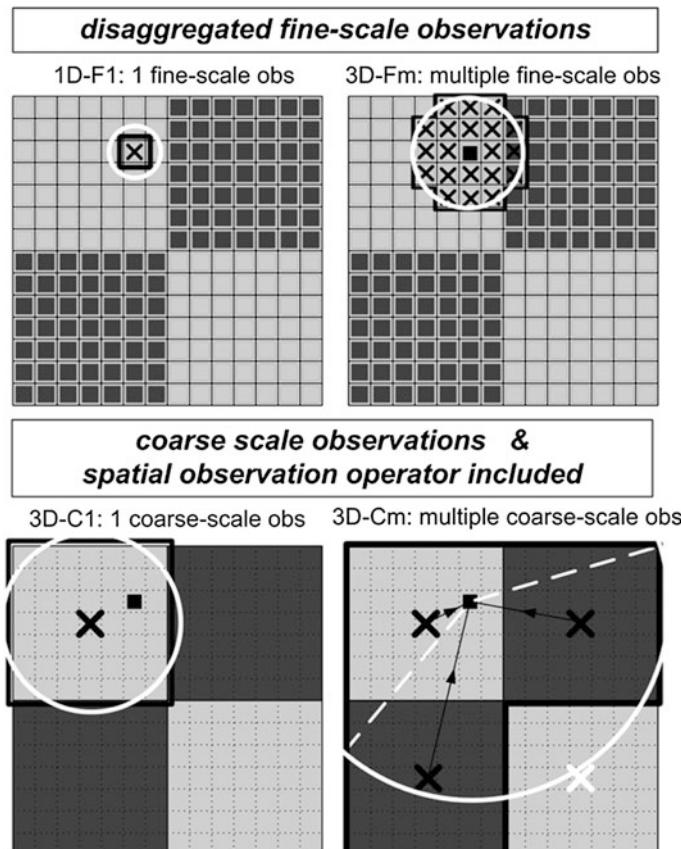


Fig. 8 Schematic of the multi-resolution EnKF proposed by De Lannoy et al. (2009)

using an observation operator for mapping between the coarse and fine scales or, alternatively, after disaggregation (re-gridding) to the finescale model resolution prior to data assimilation. Observations were assimilated either simultaneously or independently for each location. Their results indicated that assimilating disaggregated finescale observations independently is less efficient than assimilating a collection of neighboring disaggregated observations.

De Lannoy et al. (2009) also showed that direct assimilation of coarse-scale observations is superior to a priori disaggregation. Independent assimilation of individual coarse-scale observations was shown to bring the overall mean field close to the truth, without necessarily improving estimates of the finescale structure. A clear benefit to simultaneously assimilating multiple coarse-scale observations was demonstrated even as the entire domain was observed, indicating that underlying spatial error correlations can be exploited to improve the state estimate.

Although land surface models can provide useful information on spatial and temporal soil moisture variability, they are well known to produce biased forecasts (e.g., multi-model NLDAS results). A simple method for bias removal, proposed by Reichle and Koster (2004), is to match the cumulative distribution functions of the observations—typically satellite retrievals—and model predictions. However, in order to accurately estimate those functions, a long record of satellite data is required. Thus, downscaled satellite soil moisture products can largely benefit hyper-resolution modeling, as they have the potential to correct for those biases.

7 Summary

By regulating the exchange of water and energy at the interface between land and atmosphere, soil moisture plays a fundamental role in atmospheric, hydrologic, and biogeochemical processes. Soil moisture can be measured from in situ observations, remote-sensing retrievals from active and passive microwave sensors, and by integrating a land surface model. In situ measurements are limited in space and time and include direct and indirect approaches. Direct methods collect soil samples that are evaluated in the laboratory (e.g., gravimetric method), whereas indirect methods measure soil properties that are physically related to the soil water content (e.g., dielectric constant, soil suction). Surface soil moisture (top 5 cm of the soil) can be also estimated using land surface brightness temperature observed by satellite microwave sensors. These estimates are quasi-global, spatially and temporally distributed, but provided at coarse resolutions and inadequate in coastal areas and regions where the surface is frozen or densely vegetated.

Another common technique to estimate soil moisture is by integrating a land surface model. However, these predictions are affected by errors in the model parameters and formulation and uncertainties in the meteorological forcings. Land data assimilation systems combine the information from ground and/or satellite observations and model prediction into an improved estimate of surface and root-zone soil moisture. Ensemble Kalman filters and variational methods are

among the most popular for soil moisture data assimilation. The EnKF is extremely flexible in treating model and parameter errors and is suitable for nonlinear problems, while variational methods have been able to achieve optimality more efficiently, from a computational point of view.

Data assimilation has been proven to be particularly useful to estimate root zone soil moisture, which is not directly measurable from L-band microwave remote-sensing. Root-zone soil moisture controls the water and energy budgets at the interface between soil, vegetation and atmosphere through evapotranspiration. Soil moisture initialization in weather and climate models is fundamental to prevent degrading the forecast performance. Past studies show that both EnKF and variational data assimilation methods were able to improve root-zone soil moisture estimates by updating the model output with surface soil moisture from ground and satellite measurements. Additionally, LDASs have been used to downscale satellite brightness temperature observations, which are available at coarse resolutions to generate higher resolution soil moisture products.

The efficiency of assimilating soil moisture data depends on the way uncertainty in the meteorological forcings, model structure, model parameters and observations are addressed. Current methods used to perturb model forcings, variables, and parameters in LDASs are often too simplistic and do not entirely describe the uncertainty of the system. For instance, using a more complex rainfall error model than a traditional multiplicative approach was shown to improve soil moisture anomaly correlation coefficients and root mean square errors. This improvement is larger for root zone soil moisture and under non-extreme rainfall climatological conditions.

Common skill metrics to evaluate land data assimilation improvement are biases, root mean square errors, and anomaly correlation coefficients. This latter is particularly suitable as it is independent of any systematic errors, which are commonly removed in LDASs through cumulative density function transformations. Furthermore, two useful statistics are the exceedance ratio, which measure the probability of the estimated soil moisture ensemble to encapsulate the reference, and uncertainty ratio, which captures the variability in the simulated ensemble. Rank histograms are also commonly used to verify whether the estimated ensemble is statistically consistent, whether it lacks variability or whether it carries consistent biases.

References

- Andreadis KM, Lettenmaier DP (2006) Assimilating remotely sensed snow observations into a macroscale hydrology model. *Adv Water Resour* 29:872–886. doi:[10.1016/j.advwatres.2005.08.004](https://doi.org/10.1016/j.advwatres.2005.08.004)
- Bartalis Z, Naeimi V, Hasenauer S, Wagner W (2008) ASCAT soil moisture product handbook. *ASCAT Soil Moisture Rep Ser* **15**

- Beljaars ACM, Viterbo P, Miller MJ, Betts AK (1996) The anomalous rainfall over the United States during 1993: sensitivity to land surface parameterization and soil moisture anomalies. *Mon Weather Rev* 124:362–383
- Betts AK, Ball JH (1998) FIFE surface climate and site- average dataset 1987–89. *J Atmos Sci* 55:1091–1108
- Carpenter TM, Sperfslage JA, Georgakakos KP, Sweeney T, Fread DL (1999) National threshold runoff estimation utilizing GIS in support of operational flash flood warning system. *J Hydrol* 224:21–44
- Courtier P, Derber J, Errico R, Louis J-F, Vulicevic T (1993) Important literature on the use of adjoint, variational methods and the Kalman filter in meteorology. *Tellus A* 45:342–357
- Crow WT, Wood EF (2003) 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* 26:137–149
- Crow WT, Van Loon E (2006) Impact of incorrect model error assumptions on the sequential assimilation of remotely sensed surface soil moisture. *J Hydrometeor* 7:421–432
- Crow WT, Kustas WP, Prueger JH (2008) Monitoring root-zone soil moisture through the assimilation of a thermal remote sensing-based soil moisture proxy into a water balance model. *Remote Sens. Environ. (Remote Sensing Data Assimilation Special Issue)* **112**, 1268–1281. doi:[10.1016/j.rse.2006.11.033](https://doi.org/10.1016/j.rse.2006.11.033)
- Das NN, Mohanty BP, Cosh MH, Jackson TJ (2008) Modeling and assimilation of root zone soil moisture using remote sensing observations in Walnut Gulch Watershed during SMEX04. *Remote Sens Environ* 112(2):415–429
- De Lannoy GJM, Reichle RH, Houser PR, Pauwels VRN, Verhoest NEC (2007) Correcting for forecast bias in soil moisture assimilation with the ensemble Kalman filter. *Water Resour Res* 43:W09410. doi:[10.1029/2006WR005449](https://doi.org/10.1029/2006WR005449)
- De Lannoy GJM, Reichle RH, Houser PR, Arsenault KR, Verhoest NEC, Pauwels VRN (2009) Satellite-scale snow water equivalent assimilation into a high-resolution land surface model. *J Hydrometeor* 11:352–369. doi:[10.1175/2009JHM1192.1](https://doi.org/10.1175/2009JHM1192.1)
- Dirmeyer PA (2000) Using a global soil wetness dataset to improve seasonal climate simulation. *J Clim* 13:2900–2922
- Dorigo WA, Van Oevelen P, Wagner W, Drusch M, Mecklenburg S, Robock A, Jackson T (2011) A new international network for in situ soil moisture data. *Eos Trans Am Geophys Union* 92:141–142
- Draper CS, Reichle RH, De Lannoy GJM, Liu Q (2012) Assimilation of passive and active microwave soil moisture retrievals. *Geophys Res Lett* 39:L04401. doi:[10.1029/2011GL050655](https://doi.org/10.1029/2011GL050655)
- Drusch M (2007) Initializing numerical weather prediction models with satellite-derived surface soil moisture: data assimilation experiments with ECMWF's integrated forecast system and the TMI soil moisture data set. *J Geophys Res.* **112**(D03102). doi:[10.1029/2006JD007478](https://doi.org/10.1029/2006JD007478)
- Drusch M, Scipal K, de Rosnay P, Balsamo G, Andersson E, Bougeault P, Viterbo P (2009) Towards a Kalman Filter based soil moisture analysis system for the operational ECMWF Integrated Forecast System. *Geophys Res Lett* **36**(10)
- Dunne S, Entekhabi D (2006) Land surface state and flux estimation using the ensemble Kalman smoother during the Southern Great Plains 1997 field experiment. *Water Resour Res* 42: W01407. doi:[10.1029/2005WR004334](https://doi.org/10.1029/2005WR004334)
- Durand M, Margulis S (2008) Effects of uncertainty magnitude and accuracy on assimilation of multi-scale measurements for snowpack characterization. *J Geophys Res* **113**(D02105). doi:[10.1029/2007JD008662](https://doi.org/10.1029/2007JD008662)
- Entekhabi D et al (2010a) The Soil Moisture Active and Passive (SMAP) Mission. *Proc IEEE* **98**, 704–716. doi:[10.1109/JPROC.2010.2043918](https://doi.org/10.1109/JPROC.2010.2043918)
- Entekhabi D, Nakamura H, Njoku EG (1994) 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
- Entekhabi D, Reichle RH, Koster RD, Crow WT (2010) Performance metrics for soil moisture retrievals and application requirements. *J Hydrometeor* 11:832–840

- Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using monte carlo methods to forecast error statistics. *J Geophys Res* **99**(C5), 10 143–10 162
- Gao H, Wood EF, Jackson TJ, Drusch M, Bindlish R (2006) Using TRMM/TMI to retrieve surface soil moisture over the Southern United States from 1998 to 2002. *J Hydrometeorol* **7**:23–38
- García-Pintado J, Neal JC, Mason DC, Dance SL, Bates PD (2013) Scheduling satellite-based SAR acquisition for sequential assimilation of water level observations into flood modelling. *J Hydrol* **495**, 252–266
- Gebremichael M, Liao G-Y, Yan J (2011) Nonparametric error model for a high resolution satellite rainfall product. *Water Resour Res* **47**:W07504. doi:[10.1029/2010WR009667](https://doi.org/10.1029/2010WR009667)
- Gelb A (ed) (1974) Applied optimal estimation. M.I.T. Press, Cambridge, MA, 374 pp
- Gruber A, Crow W, Dorigo W, Wagner W (2015) The potential of 2D Kalman filtering for soil moisture data assimilation. *Remote Sens Environ* **171**:137–148. doi:[10.1016/j.rse.2015.10.019](https://doi.org/10.1016/j.rse.2015.10.019)
- Hamill TM, Colucci SJ (1997) Verification of eta-rsm short-range ensemble forecasts. *Mon Weather Rev* **125**:1312–1327. doi:[10.1175/1520-0493](https://doi.org/10.1175/1520-0493)
- Hamill TM (2001) Interpretation of rank histograms for verifying ensemble forecasts. *Mon Weather Rev* **129**:550–560. doi:[10.1175/1520-0493](https://doi.org/10.1175/1520-0493)
- Hillel D (1998) Environmental soil physics. Academic Press, San Diego, 771 pp
- Hoeben R, Troch PA (2000) Assimilation of active microwave observation data for soil moisture profile estimation. *Water Resour Res* **36**:2805–2819. doi:[10.1029/2000WR900100](https://doi.org/10.1029/2000WR900100)
- Hossain F, Anagnostou EN, Borga M, Dinku T (2004) Hydrological model sensitivity to parameter and radar rainfall estimation uncertainty. *Hydrol Process* **18**:3277–3299. doi:[10.1002/hyp.5659](https://doi.org/10.1002/hyp.5659)
- Hossain F, Anagnostou EN (2005) Numerical investigation of the impact of uncertainties in satellite rainfall estimation and land surface model parameters on simulation of soil moisture. *Adv Water Resour* **28**:1336–1350
- Hossain F, Anagnostou EN (2006a) A two-dimensional satellite rainfall error model. *IEEE Trans Geosci Remote Sens* **44**:1511–1522
- Hossain F, Anagnostou EN (2006b) Assessment of a multi-dimensional satellite rainfall error model for ensemble generation of satellite rainfall data. *Geosci Remote Sens Lett* **3**:419–423
- Houser PR, Shuttleworth WJ, Famiglietti JS, Gupta HV, Syed KH, Goodrich DC (1998) Integration of soil moisture remote sensing and hydrologic modeling using data assimilation. *Water Resour Res* **34**:3405–3420. doi:[10.1029/1998WR900001](https://doi.org/10.1029/1998WR900001)
- Jackson TJ (1993) Measuring surface soil moisture using passive microwave remote sensing. *Hydrol Process* **7**:139–152
- Katul GG, Wendorff O, Parlange MB, Puente CE, Folegatti MV, Nielsen DR (1993) Estimation of in situ hydraulic conductivity function from nonlinear filtering theory. *Water Resour Res* **294**:1063–1070
- Kerr YH et al (2010) The SMOS mission: new tool for monitoring key elements of the global water cycle. *Proc IEEE* **98**, 666–687, doi:[10.1109/JPROC.2010.2043032](https://doi.org/10.1109/JPROC.2010.2043032)
- Koster RD, Suarez MJ (2003) Impact of land surface initialization on seasonal precipitation and temperature prediction. *J Hydrometeorol* **4**:408–423
- Koster RD, Suarez MJ, Ducharne A, Stieglitz M, Kumar P (2000) A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *J Geophys Res* **105**:24809–24822. doi:[10.1029/2000JD900327](https://doi.org/10.1029/2000JD900327)
- Kumar SV, Peters-Lidard C, Tian Y, Reichle RH, Alonge C, Geiger J, Eylander J, Houser P (2008) An integrated hydrologic modeling and data assimilation frame-work enabled by the Land Information System (LIS). *IEEE Comput* **41**:52–59
- Lagerloef GSE, Colomb FR, Le Vine DM, Wentz FJ, Yueh SH, Ruf CS, Lilly J, Gunn J, Chao Y, de Charon A, Feldman G, Swift CT (2008) The Aquarius/SAC-D mission. *Oceanography* **21**:68–81
- Li F, Crow WT, Kustas WP (2010) Towards the estimation root-zone soil moisture via the simultaneous assimilation of thermal and microwave soil moisture retrievals. *Adv Water Resour* **33**:201–214

- Li L, Gaiser P, Jackson T, Bindlish R, Du J (2007) WindSat soil moisture algorithm and validation. In: IEEE International Geoscience and Remote Sensing Symposium, pp. 1188–1191. doi:[10.1109/IGARSS.2007.4423017](https://doi.org/10.1109/IGARSS.2007.4423017). Accessed 23–28 Jul 2007
- Liu Q, Reichle RH, Bindlish R, Cosh MH, Crow WT, de Jeu R, De Lannoy GJM, Huffman GJ, Jackson TJ (2011) The contributions of precipitation and soil moisture observations to the skill of soil moisture estimates in a land data assimilation system. *J Hydrometeor* 12:750–765. doi:[10.1175/JHM-D-10-05000.1](https://doi.org/10.1175/JHM-D-10-05000.1)
- Maggioni V, Sapiano MRP, Adler RF, Tian Y, Huffman GJ (2014) An error model for uncertainty quantification in high-time resolution precipitation products. *J Hydrometeor* 15:1274–1292. doi:[10.1175/JHM-D-13-0112.1](https://doi.org/10.1175/JHM-D-13-0112.1)
- Maggioni V, Reichle RH, Anagnostou EN (2011) The effect of satellite-rainfall error modeling on soil moisture prediction uncertainty. *J Hydrometeor* 12:413–428. doi:[10.1175/2011JHM1355.1](https://doi.org/10.1175/2011JHM1355.1)
- Maggioni V, Reichle RH, Anagnostou EN (2012a) The impact of rainfall error characterization on the estimation of soil moisture fields in a land data assimilation system. *J Hydrometeor* 13:1107–1118. doi:[10.1175/JHM-D-11-0115.1](https://doi.org/10.1175/JHM-D-11-0115.1)
- Maggioni V, Reichle RH, Anagnostou EN (2012b) The efficiency of assimilating satellite soil moisture retrievals in a land data assimilation system using different rainfall error models. *J Hydrometeor* 14(1):368–374
- Maggioni V, Anagnostou EN, Reichle RH (2012c) The impact of land model structural, parameter, and forcing errors on the characterization of soil moisture uncertainty. *Hydrol Earth Syst Sci* 16:3499–3515
- Margulis SA, McLaughlin D, Entekhabi D, Dunne S (2002) Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 Field Experiment. *Water Resour Res* 38:1299. doi:[10.1029/2001WR001114](https://doi.org/10.1029/2001WR001114)
- Martina MLV, Todini E, Libralon A (2006) A Bayesian decision approach to rainfall thresholds based flood warning. *Hydrol Earth Syst Sci* 10(3):413–426
- McLaughlin D (1995) Recent advances in hydrologic data assimilation. U.S. Natl Rep Int Union Geod Geophys 1991–1994, *Rev Geophys* 33, 977–984
- Montaldo N, Albertson JD (2003) Multi-scale assimilation of surface soil moisture data for robust root zone moisture predictions. *Adv Water Resour* 26:33–44. doi:[10.1016/S0309-1708\(02\)00103-3](https://doi.org/10.1016/S0309-1708(02)00103-3)
- Njoku EG (2011) Updated daily, AMSR-E/Aqua L2B surface soil moisture, ancillary parms, & QC EASE-Grids, June 2002 to June 2011, Boulder, CO, USA: National Snow and Ice Data Center, Digital media. (<http://nsidc.org/data/amsre>)
- Njoku EG, Jackson TJ, Lakshmi V, Chan TK, Nghiem SV (2003) Soil moisture retrieval from AMSR-E. *IEEE Trans Geosci Remote Sens* 41:215–229
- Noilhan J, Planton S (1989) A simple parameterization of land surface processes for meteorological models. *Mon Weather Rev* 117(3):536–549
- Norman JM, Kustas WP, Humes KS (1995) Source approach for estimating soil and vegetation energy fluxes in observations of directional radiometric surface temperature. *Agric Meteorol* 77 (3):263–293
- Owe M, de Jeu R, Holmes T (2008) Multisensor historical climatology of satellite-derived global land surface moisture. *J Geophys Res* 113:F01002. doi:[10.1029/2007JF000769](https://doi.org/10.1029/2007JF000769)
- Pan M, Wood EF (2006) data assimilation for estimating the terrestrial water budget using a constrained ensemble Kalman filter. *J Hydrometeor* 7:534–547
- Parajka J, Naeimi V, Bloeschl G, Wagner W, Merz R, Scipal K (2006) Assimilating scatterometer soil moisture data into conceptual hydrologic models at the regional scale. *Hydrol Earth Syst Sci* 10, 353–368. doi:[10.5194/hess-10-353-2006](https://doi.org/10.5194/hess-10-353-2006)
- Peters-Lidard CD, Houser PR, Tian Y, Kumar SV, Geiger J, Olden S, Lighty L, Doty B, Dirmeyer P, Adams J, Mitchell K, Wood EF, Sheffield J (2007) High performance earth system modeling with NASA/GSFC's Land Information System. *Innovations Syst Softw Eng* 3(3), 157–165
- Reichle RH, Koster RD (2003) Assessing the impact of horizontal error correlations in background fields on soil moisture estimation. *J Hydrometeor* 4(6):1229–1242

- Reichle RH, Koster RD (2004) Bias reduction in short records of satellite soil moisture. *Geophys Res Lett* **31**. doi:[10.1029/2004GL020938](https://doi.org/10.1029/2004GL020938)
- Reichle RH, Koster RD (2005) Global assimilation of satellite surface soil moisture retrievals into the NASA Catchment land surface model. *Geophys Res Lett* **32**. doi:[10.1029/2004GL021700](https://doi.org/10.1029/2004GL021700)
- Reichle RH, Entekhabi D, McLaughlin DB (2001) Downscaling of radio brightness measurements for soil moisture estimation: a four-dimensional variational data assimilation approach. *Water Resour Res* **37**:2353–2364. doi:[10.1029/2001WR000475](https://doi.org/10.1029/2001WR000475)
- Reichle RH, McLaughlin D, Entekhabi D (2002a) Hydrologic data assimilation with the ensemble Kalman filter. *Mon Weather Rev* **130**(1):103–114
- Reichle RH, Walker JP, Koster RD, Houser PR (2002b) Extended versus ensemble Kalman filtering for land data assimilation. *J Hydrometeorol* **3**(6):728–740
- Reichle RH, Koster RD, Dong J, Berg AA (2004) Global Soil moisture from satellite observations, land surface models, and ground data: implications for data assimilation. *J Hydrometeorol* **5**:430–442
- Reichle RH, Koster RD, Liu P, Mahanama SPP, Njoku EG, Owe M (2007) Comparison and assimilation of global soil moisture retrievals from AMSR-E and SMMR. *J Geophys Res* **112**: D09108. doi:[10.1029/2006JD008033](https://doi.org/10.1029/2006JD008033)
- Reichle RH, Crow WT, Keppenne CL (2008a) An adaptive ensemble Kalman filter for soil moisture data assimilation. *Water Resour Res* **44**(W03423). doi:[10.1029/2007WR006357](https://doi.org/10.1029/2007WR006357)
- Reichle RH, Crow WT, Koster RD, Sharif H, Mahanama SPP (2008b) Contribution of soil moisture retrievals to land data assimilation products. *Geophys Res Lett* **35**(L01404). doi:[10.1029/2007GL031986](https://doi.org/10.1029/2007GL031986)
- Reichle RH, Bosilovich MG, Crow WT, Koster RD, Kumar SV, Mahanama SPP, Zaitchik BF (2009) Recent advances in land data assimilation at the NASA global modeling and assimilation office. In Park SK, Xu L (eds) *Data assimilation for atmospheric, oceanic and hydrologic applications*. Springer Verlag, New York, pp. 407–428. doi:[10.1007/978-3-540-71056-1](https://doi.org/10.1007/978-3-540-71056-1)
- Robinson DA, Campbell CS, Hopmans JW, Hornbuckle BK, Jones SB, Knight R, Ogden F, Selker J, Wendoroth O (2008) Soil moisture measurements for ecological and hydrological watershed scale observatories: a review. *Vadose Zone J* **7**:358–389. doi:[10.2136/vzj2007.0143](https://doi.org/10.2136/vzj2007.0143)
- Robock A, Vinnikov KY, Srinivasan G, Entin JK, Hollinger SE, Speranskaya NA, Liu S, Namkhai A (2000) The global soil moisture data bank. *Bull. Am. Meteorol Soc* **81**:1281–1299. doi:[10.1175/1520-0477.2](https://doi.org/10.1175/1520-0477.2)
- Sabater JM, Jarlan L, Calvet J-C, Bouyssel F, De Rosnay P (2007) From near-surface to root-zone soil moisture using different assimilation techniques. *J Hydrometeorol* **8**:194–206. doi:[10.1175/JHM571.1](https://doi.org/10.1175/JHM571.1)
- Schaefer GL, Cosh MH, Jackson TJ (2007) The USDA natural resources conservation service soil climate analysis network (SCAN). *J Atmos Oceanic Technol* **24**(12):2073–2077
- Schmegge TJ, Kustas WP, Ritchie JC, Jackson TJ, Rango A (2002) Remote sensing in hydrology. *Adv Water Resour* **25**:1367–1385
- Seneviratne SI, Corti T, Davin EL, Hirschi M, Jaeger EB, Lehner I, Orlowsky B, Teuling AJ (2010) Investigating soil moisture–climate interactions in a changing climate: a review. *Earth Sci Rev* **99**:125–161. doi:[10.1016/j.earscirev.2010.02.004](https://doi.org/10.1016/j.earscirev.2010.02.004)
- Seuffert G, Wilker H, Viterbo P, Mahfouf J-F, Drusch M, Calvet J-C (2003) Soil moisture analysis combining screen-level parameters and microwave brightness temperature: a test with field data. *Geophys Res Lett* **30**:1498. doi:[10.1029/2003GL017128](https://doi.org/10.1029/2003GL017128)
- Shukla J, Mintz Y (1982) Influence of land-surface evapotranspiration on the earth's climate. *Science* **215**:1498–1501
- Siegert S, Broker J, Kantz H (2012) Rank histograms of stratified Monte Carlo ensembles. *Mon Weather Rev* **140**:1558–1571
- Talagrand O, Vautard R, Strauss B (1997) Evaluation of probabilistic prediction systems. In: *Proceedings, ECMWF workshop on predictability*, ECMWF, available from ECMWF, Shinfield Park, Reading, Berkshire, UK, pp. 1–25

- Thepaut J-N, Courtier P (1991) Four-dimensional variational data assimilation using the adjoint of a multilevel primitive-equation model. *Q J R Meteorol Soc* 117:1225–1254
- Wagner W, Naeimi V, Scipal K, de Jeu R, Martínez-Fernández J (2007) Soil moisture from operational meteorological satellites. *Hydrogeol J* 15:121–131
- Walker JP, Houser PR (2001) A methodology for initializing soil moisture in a global climate model: assimilation of near-surface soil moisture observations. *J Geophys Res* 106:11761–11774
- Walker JP, Willgoose GR, Kalma JD (2004) In situ measurement of soil moisture: a comparison of techniques. *J Hydrol* 293:85–99
- Wood E, Roundy JK, Troy TJ, van Beek R, Bierkens M, Blyth E, de Roo A, Döll P, Ek M, Famiglietti J, Gochis D, van de Giesen N, Houser P, Jaffé P, Kollet S, Lehner B, Lettenmaier DP, Peters-Lidard C, Sivapalan M, Sheffield J, Wade A, Whitehead P (2011) Hyper-resolution global land surface modeling: meeting a grand challenge for monitoring earth's terrestrial water. *Water Resour Res* 47(W05301). doi:[10.1029/2010WR010090](https://doi.org/10.1029/2010WR010090)
- Zhou Y, McLaughlin D, Entekhabi D (2006) Assessing the performance of the ensemble Kalman filter for land surface data assimilation. *Mon Weather Rev* 134:2128–2142

Surface Data Assimilation and Near-Surface Weather Prediction over Complex Terrain

Zhaoxia Pu

Abstract Owing to sparse observations, terrain misrepresentations, and complicated interactions between the atmosphere and complex terrain, numerical prediction of near-surface atmospheric conditions and surface data assimilation over mountainous regions present particularly challenging problems. With studies and results obtained from a recent field program, Mountain Terrain Atmospheric Modeling and Observations (MATERHORN), for both model evaluation and data assimilation, this chapter provides an overview and sample of these problems and challenges. Forecast evaluation for the mesoscale community Weather Research and Forecasting (WRF) model and results from an ensemble Kalman filter method for assimilating surface observations are presented.

1 Introduction

About 70 % of the Earth's land surface is characterized by complex topography, and thus atmospheric processes typical of complex terrain airsheds have been studied extensively in the context of air quality; reviews are given by Whiteman (2000), Fernando et al. (2010), and Zardi and Whiteman (2010). Yet the weather in heavily mountainous areas has received less attention. Accurate weather forecasting in complex terrain, especially in mountainous areas, presents a challenging problem in modern numerical weather prediction (NWP) due to a number of difficulties,

Contributed to Springer Book by Seon K. Park and Liang Xu (Eds.).

Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, Volume III.

February 2016.

Z. Pu (✉)

Department of Atmospheric Sciences, University of Utah,
135 S 1460 E, Rm. 819, Salt Lake City, UT 84112, USA
e-mail: Zhaoxia.Pu@utah.edu

including sparse observations, terrain misrepresentation in numerical models, and complicated interactions between the atmosphere and complex terrain.

In addition, near-surface weather forecasts are critical for the protection of life and property, economic and operational activities, and routine day-to-day planning. Aviation, military, wind energy, and energy consumption operations rely on accurate near-surface forecasts, and even small forecast errors can have major consequences. Temperature, wind, and precipitation are some of the more important variables to forecast, but visibility-reducing phenomena, such as dust, fog, and smog, also need to be accurately forecast. Therefore, improving forecasts for any of these near-surface variables has far-reaching significance.

However, previous studies of numerical models (Liu et al. 2008a, b; Hart et al. 2005; Mass et al. 2002; Zhang et al. 2013) have demonstrated the unavoidable errors of near-surface atmospheric simulation that are either related to or caused by complex terrain. Specifically, Liu et al. (2008a, b) performed a verification of model analyses and forecasts for five United States Army test and evaluation command ranges over a 5-year period. Their results indicated that forecast errors varied from range to range and from season to season, but larger errors were typically associated with complex terrain. Zhong and Fast (2003) compared three mesoscale numerical models and evaluated simulations over the Salt Lake Valley, Utah, in the United States for cases influenced by both weak and strong synoptic forcing scenarios. A cold bias was present in the valley extending from the surface to the top of the atmosphere. The simulated nocturnal inversion was much weaker than what was observed. Hanna and Yang (2001) evaluated four mesoscale model simulations of near-surface variables. They found that the models tended to predict weaker temperature inversions than observed and also underestimated the vertical temperature gradients in the lowest 100 m during the nighttime. They also suggested that the uncertainties regarding wind speed and direction in the lower atmosphere were primarily due to random turbulent processes that were not appropriately represented in the numerical model, and to errors in the subgrid terrain and land use. Zhang et al. (2013) recently conducted a comprehensive study of the error characteristics in numerical simulations of near-surface temperature and wind from the advanced research version of the Weather Research and Forecasting (WRF) model in regions of complex terrain. They found that forecasts not only suffered from the model's inability to reproduce accurate atmospheric conditions in the lower atmosphere but also struggled with representative issues due to mismatches between the model and the actual complex terrain. In addition, surface forecasts at finer resolutions did not always outperform those at coarser resolutions. *Forecast errors in near-surface variables were significant even when the WRF model was able to reproduce those weather phenomena reasonably well.*

In order to improve the forecasts of near-surface variables, accurate initial conditions are vital. One can easily believe that assimilation of surface observations is beneficial to numerical forecasts of near-surface atmospheric conditions. However, although surface observations are important for weather forecasts, their use in numerical weather prediction (NWP) has proven difficult. In particular, surface observations have not been used in traditional three-dimensional variational data

assimilation (3DVAR) methods at many operational centers (e.g., Kalnay et al. 1996; Mesinger et al. 2006; Simmons et al. 2004).

Recent studies show progress in surface data assimilation using an advanced data assimilation method, namely, the ensemble Kalman filter (EnKF). For instance, Hacker and Snyder (2005) showed that assimilation of surface observations in a one-dimensional column model resulted in error reductions throughout the atmospheric boundary layer. Ancell et al. (2011) demonstrated that surface analyses using an ensemble Kalman filtering method and subsequent short-term forecasts were generally better than surface forecasts from the NCEP Global Forecast System (GFS) and North American Mesoscale (NAM) model. More recently, Zhang and Pu (2014) have also demonstrated that assimilation of surface observations can potentially have a significant impact on the numerical prediction of landfalling hurricanes. Most importantly, using the WRF model and ensemble adjustment Kalman filter (EAKF) data assimilation for WRF, developed at NCAR (Anderson et al. 2009) with the Data Assimilation Research Testbed (WRF/DART) in an observing system simulation experiment (OSSE) framework, Pu et al. (2013) showed that EnKF performs better than 3DVAR in terms of analyses and short-term weather forecasts in complex terrain because it is more capable of handling surface data in the presence of terrain misrepresentation with Observing System Simulation Experiments (OSSEs).

In this chapter, the aforementioned problem and progress in the numerical prediction of near-surface atmospheric conditions and in surface data assimilation are *highlighted* and demonstrated by the author of this chapter and her students' research as well as results obtained from the recent field program Mountain Terrain Atmospheric Modeling and Observations (MATERHORN).

Specifically, the WRF model (Skamarock et al. 2008) and its 3DVAR data assimilation system (Barker et al. 2004), as well as WRF/DART (Anderson et al. 2009), are used to obtain the numerical prediction and data assimilation results in this chapter.

2 Characteristics of Errors in Near-Surface Atmospheric Conditions

In order to understand the characteristics of the errors in numerical weather prediction over complex terrain, Zhang et al. (2013) conducted a comprehensive study to evaluate the accuracy of numerical weather prediction, with attention to *the near-surface atmospheric conditions, specifically the 2-m temperature and 10-m wind speed and direction*, in numerical weather prediction. Numerical simulations produced by Version 3.3 of the WRF model were first evaluated for three typical weather events (i.e., a low-level jet, a cold front, and a wintertime persistent

inversion) over the Southern Great Plains (SGP) and the Intermountain West of the United States, and then for a 1-month-long forecast over a region of complex terrain.

Based on the three typical weather events, it was found that the WRF model was able to reproduce all three weather phenomena reasonably well. Verification of near-surface conditions (i.e., 2-m temperature and 10-m wind) indicated the complexity in forecasting surface variables. For the frontal case and low-level jet case over the central United States (flat terrain), the model terrain matched the actual terrain and thus mitigated representative errors. The forecasts of surface variables generally agreed well with the observations. However, errors still occurred, depending on the model's ability in forecasting the structures in the lower-atmospheric boundary layer. For the inversion case over the Salt Lake Valley (over complex terrain), different error characteristics were found over the mountain and valley stations. Forecasts not only suffered from the model's inability to reproduce accurate atmospheric conditions in the lower atmosphere but also struggled with representative issues due to mismatches between the model and the actual terrain.

In addition, it is also found that forecasts of near-surface variables at finer resolutions did not always outperform those at coarser resolutions. Increasing the vertical resolution did not help predict the near-surface variables, although it did improve the forecasts of the structure of mesoscale weather phenomena. Numerical forecasts of near-surface atmospheric conditions were sensitive to the planetary boundary layer (PBL) scheme in the WRF model, but there was no single PBL scheme that performed better than the others. More importantly, forecast errors in near-surface atmospheric variables showed flow-dependent features in all three of the individual cases when strong synoptic forcings were present.

Then, to further understand the general characteristics of the errors in near-surface forecasts in complex terrain, additional verification was conducted for forecasts over a 1-month period (from 15 September to October 2011, prior to the MATERHORN field program) in the Dugway Proving Ground (DPG), Utah. The major findings are the following: (1) The forecast errors of surface variables (2-m temperature and 10-m winds) depended to a large degree on the diurnal cycle of the surface variables themselves, especially when the synoptic forcing was weak. Specifically, the forecast errors for 2-m temperature reached two daily maxima at 0300 and 1500 local time, and two daily minima at 0700 and 1900 local time. Errors in wind speed and direction followed the same trends, with a maximum at night and a minimum in the afternoon. Forecast errors were independent of the initialization time and forecast lead time. (2) *The model forecasts showed positive (warm) temperature biases at night and negative (cold) biases during the daytime.* In contrast to the 2-m temperature, wind direction and speed had no systematic biases from a long-term perspective. (3) Under strong synoptic forcing, diurnal patterns in forecast errors were broken, while flow-dependent errors were clearly shown (See details in Zhang et al. 2013).

3 Surface Data Assimilation: 3DVAR Versus EnKF

The aforementioned results from Zhang et al. (2013) along with those of other studies (as mentioned in the introduction) illustrate the complexity and challenges involved in near-surface simulation/forecasting over complex terrain. Since surface observations are the main conventional meteorological observations, it is easy to claim that assimilation of surface observations can result in improvements in the numerical prediction of near-surface atmosphere variables. However, as shown in early studies (Kalnay et al. 1996; Mesinger et al. 2006; Simmons et al. 2004) and as overviewed by Pu et al. (2013), the traditional 3DVAR method has problems assimilating surface observations, especially 2-m temperature, into the NWP model. Thus, in both the NCEP reanalysis and ECMWF global reanalysis (ERA-40), surface data were not assimilated.

In order to examine the problem with surface data assimilation using 3DVAR and also to evaluate the ability of the ensemble Kalman filter to assimilate surface observations, Pu et al. (2013) performed a series of data assimilation experiments using a configuration of Observing System Simulation Experiments (OSSEs) with the WRF model and its data assimilation system (e.g., 3DVAR and WRF/DART, as mentioned in the introduction). A series of data assimilation experiments with both a single observation in a mountain valley and multistation observations over the continental U.S. were conducted.

For a quick background, the WRF 3DVAR system was developed based on the NCAR/Penn State University Mesoscale Model Version 5 (MM5) 3DVAR system (Barker et al. 2004). 3DVAR provides an analysis \mathbf{x}^a via the minimization of a prescribed cost function $J(x)$,

$$J(x) = J^b + J^o = \frac{1}{2}(\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1}(\mathbf{x} - \mathbf{x}^b) + \frac{1}{2} \sum_{i=1}^n (\mathbf{y} - \mathbf{y}^o)^T \mathbf{O}_i^{-1}(\mathbf{y} - \mathbf{y}^o) \quad (1)$$

In Eq. (1), the analysis \mathbf{x}^a represents an a posteriori maximum likelihood (minimum variance) estimate of the true atmospheric state given two sources of information: the background (previous forecast) \mathbf{x}^b and observations \mathbf{y}^o (Lorenc 1986). The analysis fit to these data is weighted by the estimates of their errors: \mathbf{B} and \mathbf{O} are the background and observational error covariance matrices, respectively; $\mathbf{y} = \mathbf{H}(\mathbf{x})$, and \mathbf{H} is a linear or non-linear operator projecting the grid point state \mathbf{x} to estimated observations. In many studies, the background error covariance term (\mathbf{B}) is generated with the so-called NMC (National Meteorological Center, now known as NCEP) method (Parrish and Derber 1992; Wu et al. 2002; Barker et al. 2004) to fit the specific region and season. For instance, statistics of the differences between two short-range forecasts (e.g., 24-h and 12-h) valid at 0000UTC and 1200UTC for 1 month (June 2008) are paired (i.e., a total of 60 samples) to generate the background error covariance. Since the \mathbf{B} in 3DVAR is generated by statistics, it is *static* throughout the data assimilation experiment.

In contrast to the 3DVAR method, the background error covariance term is estimated using an ensemble of forecasts in an ensemble Kalman filter. The ensemble mean is supposed to be the best estimate of the true state. The analysis is updated via the equation

$$x^a = x^f + \mathbf{K}[y^o - \mathbf{H}(x^f)] \quad (2)$$

with a Kalman gain matrix

$$\mathbf{K} = \mathbf{P}^f \mathbf{H}^T (\mathbf{H} \mathbf{P}^f \mathbf{H}^T + \mathbf{R})^{-1} \quad (3)$$

where y^o is the observation vector, and the superscripts f and a indicate, respectively, forecast (background) and analysis. \mathbf{H} is a linearized observation operator, which relates the model state x and observation y^o by

$$y = \mathbf{H}x + \epsilon \quad (4)$$

and ϵ is a Gaussian variable with mean zero and covariance \mathbf{R} . \mathbf{P}^f is the background error covariance and is estimated using an ensemble of k forecasts $x_k^f(t_i)$ (Evensen 1994).

$$\mathbf{P}^f \approx \frac{1}{K-1} \sum_{k=1}^K (x_k^f - \bar{x}^f)(x_k^f - \bar{x}^f)^T \quad (5)$$

where the overbars represent the ensemble average. As ensemble forecasts are used in generating the background error term, the background error covariance in the ensemble Kalman filter is flow-dependent.

Owing to the fundamental differences between the 3DVAR and ensemble Kalman filter methods, Pu et al. (2013) found that the two methods behave differently regarding surface data assimilation. Specifically, it is found that there are fundamental difficulties in assimilating surface observations with 3DVAR, especially when assimilating surface observations in a mountain-valley region. Results from the assimilation of observations at a single observation station demonstrate that EnKF can overcome some of the fundamental limitations of 3DVAR in assimilating surface observations over complex terrain. Specifically, through its flow-dependent background error term, EnKF produces more realistic analysis increments over complex terrain in general. Figure 1 shows sample structures of the estimated background error standard deviation of the streamfunction in 3DVAR (static) and the EnKF over the assimilation period for a case presented in Pu et al. (2013). In 3DVAR, the error variance is homogeneous in each statistical bin and has no correlation with terrain and the synoptic situation. In contrast, error variances

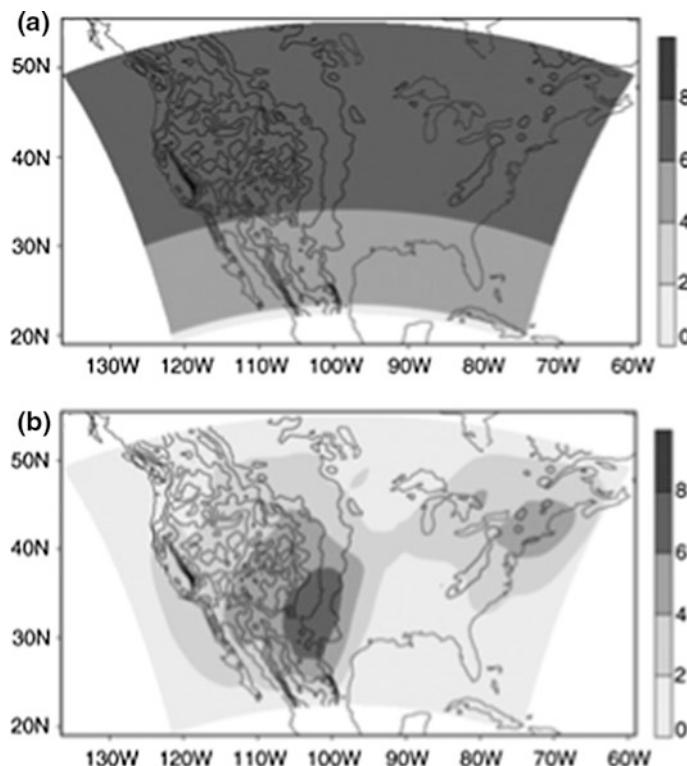


Fig. 1 Sample figure comparing the structures of the 3DVAR and the EnKF background terms. **a** Estimated background error standard deviation of the streamfunction (shaded contour; unit: $10^5 \text{ m}^2 \text{ s}^{-1}$) in 3DVAR (static in time) and **b** in EnKF [averaged over the data assimilation period (from 0000 UTC to 0600 UTC 5 June 2008)] near 800 m AGL. Contour lines denote the terrain heights (interval: 500 m). [Figure adopted from Pu et al. (2013)]

in the EnKF reflect the structure of the synoptic system. As a result, for analysis increments from a single observation in the mountain valley, the shapes of the analysis increments from 3DVAR follow classical correlation and cross-correlation functions of variables using the geostrophic increment assumption (Fig. 2). Analysis increments from the observation station within the mountain valley area have been spread across the mountains (Fig. 3). With the EnKF, the analysis increments resulting from the assimilation of the same single observation in the valley remain inside the valley (Fig. 4). No cross-mountain analysis increment is found. Since it is expected that the air temperature and wind conditions can be inhomogeneous over the mountain valley and cross-mountain areas, the cross-mountain analysis

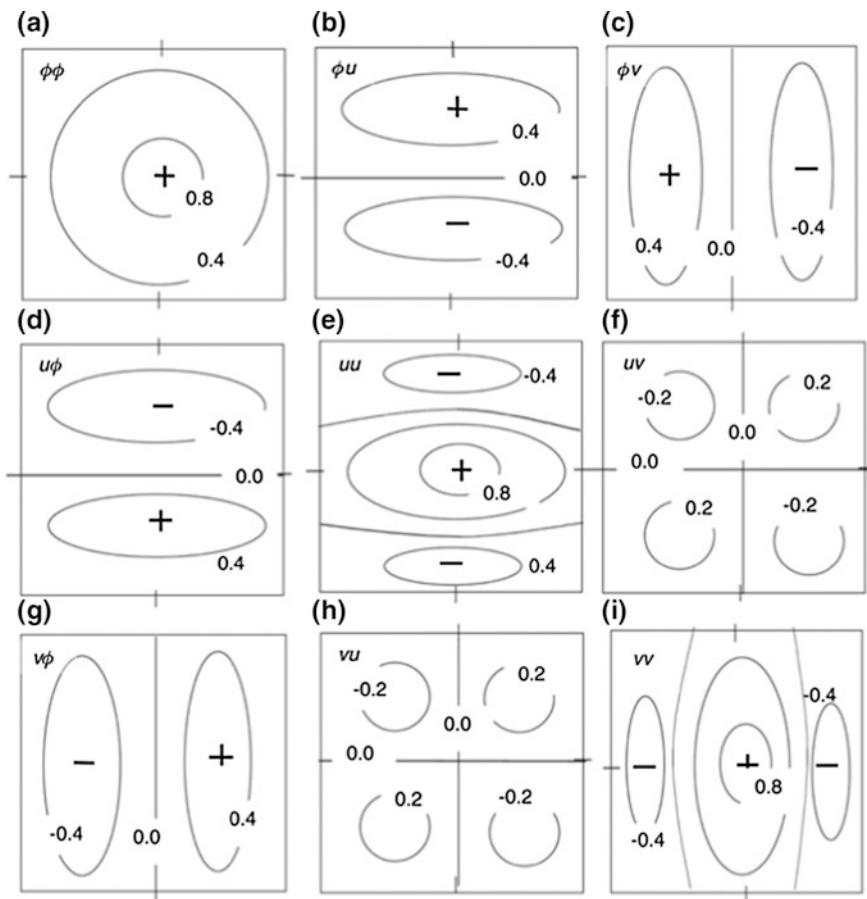


Fig. 2 Schematic illustration of the correlation and cross-correlation functions for multivariate OI analysis derived using the geostrophic increment assumption (courtesy Gustafsson 1981 and Kalnay 2003). ‘ φ ’ is a thermodynamic variable related to temperature, and u and v denote the horizontal components of wind

increments in 3DVAR could be unrealistic. The EnKF helps mitigating the problem by reproducing more reasonable analysis increments.

More comprehensive comparisons are conducted in a short-range weather forecast using synoptic cases with multiple station surface observations. EnKF is better than 3DVAR for the analysis and forecast of the weather system over flat terrain. Over complex terrain, EnKF clearly performs better than 3DVAR, because it is more capable of handling surface data in the presence of terrain misrepresentation. Results also suggest that caution is needed when dealing with errors due

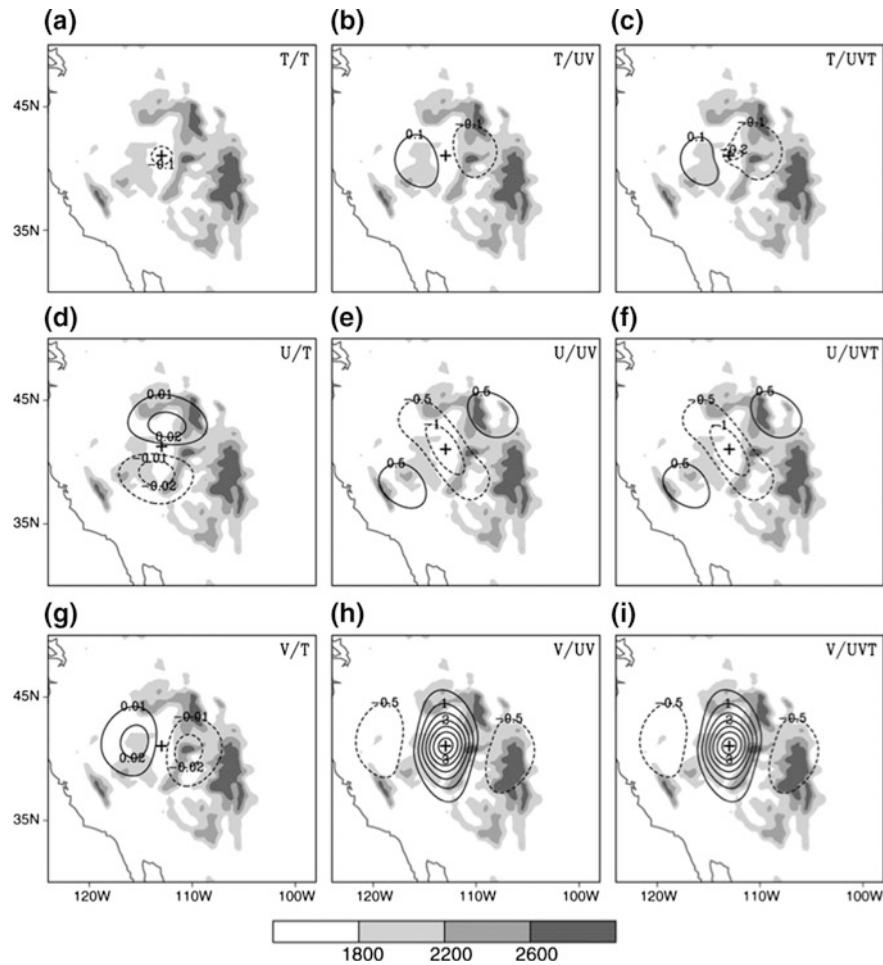


Fig. 3 3DVAR analysis increments of temperature (K; top row), u-component (m s^{-1} ; middle row), and v-component (m s^{-1} ; bottom row) of wind at the lowest model level with assimilation of 2-m temperature (left column), 10-m winds (middle column), and both 2-m temperature and 10-m winds (right column) from a single observation station over complex terrain. The shaded contours show the terrain heights (unit: m). '+' denotes the location of the observation station. [Courtesy Pu et al. 2013]

to model terrain representation. Data rejection may cause degraded forecasts because data are sparse over complex terrain (see Pu et al. 2013 for details).

These major findings and conclusions from the aforementioned early studies by Zhang et al. (2013) and Pu et al. (2013) were further corroborated during the MATERHORN field program with observations obtained.

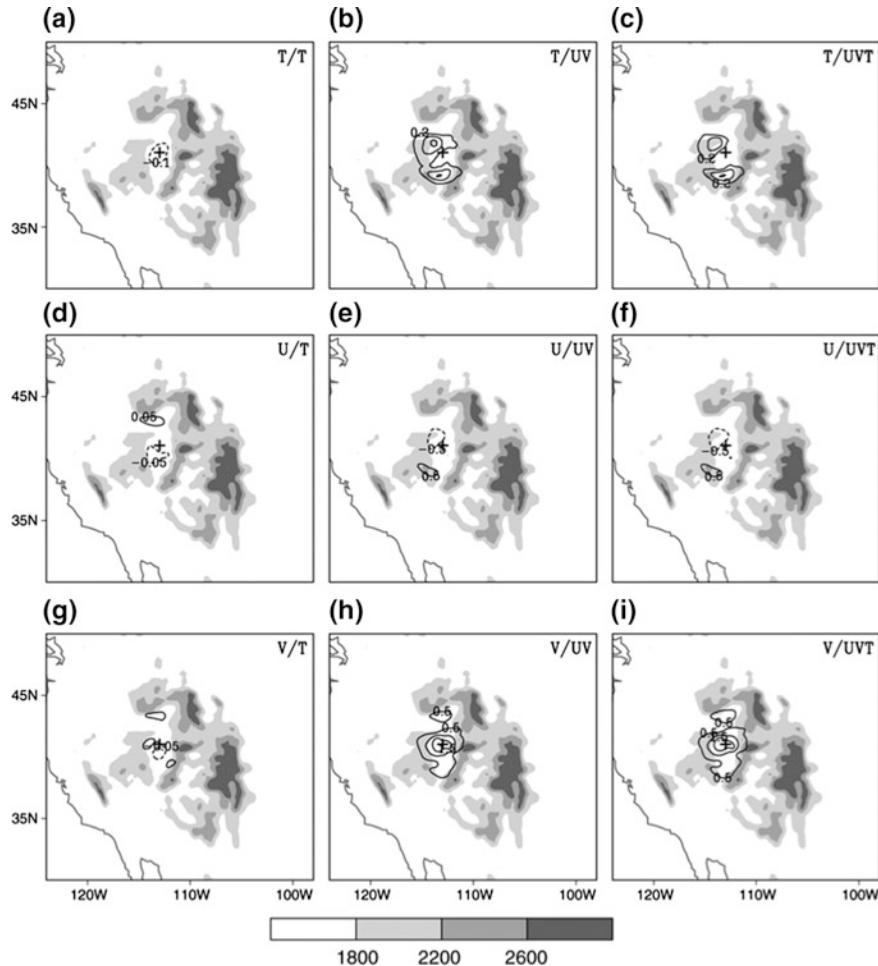


Fig. 4 Same as Fig. 3, except for the EnKF analysis increments. [Courtesy Pu et al. 2013]

4 Results from the MATERHORN Field Program

To address scientific needs and help improve the prediction of mountain weather, the US Department of Defense has funded a research effort—the Mountain Terrain Atmospheric Modeling and Observations (MATERHORN) Program—that draws on the expertise of a multidisciplinary, multi-institutional, and multinational group of researchers. The ultimate goal of MATERHORN is to identify and study the limitations of current state-of-the-science mesoscale models for mountain terrain weather prediction and develop scientific tools to help realize leaps in predictability. The program has four principal thrusts, encompassing Modeling, Experimental, Technology, and Parameterization components, directed at diagnosing model

deficiencies and critical knowledge gaps, conducting experimental studies, and developing tools for model improvements (Fernando et al. 2015).

During the MATERHORN program, two major field campaigns were conducted, from September to October 2012 (Fall Campaign) and in May 2013 (Spring Campaign), respectively, over the Dugway Proving Ground (DPG). The DPG is located approximately 80 miles southwest of Salt Lake City, Utah. It is characterized by complex terrain not only because of it is surrounded on three sides by mountain ranges but also due to the various land covers (e.g., playa and sagebrush, etc.). Figure 5a shows the DPG area map and land features as well as a total of 31 automatic surface stations in the area. Considering the height of its mountains relative to its flat regions, the DPG area is representative of common complex terrain features.

Comprehensive observations were collected of near-surface atmospheric conditions, profiling measurements from multiple platforms (e.g., tethersondes, lidar, radiosondes, etc.), soil states, and surface energy budgets during the field program. They offer a great opportunity for evaluating numerical prediction and data assimilation over complex terrain.

4.1 *Evaluation of Real-Time WRF Forecasts During MATERHORN*

A real-time forecast was performed by the author at the University of Utah with the WRF model at high resolution (\sim 1 km horizontally) four times a day (at 00, 06, 18, and 24 UTC) during both MATERHORN Fall (September to October 2012) and Spring (May 2013) campaigns. The purpose of this real-time forecasting was not only to support decision-making during the field program but also to provide a useful database to evaluate the performance of WRF model in predicting synoptic flows over mountainous terrain. During the field program, a series of 48-h forecasts were produced for over 200 forecast lead times and for all Intensive Observational Periods (IOPs) during September–October 2012 and May 2013. After the field experiments, these forecast results were compared with observations collected from the field experiments. The results presented in the following section are from the MATERHORN Fall Campaign (September to October 2012).

The real-time forecasting system was built using Version 3.3 of WRF. Four one-way nested domains, with horizontal grid spacings of 30, 10, 3.33, and 1.11 km, were used (Fig. 5b). The innermost domain (1.11 km) focused on the DPG area. The WRF model includes 41 vertical levels; the top of the model was set to 50 hPa. The physical parameterization schemes included the Kain-Fritsch cumulus scheme (for 20 and 10 km grid spacings only), the Thompson micro-physics scheme, the Rapid Radiative Transfer Model (RRTM) for longwave radiation and Dudhia for shortwave radiation, and the YSU PBL scheme (see details of the physical schemes in Skamarock et al. 2008). A topography dataset with a 30

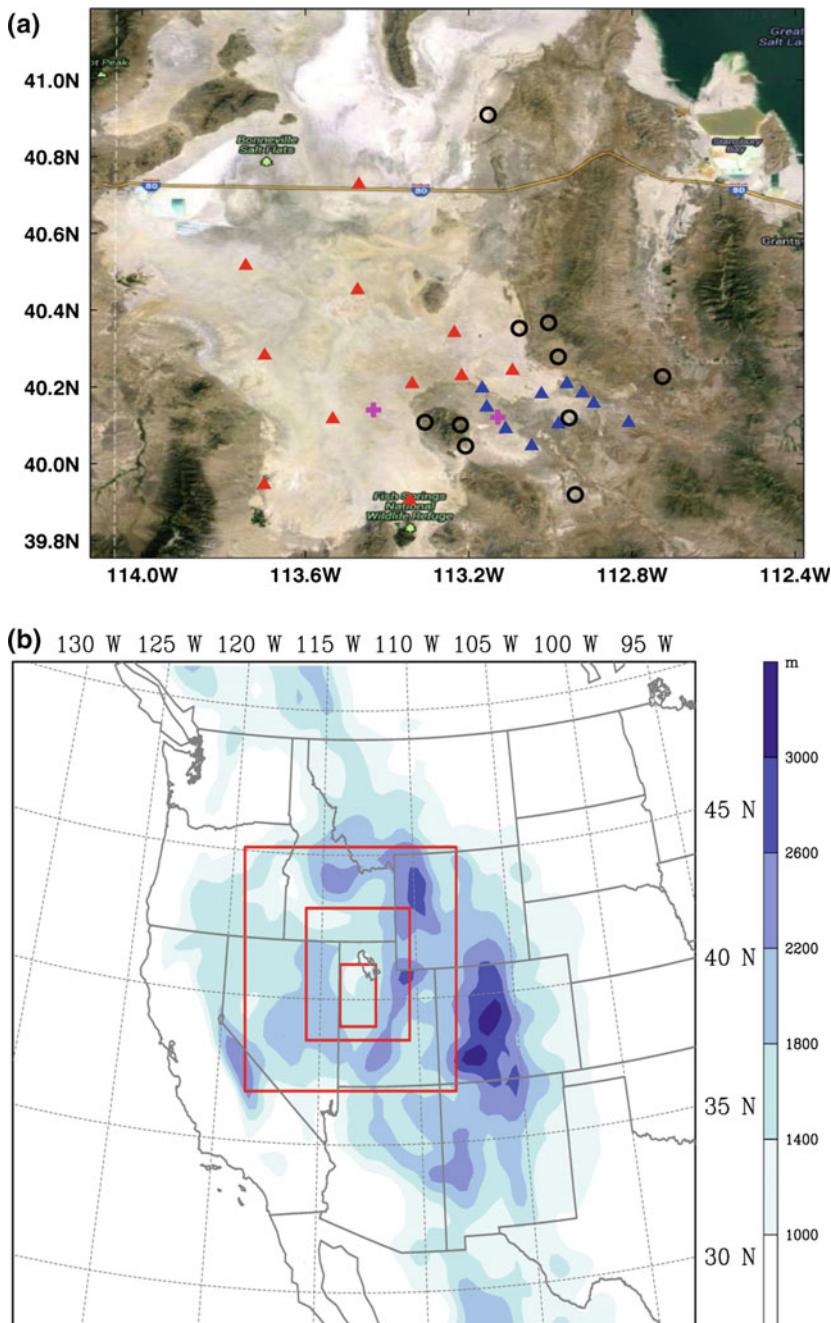


Fig. 5 **a** The area map of Dugway Proving Ground (DPG) with surface stations. Red triangles represent “Playa” stations and blue triangles represent “Sagebrush” stations. Black circles denote those stations that do not count as either Playa or Sagebrush. The two purple plus signs represent the two sounding stations during the MATERHORN field program over Playa and Sagebrush regions. **b** Locations of model domains for real-time WRF forecasting during MATERHORN. The innermost domain covers DPG

arc-second (about 1000 m) resolution and an updated land-use dataset with 27 land-use categories (instead of the 24 land-use categories provided by the WRF Version 3.3 release) from the U.S. Geological Survey (USGS) were used in order to ensure more accurate surface conditions, especially for playa and desert regions in the western United States. The Noah land surface model was used because it predicts the land states, such as surface temperature and soil moisture and temperature, in each layer with time. Initial and boundary conditions were derived from the analyses and forecasts produced by the NCEP North American Model (NAM) forecast system at 0000, 0600, 1200, and 1800 UTC. Over a 1-month period from 21 September to 20 October 2012, a total of 120 forecasts were generated.

Early evaluation of the real-time forecasts was conducted after the field program. It was found that the WRF model was capable of producing reasonable forecasts in large-scale synoptic conditions and average mesoscale conditions over the DPG region. However, forecast errors were present in small-scale and local flows and their associated near-surface conditions. Some results are highlighted in the following two subsections.

4.1.1 General Statistics and Diurnal Variations

We calculated various statistical metrics to characterize the error of near-surface variables in forecasts. For instance, root-mean-square errors (RMSEs), mean absolute errors (MAEs), and bias errors (BEs) were calculated against surface observations to evaluate the overall performance of WRF forecasts.

The statistical calculations are as follows:

$$MAE = \frac{1}{n} \sum_{i=1}^n |F_i - O_i|$$

$$BE = \frac{1}{n} \sum_{i=1}^n (F_i - O_i)$$

where i denotes the i th observation, t denotes the observation at time t , O_i represents the value of the observation at the i th location, F_i denotes the forecast value interpolated to that observation location, and n is the total number of stations.

The MAEs and mean BEs were calculated for each of the four initialization times. Specifically, all forecasts initialized at 00 UTC, 06 UTC, 12 UTC, and 18 UTC during the month were averaged over all stations to calculate the MAEs and BEs as a function of forecast lead time. Figure 6a–c show the MAEs for 2-m temperature and 10-m wind speed and direction calculated in the DPG area in the 1.11 km domain for forecasts initialized at 00 UTC, 06 UTC, 12 UTC, and 18 UTC. A clear diurnal pattern is found in the errors of all variables. Specifically, the temperature error peaks twice per day, around 0300 mountain standard time (MST) and 1500 MST (corresponding to 1000 and 2200 UTC). There are also two error minima for temperature, at around 0700 and 1900 MST (corresponding to

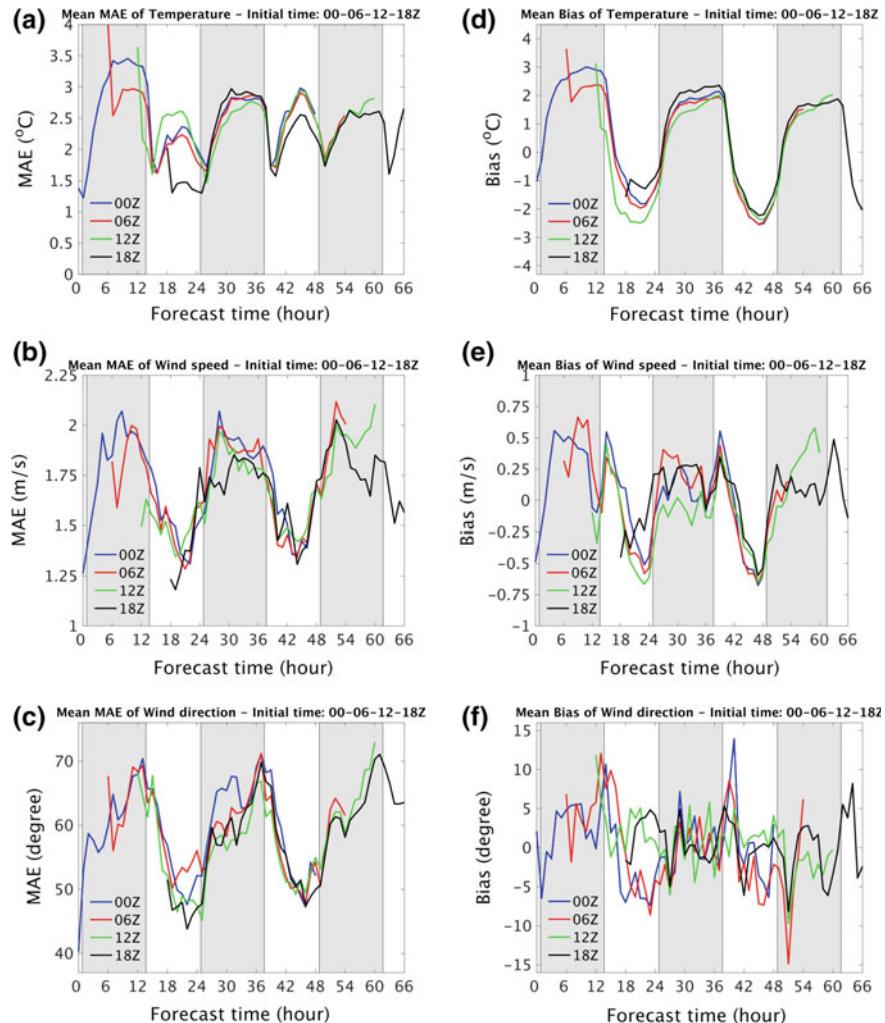


Fig. 6 Variation of mean absolute errors for **a** temperature, **b** wind speed, and **c** wind direction and mean biases for **d** temperature, **e** wind speed, and **f** wind direction along with forecast time, calculated for each forecast initial time from 00UTC, 06UTC, 12UTC, and 18 UTC and averaged over the whole month (21 September to 20 October 2012) over all surface stations

1400 and 0200 UTC). Wind speed and direction follow the same error trends, with a maximum in the early evening or before sunrise and a minimum in the afternoon. The dependence of the surface forecasts on initialization time was examined. Figures 6a-c show that the error trends are independent of initialization time and forecast lead time and follow the same diurnal variation. However, compared with the forecasts initialized during the daytime (0000 and 1800 UTC), relatively large errors occur in the first 2–3 h in 2-m temperature for the forecasts initialized at night

(0600 and 1200 UTC). The large errors in the nighttime-initiated forecasts could be caused by the erroneous surface and soil conditions in the stable boundary layer in the North American Mesoscale Forecast System (NAM) analysis. Apparently, the large errors associated with initial conditions in the nighttime-initiated forecasts do not persist beyond a few hours.

Figure 6d–f further demonstrate the diurnal patterns of the bias errors in 2-m temperature over the whole month, averaged over all stations. Positive (warm) biases are found at night and negative (cold) biases are present during the daytime. No systematic biases are found in wind direction and speed.

So far, the above results are consistent with the findings in Zhang et al. (2013), as the statistics were done in the same season. Because data are available, additional comparison is conducted.

4.1.2 Playa Versus Sagebrush

Although the real-time WRF forecasts reasonably characterize the overall statistics of the near-surface variables, because the near-surface atmospheric conditions vary with underlying land surfaces, it is very important to evaluate whether the WRF model can distinguish the characteristics of atmospheric boundary layer structures and near-surface variables in these areas. Over the DPG area, the underlying land surface is dominated by two types: playa and sagebrush. Fortunately, during the intensive observation periods (IOPs) of the field campaign, two sounding stations were operated over the DPG, one in a Playa area and the other in a Sagebrush area (Fig. 5a). There were also tethersonde balloons.

Figure 7 compares the sounding profiles from WRF forecasts and radiosonde observations of temperature and wind at 2030 UTC 3 October 2012 and 0030 UTC 7 October 2012. Although discrepancies are found in model-simulated and observed soundings, it is apparent that the model was able to distinguish the atmospheric conditions over both Sagebrush and Playa. In both sounding observations, the Sagebrush was warmer (colder) than the Playa below (above) about 750 hPa. The model was able to capture these features well, although biases were present in the temperature forecasts. Meanwhile, the model was not able to capture the warm surface layer revealed by soundings, perhaps due to the lack of vertical resolution between the surface and the lowest model level. This can be clearly seen in Fig. 8, where the model results are compared with the tethersonde balloon observations. In fact, Fig. 8 clearly reveals that the WRF model completely missed the larger gradients of temperature and wind near the bottom of the atmosphere or very close to the ground, implying that the mesoscale model is incapable of resolving the rapid change in temperature and wind conditions near the bottom of the atmosphere.

Figure 9 illustrates the overall comparison between observations from surface stations and WRF forecasts for temperature over the Playa stations and the

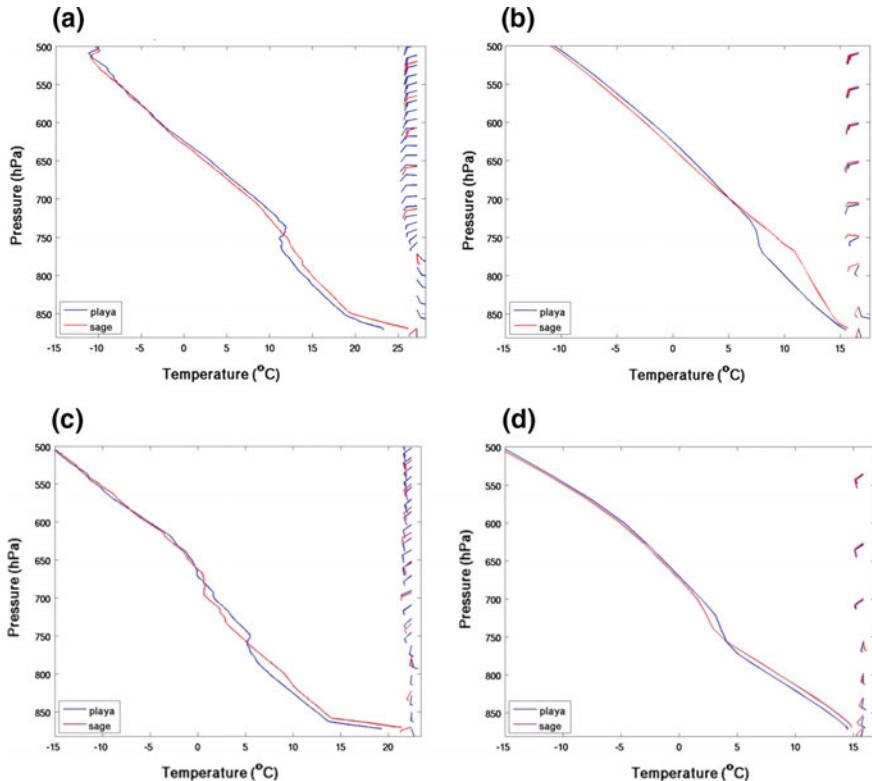


Fig. 7 Comparison of the temperature and wind profiles from (a and c) radiosonde observations and (b and d) WRF forecasts at (a and b) 2030 UTC 3 October 2012 and (c and d) 0030 UTC 7 October 2012. The *blue* and *red* curves denote the conditions over Playa and Sagebrush areas, respectively

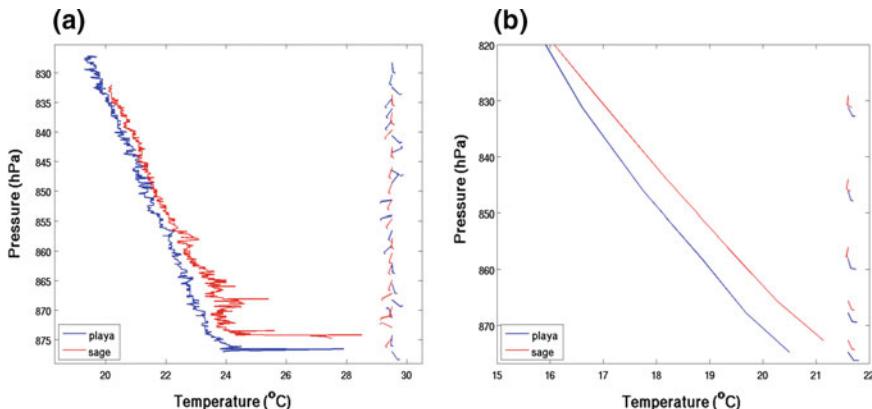


Fig. 8 Comparison of the temperature and wind profiles from **a** tethered balloon observations and **b** WRF forecast at 2000 UTC 01 October 2012. The *blue* and *red* curves denote the conditions over Playa and Sagebrush areas, respectively

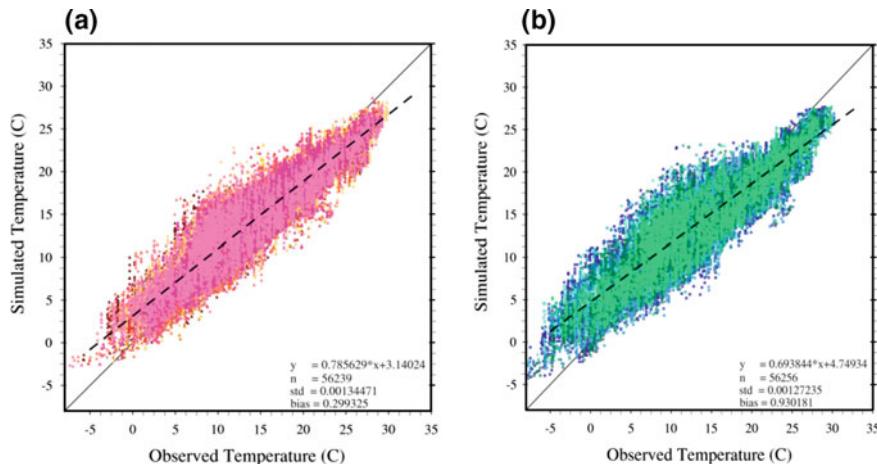


Fig. 9 Scatterplots for 2-m temperature of WRF forecast/simulations and surface observations for the **a** Playa area and **b** Sagebrush area over the whole month. In each pane, the *thin gray line* denotes $x = y$; the *dashed line* denotes a linear fit of the data

Sagebrush stations during the whole month at all forecast lead times. It indicates that the Sagebrush area has larger biases in temperature forecasts than the Playa region does, perhaps due to its complex terrain features relative to the flat Playa region.

4.2 Surface Data Assimilation with EnKF

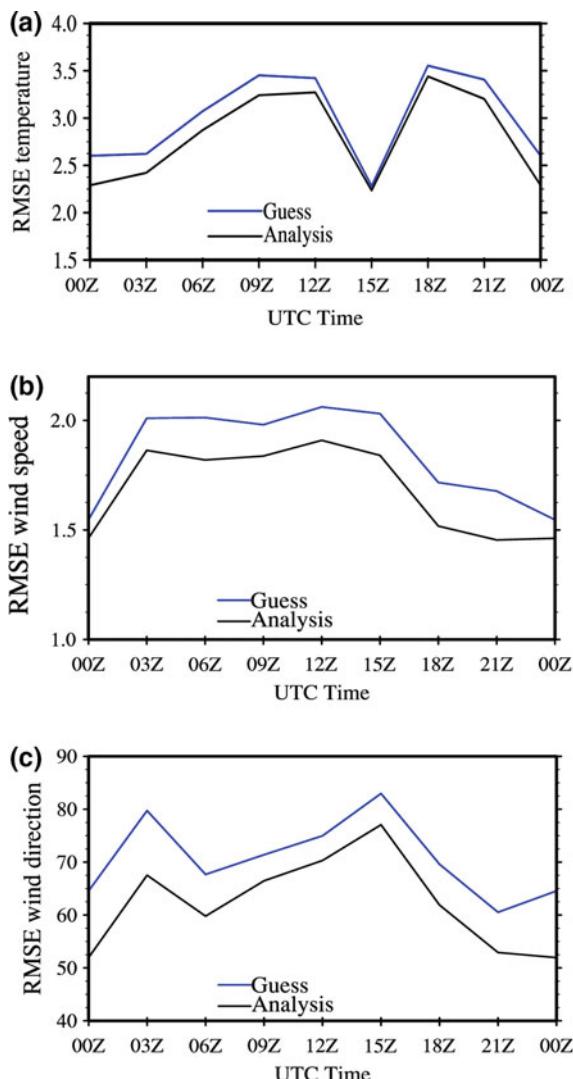
The above evaluation proves that errors in near-surface weather prediction can be serious due to complex terrain. The errors not only have diurnal patterns but also depend on the terrain features. To further understand the impact of assimilating surface observations on numerical prediction of near-surface atmospheric conditions, and also to confirm the findings from Pu et al. (2013), a month-long EnKF data assimilation was conducted for the fall 2012 MATERHORN field campaign, covering the whole period from 21 September to 21 October 2012. The first several days can be considered a spin-up period because the first IOP started at 2000 UTC 25 September.

A continuous, 3-hourly cycled EnKF data assimilation was performed for the whole period with 60 ensemble members. The initial perturbations were generated at 0000 UTC 21 September 2012 by adding ensemble perturbations to the deterministic initial conditions using fixed covariance perturbations (Torn et al. 2006). The data assimilation was conducted with the DART EnKF system (Anderson et al. 2009), and the 3-h forecasts during the cycles were integrated with the WRF model. The assimilated observations included NCEP PREBUFR (NCEP Automated Data

Processing Global Upper Air and Surface Weather Observations in the PREBUFR format) conventional observations, Mesonet surface observations (including the 31 stations over the DPG area as shown in Fig. 5a), and radiosondes from the MATERHORN field campaign (see Fig. 5a). A covariance inflation that varies temporally and spatially was used in this data assimilation system (DART/WRF) to avoid filter divergence and to reduce the impact of model error (Anderson 2007).

The RMSEs and BEs were calculated against surface observations during the 1-month period to evaluate the overall performance of the EnKF analysis. Figure 10 compares the RMSEs of the 3-hourly WRF forecast (guess field) and analyses,

Fig. 10 The diurnal variation of root-mean-square error (RMSE) of **a** temperature ($^{\circ}\text{C}$), **b** wind speed (m s^{-1}), and **c** wind direction (degrees), calculated every 3 h for EnKF analysis (black curve) and WRF 3-h forecast (Guess, blue curve) averaged over the whole month (21 September to 20 October 2012) over all surface stations and all 60 ensemble members



averaged over all stations and all ensemble members as well as during the entire evaluation period (00 UTC 25 September to 00 UTC 21 October 2012) for 10-m wind speed and direction and 2-m temperature. It is clear that the DART/WRF analyses fit the observations better than the guess field does, indicating improvement in the atmospheric state as a result of the EnKF analysis. In addition, since the model forecasts tend to underestimate the variations in temperature during the day (e.g., warm biases are present during the night and cold biases are present during the day), the positive impact of using EnKF data assimilation can at least partly overcome this problem.

Figure 11 further examines the diurnal variation of mean bias errors in 2-m temperature from the EnKF analysis. The errors were averaged over the entire evaluation period for all stations and all ensemble members. Biases over all stations in DPG, including Playa and Sagebrush, are compared. Note that although the EnKF analysis has reproduced the improved temperature analysis, by assimilating surface observations and two soundings over the area, it still cannot overcome the cold and warm biases in temperature. Meanwhile, the bias error over the Playa region is smaller than the averages over all stations and Sagebrush stations, while the bias error over the Sagebrush area is the largest. The results suggest that surface data assimilation can result in improved diurnal temperature, but it still cannot reduce the intrinsic errors and biases over complex terrain. Other factors that should be considered for further improvements are model physical parameterizations and the impact of land surface processes.

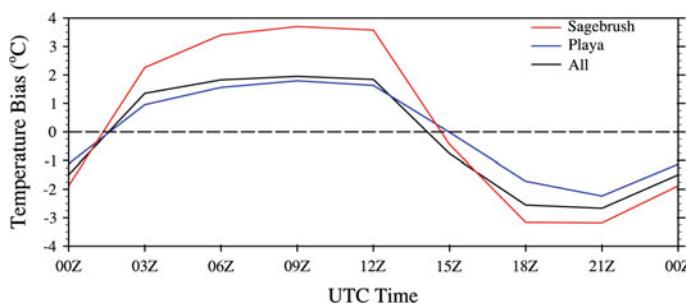


Fig. 11 The diurnal variation of mean bias errors of temperature ($^{\circ}\text{C}$) of the EnKF analysis averaged over the whole month (21 September to 20 October 2012) over all 60 ensemble members based on the average of all surface stations (All, black line), Playa stations (Playa, blue line) and Sagebrush stations (Sagebrush, red line)

5 Concluding Remarks

Numerical prediction of near-surface atmospheric conditions and surface data assimilation over mountainous regions present particularly challenging problems. This chapter first overviews the specific challenges in numerical prediction of near-surface variables and data assimilation found in the author's previous studies. Then, case studies and results obtained from the recent MATERHORN field program for both model evaluation and data assimilation are presented. It is found that the forecast errors of near-surface atmospheric conditions, such as 2-m temperature and 10-m winds, are characterized by a diurnal pattern under weak synoptic constraints. Errors are generally large during the morning and evening transition period, while at the same time the model underestimates the range of the diurnal variation in temperature, as a warm (cold) bias is present during the night (day) time. Under strong synoptic constraint cases, the errors are flow-dependent, with complicated interactions between the near-surface atmospheric conditions and the large-scale environment.

Surface observations are the main source of data over complex terrain. It is found that the traditional 3DVAR data assimilation method has difficulty assimilating surface observations over complex terrain. With its flow-dependent background covariance term, the EnKF method can overcome the difficulties experienced with the 3DVAR method. Results from a month-long EnKF analysis during the MATERHORN fall campaign not only show that EnKF is a promising method for surface data assimilation over complex terrain, but also demonstrate that surface data assimilation can result in improved numerical analysis and prediction of near-surface variables. However, it appears that data assimilation can overcome only part of the problem, as it can reduce the error only to a degree. Large errors during the morning and evening transition periods still remain. Meanwhile, due to the complex terrain with various land uses, some intrinsic errors, especially those associated with terrain features (e.g., those related to model physical parameterizations and land surface processes) could be large sources of error in the numerical prediction of near-surface weather. Therefore, even if more-comprehensive observations are obtained over complex terrain, model errors (e.g., those from physical parameterization, terrain representation, etc.) must be considered. In addition, near-surface variables are strongly influenced by underlying surface conditions and soil states (Massey et al. 2014). Therefore, future work should also emphasize improving coupling between atmospheric and land surface models.

Acknowledgements This research was supported by the Office of Naval Research Award # N00014-11-1-0709, Mountain Terrain Atmospheric Modeling and Observations (MATERHORN) Program. The author would like to express her appreciation to the MATERHORN-X team members who worked hard to collect observations during the field campaign. Early discussion with Drs. Jim Steenburgh, Sebastian Hock, Dragan Zajic, Silvana DiSabatino, and Eric Pardyjak was very helpful. Data handling and initial work by several of the author's graduate students and research associates are also appreciated. Computer support from the Center for High-Performance Computing (CHPC) at the University of Utah is gratefully acknowledged.

References

- Ancell BC, Mass CF, Hakim GJ (2011) Evaluation of surface analyses and forecasts with a multiscale ensemble Kalman filter in regions of complex terrain. *Mon Weather Rev* 139:2008–2024
- Anderson JL (2007) An adaptive covariance inflation error correction algorithm for ensemble filters. *Tellus A* 59:210–224
- Anderson JL, Hoar T, Raeder K, Liu H, Collins N, Torn N, Avellano A (2009) The data assimilation research testbed: a community facility. *Bull Am Meteor Soc* 90:1283–1296
- Barker DM, Huang W, Guo YR, Bourgeois AJ, Xiao QN (2004) A three dimensional data assimilation system for use with MM5: implementation and initial results. *Mon Weather Rev* 132:897–914
- Efron B, Tibshirani RJ (1993) An introduction to the bootstrap. Chapman and Hall, 436 pp
- Evensen G (1994) Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics. *J Geophys Res* 99:10143–10162
- Fernando HJS, Zajic D, Di Sabatino S, Dimitrova R, Hedquist B, Dallman A (2010) Flow, turbulence and pollutant dispersion in urban atmospheres. *Phys Fluids* 22:051301–051319
- Fernando HJS and coauthors (2015) The MATERHORN—unraveling the intricacies of mountain weather. *Bull Am Meteor Soc* 96:1945–1967
- Gustafsson N (1981) A review of methods for objective analysis. In: Bengtsson L, Ghil M, Källen E (eds) *Dynamic meteorology: data assimilation methods*. Springer, New York, pp 17–76
- Hacker JP, Snyder C (2005) Ensemble Kalman filter assimilation of fixed screen-height observations in a parameterized PBL. *Mon Weather Rev* 133:3260–3275
- Hanna SR, Yang R (2001) Evaluations of mesoscale models' simulations of near-surface winds, temperature gradients, and mixing depths. *J Appl Meteor* 40:1095–1104
- Kalnay E, Kanamitsu M, Kistler R, Collins W, Deaven D, Gandin L, Iredell M, Saha S, White G, Woollen J, Zhu Y, Leetmaa A, Reynolds R, Chelliah M, Ebisuzaki W, Higgins W, Janowiak J, Mo KC, Ropelewski C, Wang J, Jenne R, Joseph D (1996) The NCEP/NCAR 40-year reanalysis project. *Bull Am Meteor Soc* 77:437–471
- Kalnay E (2003) Atmospheric modeling, data assimilation and predictability. Cambridge University Press, Cambridge, UK, 341 pp
- Liu Y et al (2008a) The operational mesogamma-scale analysis and forecast system of the U.S. Army test and evaluation command. Part I: Overview of the modeling system, the forecast products, and how the products are used. *J Appl Meteor Climatol* 47:1077–1092
- Liu Y et al (2008b) The operational mesogamma-scale analysis and forecast system of the U.S. Army test and evaluation command. Part II: Interrange comparison of the accuracy of model analyses and forecasts. *J Appl Meteor Climatol* 47:1093–1104
- Lorenc A (1986) Analysis methods for numerical weather prediction. *Q J R Meteorol Soc* 112:1177–1194
- Mass CF, Ovens D, Westrick K, Colle BA (2002) Does increasing horizontal resolution produce more skillful forecasts? *Bull Am Meteor Soc* 83:407–430
- Massey JD, Steenburgh WJ, Hoch SW, Knievel JC (2014) Sensitivity of near-surface temperature forecasts to soil properties over a sparsely vegetated dryland region. *J. Appl. Meteor. Climatol.* 53:1976–1995
- Mesinger F, DiMego G, Kalnay E, Mitchell K, Shafran P, Ebisuzaki W, Jović D, Woollen J, Rogers E, Berbery E, Ek M, Fan Y, Grumbine R, Higgins W, Li H, Lin Y, Manikin G, Parrish D, Shi W (2006) North American regional reanalysis. *Bull Am Meteor Soc* 87:343–360
- Parrish DF, Derber JC (1992) The National Meteorological Center's spectral statistical interpolation analysis system. *Mon Weather Rev* 120:1747–1763
- Pu Z, Zhang H, Anderson JA (2013) Ensemble Kalman filter assimilation of near-surface observations over complex terrain: comparison with 3DVAR for short-range forecasts. *Tellus 65A*:19620

- Simmons AJ, Jones P, Bechtold V, Beljaars A, Kallberg P, Saarinen S, Uppala S, Viterbo P, Wedi N (2004) Comparison of trends and low-frequency variability in CRU, ERA-40, and NCEP/NCAR analyses of surface air temperature. *J Geophys Res* 109:D24115
- Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker M, Duda KG, Huang X-Y, Wang W, Powers JG (2008) A description of the advanced research WRF Version 3. NCAR Tech. Note, NCAR/TN-475+STR, 113 pp
- Torn RD, Hakim GJ, Snyder C (2006) Boundary conditions for limited-area ensemble Kalman filters. *Mon Weather Rev* 134:2490–2502
- Whiteman CD (2000) Mountain meteorology: fundamentals and applications. Oxford University Press, 355 pp
- Wu W-S, Purser RJ, Parrish DF (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon Weather Rev* 130:2905–2916
- Zardi D, Whiteman CD (2010) Diurnal mountain wind systems, Chapter 2. In: Mountain weather research and forecasting. Meteorological monographs. American Meteorological Society, Boston, MA
- Zhang H, Pu Z, Zhang X (2013) Examination of errors in near-surface temperature and wind from WRF numerical simulations in regions of complex terrain. *Weather Forecast* 28:893–914
- Zhang H, Pu Z (2014) Influence of assimilating surface observations on numerical prediction of landfalls of Hurricane Katrina (2005) with an ensemble Kalman filter. *Mon Weather Rev* 142:2915–2934
- Zhong S, Fast J (2003) An evaluation of the MM5, RAMS, and Meso-Eta models at subkilometer resolution using VTMX field campaign data in the Salt Lake Valley. *Mon Weather Rev* 131:1301–1322

Recent Developments in Bottom Topography Mapping Using Inverse Methods

Edward D. Zaron

Abstract The problem of identification and mapping of underwater topography, in the form of river channel depth, littoral zone depth profiles, and spatially-resolved river, estuary, and ocean bottom topography, has received attention in recent years in tandem with the increasing availability of remotely-sensed data for hydrologic and hydrodynamic modeling. A variety of inverse methods have been successfully applied in order to estimate the bottom topography from diverse data, typically by using variants of the ensemble extended Kalman filter, but variational methods and non-parametric filters have also been used. The types of measurements used include remotely-sensed and in situ water level, surface currents, surface wave celerity, and measurements of surface wave direction and wave breaking. The dynamics employed to relate bottom depth to the measured variables have, to date, been based on the vertically integrated shallow equations, the Saint-Venant equations, with either Chézy or Manning frictional representation; and coastal zone applications have additionally coupled these dynamics with the wave radiation stress and dissipation from models of phase-averaged surface waves. The relevance of three-dimensional dynamics associated with vertical shear and baroclinicity are recognized but not yet incorporated into inverse methods for topographic estimation. A scale analysis of the shallow water equations is proposed as a guide to understanding how the dynamics, spatial correlation scales, and data types are related to length and time scales of the given application.

1 Introduction

Underwater topography, i.e., the topography of the sea floor or river bed, plays a key role in the kinematics and dynamics of rivers, estuaries, and the oceans. Knowledge of the bottom topography is foundational data for geologists who seek to understand

E.D. Zaron (✉)
Department of Civil and Environmental Engineering,
Portland State University, Portland, OR 97208-0751, USA
e-mail: ezaron@pdx.edu

the history of the Earth, geodesists who chart the shape and gravity field of the Earth, and hydrodynamic modelers who seek to understand and predict the movement of water on the Earth's surface. Practical applications of maps of underwater topography are numerous and include shipping and navigation, river and flood management, and construction and geotechnical engineering.

The above-mentioned applications highlight the importance of bottom topography and motivate the development of advanced mapping techniques for circumstances where depth cannot be directly measured. For example, the planned Surface Water Ocean Topography wide-swath altimeter mission (SWOT) will conduct near-global observations of water levels from space, making measurements of many rivers not previously gauged (Pavelsky et al. 2014). In order to make the water level measurements useful to hydrologists, dynamical models are needed to relate river flow rate to water level, and water depth, i.e., surface level minus bottom topography, is a key quantity. Thus, there is a need to use remotely sensed observations to determine bottom topography. Other applications concern knowledge of time-variable topography, for example, the morphodynamic response of sand bar systems in the coastal ocean to the passage of storms, where conducting repeated *in situ* surveys may be prohibitively expensive. These applications motivate the development of inverse methods, in which the relationships between water level, waves, currents, and bottom topography are used to infer topography from remote sensing or a combination of remotely-sensed and *in situ* data.

Water depth and the shape of the sea floor or river bed provide kinematic constraints and can influence the flow through vorticity dynamics and boundary layer processes. This review emphasizes problems involving assimilation of remote sensing data into models based on fluid kinematics and dynamics. Approaches to remote identification of underwater topography using gravimetric techniques are not discussed (e.g., Smith and Sandwell 1994).

As a result of the diversity of applications and dynamical processes connected with topography, the literature on bottom topography estimation by inverse methods is spread over a wide range of research journals. Because of this, and the sometimes different nomenclature referring to bed level, bathymetry, and bottom topography, it can be difficult to translate the findings or methodologies from one domain (e.g., river hydrology) to another (e.g., oceanography). For example, within oceanography the slope of the water surface is typically much less than the slope of the seafloor topography, and it is common to reference height data to a reference ellipsoid or geoid. In contrast, many hydrologic applications involve rivers where the slope of the water surface is nearly equal to the slope of the river bed, and elevations are frequently referenced to a local datum. The use of different reference surfaces for vertical datums can be a source of misunderstanding among researchers in different fields. The purpose of the present review is to identify the commonalities and unique features of the bottom topography estimation problem in the recent literature. By comparing the approaches used in the different application domains, it should make it clearer how findings may be interpreted and generalized.

The paper is organized by reviewing recent literature concerning the topographic estimation problem in rivers, in estuaries, in the nearshore zone, and in the coastal

ocean. In each case the main attributes and methodologies are summarized in order to indicate the purposes for estimation of bottom topography, the primary types of data used, and the assimilation or estimation methodologies employed. In the Discussion section a regime diagram is proposed as a guide to understanding the relationships between dynamics, observed data, and spatial correlation scales. Finally, some common themes are highlighted and areas for future research are suggested.

2 Bottom Topography Estimation in Various Domains

There is more than one way to categorize recent work on the bottom identification problem. Our concern is with inverse methods in which the underwater topography is inferred by combining kinematic and/or dynamical models with measurements of water velocity, volume transport, or water surface elevation. Thus, the dynamical assumptions and type of data assimilated provide natural categories for organization. Of course, the dynamical assumptions are closely tied to the range of space and time scales of concern, which serve to distinguish one application area from another. A closely-related issue concerns the purpose for identification of bottom topography. In some instances, the topography is simply treated as a distributed control parameter which, along with other parameters such as bottom roughness, boundary conditions, etc., is to be calibrated in order to obtain an optimized hydrodynamic forecast model. In other instances, the topography may be of interest in itself, such as in studies of beach morphodynamics. Finally, for applications concerned with navigation, the depth envelope containing the shallowest topography will be of interest, rather than the bottom topography *per se*.

The assumed dynamics of the body of water under consideration, the type and quantity of the data used, and the intended application influence how the topographic estimation problem is posed. For example, for making flow estimates from satellite observations of water level in a coastal river, it may be perfectly acceptable to use a one-dimensional continuity equation coupled with an assumed balance between the along-channel pressure gradient and bottom stress. However, this approach would be completely inappropriate if it was desired to obtain a space-and-time resolved picture of the flow, and cross-channel topography, for the purpose of river navigation, using, say, remotely-sensed maps of water currents. For these reasons, the discussion below is organized around specific applications, ranging from river reach-scale to the coastal ocean.

2.1 *Rivers*

There are three main purposes for estimating river bathymetry, each associated with different requirements for resolution and accuracy of the final product:

- Topography for hydrologic monitoring and discharge modelling. The bottom depth is one of several parameters which must be known in order to make accurate estimates of river discharge.
- Topography for hydrodynamic modeling of flood planes or spatially-resolved river flows. Topography, topographic slopes, and frictional parameters are needed for accurate modeling of time-dependent flooding events and land-surface runoff.
- Topography for navigation or underwater morphology. Spatially-resolved underwater topography is needed for purposes that might involve safe transport or multi-purpose environmental modeling.

There is a large and growing literature on flow estimation and bottom topography determination in inland rivers, much of which is connected with planning for the SWOT mission (expected launch date in 2020). The key question uniting this work is how to estimate surface water storage and transport using the measurements of water surface elevation anticipated from the mission. Maps of water elevations are expected at a resolution of 100 m with centimeter vertical accuracy when averaged over 1 km² (Rodriguez 2015). The near-global spatial coverage and temporal resolution of SWOT observations, 1 to 4 times per 22-day repeat cycle depending on location, will provide an unprecedented view of inland water surfaces from space (Biancamaria et al. 2010).

One starting point for the utilization of SWOT or SWOT-like data are the hydraulic models consisting of a constitutive relation between discharge, Q , and remotely-sensed parameters such as the river width W , water depth D , and channel slope S , assumed to equal the water surface slope. For example, using a large sample of training data, Bjerklie et al. (2005) used regression to determine a model,

$$Q = 7.22W^{1.02}D^{1.74}S^{0.35}, \quad (1)$$

which showed skill for making aggregate, area average, discharge estimates with an accuracy of 10 %; however, the performance at specific locations was much poorer. This particular constitutive relation is dimensionally inhomogeneous, i.e., the values of the exponents and leading coefficient depend on the system of units employed, and this serves to highlight the fact that the underlying physics depend on additional parameters which cannot be represented in hydraulic relations of the form (1), and the lumped nature of these models limits their use for spatially-resolved estimates of bottom topography. Nonetheless, there is a large body of literature that seeks to use similar approaches to identification of discharge and depth from remotely sensed data.

A recent article founded in the above approach is Gleason and Smith (2014), where hydraulic geometry is used to define dimensionally homogeneous scaling laws among Q , W , D , and water velocity (V),

$$W = aQ^b, \quad D = cQ^f, \quad V = kQ^m, \quad (2)$$

for unknown coefficients, a, b, c, f, k , and m . Mass conservation, $Q = WDV$, is used to provide constraints, $ack = 1$ and $b + f + m = 1$. The authors note previously overlooked correlations between the multiplicative coefficients and the exponents, e.g., between a and b , which effectively reduces by half the number of model parameters to be identified. These correlations, together with the mass conservation along river reaches between elevation measurements, provide sufficient data for parameter estimation. The authors successfully employ a genetic algorithm to identify model parameters leading to time-variable discharge estimates with 20 to 40 % accuracy, entirely from the remotely sensed data (river width). Although it is not the focus of the work, one can surmise the reach-average depth estimates obtained from the model would be of similar accuracy.

Another line of work uses dynamical constraints to describe river flow. Within the hydraulic literature, the cross-sectionally averaged shallow water equations, the Saint-Venant equations (Chow 1959), form the basis for dynamical models. These equations may be expressed as,

$$\frac{\partial D}{\partial t} + \frac{\partial Q}{\partial x} = 0 \quad (3)$$

$$\frac{\partial V}{\partial t} + V \frac{\partial V}{\partial x} = -g \frac{\partial D}{\partial x} - \frac{\tau_w}{\rho R_H} + gS_0, \quad (4)$$

where g is gravitational acceleration, τ_w is the cross-sectional average wall shear stress (the frictional stress along the bottom and sides of the channel), ρ is water density, R_H is the hydraulic radius (the cross-sectional area divided by the wetted perimeter; approximately equal to water depth for wide rectangular channels), and S_0 is bed slope. Unlike most oceanographic applications, the pressure gradient force, which depends only the slope of the free surface, is expressed as a sum of two terms, the gradient of the water depth and the gradient of the bed. Frictional parameters are included in the parameterization of the wall stress, according to either the Manning or Chézy parameterizations (Chow 1959), the latter being equivalent to the quadratic bottom drag law commonly in oceanic applications (Gill 1982). Approximations employed in the river modeling literature include the “diffusive wave” approximation in which the momentum equation is simplified by neglecting the Lagrangian acceleration, leaving a balance between the pressure gradient and the wall stress, and is valid for steady low-Froude-number flows. The “kinematic wave” approximation involves the further neglect of $\partial D/\partial x$ by assuming the free surface is parallel to the bottom.

Several studies have applied variational assimilation, the Kalman filter and their extensions to the parameter identification problem for the Saint-Venant equations and its simplifications, seeking to identify river transport, friction parameters (e.g., roughness), and water depth from remotely sensed data. For example, Roux and Darたus (2005) investigated different formulations of variational assimilation (nonlinear optimization given all data at once) and the extended Kalman filter (sequential assimilation) for the steady Saint-Venant equations in the low-Froude-number limit, using the Manning frictional parameterization (quadratic drag coefficient proportional to

$R_H^{-1/3}$). Using an identical-twin strategy to validate the methodology, river width data were assimilated into a one-dimensional river model. The authors investigated different approaches to weighting data in the cost function defining the variational methods and found that the usual (unweighted) minimum variance criteria performed better than weighting by the inverse of the either the model forecast or the data value. The extended Kalman filter was also useful for reconstructing the flow, bottom depth, and frictional parameters; however, dependence on the sequence of data assimilated was noted.

In a more recent application Yoon et al. (2012) used an ensemble smoother, the local ensemble batch smoother, to assimilate SWOT-like measurements of river slope, width, and water surface elevation into a realistic branched river and floodplain model employing diffusive wave dynamics. The local ensemble batch smoother used was essentially a localized ensemble Kalman smoother (Hamill et al. 2001; Evensen 2006), with observations batched within a moving smoother window (Dunne and Entekhabi 2005). One noteworthy aspect of this work is the explicit attention given to modeling the spatial statistics of the unknown topography. They use an exponential model with very long, 100 km, correlation scale and spatially uniform variance. Simplification is achieved by treating the Manning friction coefficient as spatially constant. With these assumptions the topography obtained within the one-dimensional model showed good convergence to the true topography as more data were ingested by the system, corresponding to multiple satellite overpasses. Their paper also highlights the significance of boundary condition noise, which in this case results from discharge estimates provided by a larger-scale precipitation runoff model.

Garambois and Monnier (2015) present a one-dimensional analysis of the kinematic wave equation through synthetic data experiments. Their approach combines a one-dimensional mass conservation equation with dynamics reduced to a balance between the pressure gradient and the bottom friction,

$$\frac{\partial Q}{\partial x} = 0 \quad \text{and} \quad Q = KAR_H^{2/3} \sqrt{S_0}, \quad (5)$$

where K is the Strickler-Manning roughness coefficient ($\text{m}^{1/3}\text{s}^{-1}$), A is the cross-sectional area, and other terms are as used previously. Three key simplifications permit independent identification of roughness and water depth. First, the hydraulic radius is eliminated in favor of cross sectional area, A , and river width, W , and the defining relation between A and W is used,

$$A(h) = A_0 + \int_{D_0}^D W(h') dh', \quad (6)$$

where A_0 is the wetted area below the lowest observed water level D_0 . The second simplification is that roughness K is assumed to be spatially constant. Lastly, it is assumed that the channel cross section is rectangular, i.e., W is independent of depth. With the above representation, Garambois and Monnier (2015) are able to

unambiguously identify the model parameters, K , Q , and bathymetry, D , from synthetic SWOT-like observations of water level, along-channel water slope, and river width.

Honnorat et al. (2009) use a variational approach to identify spatially-resolved river bottom topography within a two-dimensional shallow water model using the Manning representation of bottom stress. Their focus is on assimilation of Lagrangian particle trajectories, as well as Eulerian velocity measurements, within an idealized 100 m \times 16 m rectangular channel with a 0.25 m bump of width 30 m on the bottom, uniform in the across-channel direction. The variation formulation of Honnorat et al. (2009) also imposes a smoothness penalty on the gradient of the topography, without explicitly modeling its spatial statistics. The authors find that observations of water depth, alone, are not sufficient to determine the topography, while their combination with trajectory observations makes it possible to significantly improve the issue.

The follow-up work of Honnorat et al. (2010) is noteworthy in that it presents one of the few applications using real-world data acquired in a laboratory flume. The results illustrate the degree to which the three-dimensional flows around the submerged topography lead to systematic errors in the model dynamics, and, consequently, errors in estimated topography. In the case studied, which involved flow over a 4 cm weir in a 10 cm-deep channel, the flow separation upstream and recirculation downstream of the weir led to identification of an “equivalent” topography which apparently smoothed over the actual weir shape and encompassed the three-dimensional flow region. Significant reduction of the estimated Manning drag coefficient was also associated with these regions. Thus, even in this relatively shallow experimental setup, the impacts of three-dimensional flow dynamics were quite significant to topography identification with a shallow water model. Notably, the three-dimensional recirculation zone persisted downstream a distance equal to about ten times the weir height.

Using a variational approach, Zaron et al. (2011) developed the generalized inverse of a shallow-water model which included the bottom depth as a distributed control parameter. The model domain consisted of a several hundred kilometer long stretch of the Hudson River and assimilated remotely-sensed near-surface velocity data confined to one region within the river, Haverstraw Bay. Although tidal currents were reversing and variable within the Bay, it was found that data from maximum current periods had the largest impact on the topographic estimation, which was explained by analysis of terms in the Euler-Lagrange equations for the topographic estimator. Also, the accuracy of along- and across-channel topographic estimates depended on making a plausible estimate of the spatial statistics of the topography, including unequal spatial correlation scales in the along- and across-channel directions. It was also important to account for vertical shear of the flow to transform the measured near-surface velocities into the depth-averaged velocities governed by the shallow water dynamics. In this case, both in situ observations and a validated three-dimensional model were available to develop the needed transformations.

Wilson and Özkan-Haller (2012) apply the ensemble Kalman filter to a hydrodynamic model in which the prognostic state variables (D and Q) are augmented with the bottom depth. Their application is based on shallow water dynamics simulated

by the Regional Ocean Modeling System (ROMS; Shchepetkin and McWilliams 2005), making the quasi-steady assumption (flow in equilibrium with prescribed time-variable boundary conditions), with Chézy bottom drag and spatially constant drag coefficient. This paper also conducts a set of observing array experiments to compare the efficacy of observations in various configurations. Not surprisingly, more data leads to better topographic estimates. A detailed analysis of a purely kinematic model is used to assess ensemble size for prescribed correlation scales of model errors. Landon et al. (2014) went on to apply the same ensemble approach with ROMS to assimilate GPS drifter tracks in the Kootenai River, Idaho, and demonstrate accurate topographic reconstructions in a realistic setting. They also examine carefully the sensitivity of their results to the number of ensemble members, ensemble localization, data quantity, and data accuracy.

Simeonov et al. (2013) demonstrate that the combination of water level and velocity observations can be utilized to calibrate not only bed friction but also datum offsets and discharge. They use a two-dimensional shallow water model for a 70 km reach of the Kootenai River, Idaho, as the basis for the experiments. The bottom drag and datum offset are calibrated by a systematic trial-and-error approach which permits intercomparison of different weighting schemes in the objective function. The topographic correction consists of a single datum offset. The other aspect considered in detail is the relationship between the remotely sensed surface current and the vertical average current which is simulated in the hydrodynamic model. The calibration between these quantities depends on the vertical distribution of flow which varies spatially. The practical difficulties involved in determining these calibration factors are analyzed with in situ transects and are found to be caused by the repeatability of boat tracks over the irregular river bed, the vertical averaging of the ADCP, and the side-lobe interference that leads to loss of in situ data near the bottom. Simeonov et al. (2013) also carefully consider data accuracy and model errors in order to weight terms in the objective function. They find that the calibrated drag coefficient agrees well with an independent estimate based solely on the in situ ADCP data which further validates their approach.

2.2 *Estuaries*

Bottom topography is important to estuary dynamics, and inaccurate topography, even in well-surveyed areas, can be a leading cause of errors in forecast models (Blumberg and Georgas 2008). Consequently, many studies of estuarine dynamics involve careful development of bathymetric grids, including calibration of topography and roughness coefficients. Thus, while determination of bottom topography is not usually the end goal, it is typically regarded as one necessary step in the development of useful estuarine models. Although formal inverse or data assimilation methods have not been widely used in an estuarine modeling context, some studies of model calibration provide a useful context for understanding the topographic estimation problem.

For example, Cea and French (2012) used a Monte Carlo approach to study sensitivity to topography and bottom roughness parameters in a two-dimensional shallow water model of the Crouch-Roache estuary, Essex, UK. By carefully analyzing the sources of error in bottom topography they developed a decomposition of topography errors in terms of an overall datum offset, a regional perturbation, and a “dynamic error” related to the physics of the mobile bed. This parameterization of the magnitude of the bottom topography errors provided a spatially variable scaling for the Monte Carlo perturbations as well as criteria for the definition of morphological zones within the estuary.

Falcao et al. (2013) address the problem of merging multiple bathymetric (under-water elevation) and topographic (land elevation) models, consisting of multiple elevation sub-models and raw data, for the purpose of tide and inundation modeling within an estuary using the ADCIRC model (Luettich et al. 1991). While not specifically addressing the inverse problem, their analysis and discussion highlight key areas which must be addressed in blending disparate overlapping topography/bathymetry data and in modeling spatial error statistics. One key issue is the harmonization of vertical datums; in their case the topographic (land surface) elevations were relative to mean sea level (MSL) at a specific tide gauge (Cascais) and historical period (1882–1938), while other elevations were expressed with respect to either a MSL field or with respect to a chart datum. Also, in this application, the mapped bathymetry was to be merged with recently collected GPS water level data reported relative to the WGS84 ellipsoid. Because the dynamically significant component of the water surface slope is the slope relative to a gravitational equipotential surface, meaningfully blending these data required the use of a geoid model to reference the GPS data to the MSL datum. Finally, the authors analyzed the significance of datum offsets by conducting a small Monte Carlo study in which the model was run with artificially introduced offsets. Although the results of these studies cannot be easily generalized, they provide useful case studies for the optimization and blending of topography from diverse sources.

2.3 *Nearshore/Surf Zone*

In the nearshore or surf zone wave-averaged flows are coupled to surface wave processes through the Stokes drift and wave set up pressure gradient, through bottom boundary layer processes where the friction velocity is determined by both the wave-orbital and wave-averaged velocities, and through dissipation driven by breaking waves. Likewise, the wave field is influenced by refraction, straining, and shearing by the wave-average current. And both wave and wave-averaged flows are influenced by bottom topography. The significant role of topography is demonstrated in the work of Plant et al. (2009), in which a series of numerical experiments with a coupled wave-flow model were conducted to examine the sensitivity to bathymetric filtering and grid resolution. The authors found that a comparison of maximum-resolution versus intentionally-degraded-resolution model runs was effective at identifying sensitivi-

ties in practice. Plant et al. (2009) concluded that wave processes are most sensitive to cross-shore topographic variability, sand bar location and profiles, while wave-averaged flow is more sensitive to larger-scale alongshore topographic variability.

One attribute of the nearshore region that distinguishes it from rivers and estuaries is the availability of diverse types of remotely-sensed wave data which are usefully related to bottom topography (Holthuijsen 1983; Holland et al. 2001; Holman and Haller 2013). For example, the frequency, σ , and wavenumber, k , of linearly propagating waves can be inferred from visible imagery and are related by a dispersion relation,

$$\sigma^2 = gk \tanh(kD), \quad (7)$$

which directly involves the bottom depth, D . Other observables, particularly the spatial pattern of breaking waves, from which wave dissipation can be estimated, are also strongly related to bottom depth.

The Beach Wizard software described in van Dongeren et al. (2008) exemplifies a novel approach to computing bathymetry, possibly time-variable, in a nearshore model. The estimation methodology is analogous to the “nudging” formerly employed in the atmospheric data assimilation community (Zou et al. 1992). Diverse data may be employed, such as wave celerity and dissipation estimated from remote sensing, and spatially local relationships between the observations and control parameters are employed to nudge model parameters towards agreement with observations. The approach is apparently successful if data are sufficiently dense in space and time.

Insight has been gained from the analysis of simplified models of surf zone dynamics in which wave steepening and dissipation are strongly coupled to the bathymetry. The one-dimensional model of wave propagation based on conservation of wave energy is similar to the one-dimensional kinematic wave models discussed above in the riverine context. In river models the volume transport, Q , is approximately conserved, while in nearshore models the wave energy flux, $F = C_g E$, is approximately conserved, where E is wave energy, and C_g is group velocity. Sources and sinks of wave energy are strongly nonlinear functions of water depth, wave height, and other factors (Thornton and Guza 1982). Dynamical nonlinearities lead to statistical nonlinearities and motivate the use of data assimilation techniques which do not rely on assumed normality of errors. Plant and Holland (2011) developed a discrete Bayesian network model based on the Thornton and Guza (1982) dynamics for three spatial locations across shore. The extreme reduction in spatial resolution made it computationally feasible to implement the Bayesian network to describe the coupled probability density of the entire suite of model boundary conditions, prognostic outputs, and parameters. Because of the extreme reduction in spatial dimensionality and the need for an extensive Bayesian network training data set, this technique is unlikely to be applicable to mapping real-work bathymetry. Nonetheless, the authors’ approach is noteworthy for the careful analysis of error propagation in this model.

The mapping of nearshore bathymetry using wave celerity measurements relies on the spatially local relationship between water depth and wave speed (7). Holman

et al. (2011) investigate the issues and methods for estimating nearshore bathymetry using time series imagery from small un-manned aircraft systems. Implementations of this approach by Holman and colleagues use a Fourier transform in time, and cross-correlation analysis in space, to optimize the spatial resolution of the D estimates (Plant et al. 2008; Holman et al. 2013). The analysis of Holman et al. (2011) finds that the most significant difficulty in making these observations from aircraft is geo-referencing of the acquired imagery, which limits applications to regions where the coastline or other fixed control points are visible.

Spatially resolved coupled wave and flow models have recently been applied in the nearshore. Dynamics involve simplified, one-way, coupling of wave radiation stresses and dissipation to the flow model, without feedbacks of currents on the wave dynamics. The series of papers by Wilson and collaborators (Wilson et al. 2010, 2014) applies the ensemble Kalman Filter to demonstrate the feasibility of developing a nearshore forecasting model solely with information obtained via remote sensing. Assimilated data included horizontal surface currents, wave celerity, and shoreline location from optical remote sensing; currents from infrared remote sensing; and wave celerity from X-band marine radar; with wave celerity measurements being the most spatially extensive. The coupled model was able to provide a dynamical basis for fusing the diverse measurements to produce improved bathymetry, which resulted in increased skill for current predictions.

A variational approach to identification of nearshore topography was taken by Kurapov and Özkan-Haller (2013). Once again, the dynamical system involved a shallow water flow model coupled with a phase-averaged wave model with one-way coupling from the wave model to the flow model. Experiments were conducted which illustrated the efficacy of spatially gridded velocity data for identification of topography, in both weakly- and strongly-nonlinear flow regimes. The particular solution method for the variational problem permitted an analysis of resolution and conditioning of the inverse, which showed how wave dynamics led to identification of the cross-shore structure of topography, while the flow dynamics constrained the along-shore structure. Modeling the spatial statistics of the unknown topography remains a key issue; although, with sufficiently dense data it appears that results are not sensitive to spatial covariance models employed.

3 Discussion

This review has outlined the diverse approaches to bottom topography estimation found in the recent literature. From this survey of applications it is apparent that both ensemble-based and variational approaches have utility. Model resolution or the number of spatial degrees of freedom is an important issue in relation to whether topography alone or topography and a friction parameter are to be determined. The Bayesian network of Plant and Holland (2011) provides a complete representation of the statistical dynamics of a simplified model of the surf zone, while other models, e.g., Wilson et al. (2014) and Kurapov and Özkan-Haller (2013), are based on more

restrictive dynamical and statistical assumptions but resolve the two-dimensional spatial fields. Nonetheless, the assumed quasi-linear forms of the statistical estimators in variational and Kalman-filter-based approaches do not appear to be problematic; the estimators employed are apparently acceptable approximations to the non-linear dependence of water level and velocity on bottom depth within the shallow water approximation.

Studies employing field data and idealized identical twin experiments have reported affirmative results concerning the convergence of estimated topography towards the true topography, within the limits imposed by the accuracy of the dynamics. For example, the study of Honnorat et al. (2010) found that the three-dimensional nature of flow separation around a submerged obstacle led to errors in estimated topography when a variational approach based on vertically-integrated shallow water dynamics was used. Likewise, studies of the nearshore zone have mentioned the potential influence of stratification and three dimensional flow features which are not currently represented (Kurapov and Özkan-Haller 2013).

Is it possible to delimit the expected range of scales in which the Saint-Venant or shallow water equations should be useful for the determination of bottom topography? The three main assumptions justifying the use of the shallow water equations are (1) shallow aspect ratio, (2) hydrostatic balance, and (3) vertically homogeneous flow. Taking L as the horizontal scale of motion and H as water depth, and U and W as typical horizontal and vertical velocities, one anticipates $W = UH/L$ assuming the flow to be incompressible. The advective terms in the vertical momentum equation can be compared with the acceleration of gravity, which suggests the condition $L > HF$ for the validity of the hydrostatic approximation, where $F = U/(gH)^{1/2}$ is the Froude number. Finally, in order for the flow to be vertically homogenous (outside of the bottom boundary layer) it is necessary for the turbulent stress at the bottom to be distributed over the full depth of the fluid. This last condition can be analyzed by defining the friction velocity as $u_*^2 = C_d U^2$, and assuming that the time for the influence of the bottom boundary layer to penetrate the water depth is H/u_* (Hinze 1975). The lateral distance associated with this vertical mixing time is $L = HU/u_*$, or $L = H/C_d^{1/2}$. The analysis for Manning-type friction is similar, but includes more complex dependence on H resulting in $L = HK(H/r)^{1/6}$.

A frictional Reynolds number may be defined as the ratio between the nonlinear acceleration and the bottom stress, $Re_f = H/(C_d L)$. One anticipates that friction is dynamically important when $Re_f < 1$, or $L > H/C_d$.

In the case of steady flow, the above analysis suggests the following sequence of scales: $L_1 = FH$, the smallest horizontal scale at which the flow is hydrostatic; $L_2 = H/C_d^{1/2}$, the horizontal scale at which the flow becomes vertically homogenous; and $L_3 = H/C_d$, the horizontal scale at which bottom drag is dynamically dominant. If the flow is unsteady then the time scale, T , of the forcing sets the length scale, $L_{wave} = T\sqrt{gH}$, assuming small F . The scale at which friction becomes significant is then $L_4 = L_3/F$.

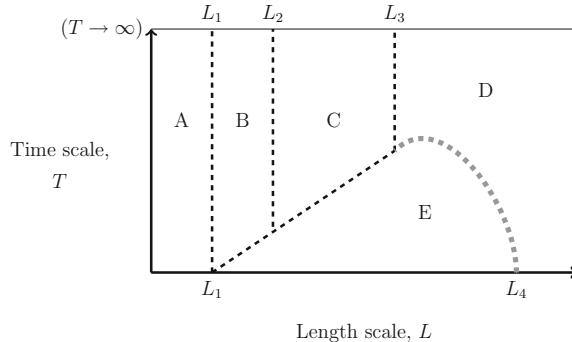


Fig. 1 Scales of length, L , and time, T , are used to categorize dynamical regimes in which the Saint-Venant or shallow water equations are or are not applicable to bottom topography mapping. Indicated length scales are given by $L_1 = HF$, $L_2 = H/C_d^{1/2}$, $L_3 = H/C_d$, and $L_4 = L_3/F$, where H is water depth, F is Froude number, and C_d is Chézy friction coefficient. In region A ($L < L_1$) dynamics are non-hydrostatic and the conventional shallow water equations do not apply. In region B ($L_1 < L < L_2$ and $T > L/\sqrt{gH}$) the flow is quasi-steady, but does not satisfy the assumption of vertical homogeneity as required for shallow water dynamics. In region C ($L_2 < L < L_3$ and $T > L/\sqrt{gH}$) the flow is quasi steady, frictional effects are small, and shallow water dynamics are applicable. In region D ($L > L_3$) the flow is strongly influenced by bottom friction, and the kinematic wave approximation applies. Region E ($L_1 < L < L_4$ and $T < L/\sqrt{gH}$) is governed by weakly-damped shallow water wave dynamics. The boundary between regions D and E is not clear from scaling arguments alone, but it defines the region D in which frictional effects are dominant

These scales are sketched in Fig. 1 to indicate the various domains in which the shallow water or Saint-Venant equations are or not valid. Within region A, defined by $L < L_1$, the dynamics are non-hydrostatic and the conventional shallow water equations do not apply. In region B, $L_1 < L < L_2$ and $T > L/\sqrt{gH}$, the flow does not satisfy the assumption of vertical homogeneity as required for shallow water dynamics. In region C, $L_2 < L < L_3$ and $T > L/\sqrt{gH}$, the flow is quasi-steady, frictional effects are small, and shallow water dynamics are applicable. In region D, $L > L_3$, the flow is strongly influenced by bottom friction, and the kinematic wave approximation applies. In region E, $L_1 < L < L_4$ and $T < L/\sqrt{gH}$, weakly-damped shallow water wave dynamics are applicable. The boundary between regions D and E is not clear from scaling arguments alone, but it defines one boundary of region D in which frictional effects are dominant. The regimes are broadly distinguished by the applicability of shallow water wave dynamics in region E, with time scale, $T < L/\sqrt{gH}$, versus the quasi-steady regime C, $T > L/\sqrt{gH}$, and the frictional regime D.

The dynamical regimes can be used to understand some of the results obtained in the literature in relation to the type of measurements assimilated. For example, wave-celerity measurements are used primarily in the nearshore applications. The measurement principle involves making measurements of wave period and wavelength, which are converted to depth measurements via the dispersion relation (7). The measurement technique thus resolves a length scale $L_d = 2\pi k^{-1}$ equal to the

wavelength of the surface waves, and the assumed dynamics play a secondary role. Consequently, the nearshore studies utilizing wave celerity measurements are able to identify topographic features at scales between L_d and L_2 , in regime B, where shallow water dynamics are not formally valid.

In regime D, when conditions for the validity of the kinematic wave equation are satisfied, there results in a balance, for Chézy-type friction,

$$gHS = -C_d|U|U, \quad (8)$$

so there is a linear relationship between squared water speed, U^2 , and the water depth, H , with proportionality constant defined by the ratio of the river slope and friction coefficient, gS/C_d . Obviously, if C_d and S were known, then it would be possible to estimate H from measurements of U in this regime, even in steady conditions; and if the river slope were known, then the drag coefficient could also be determined. In this one-dimensional framework, the kinematic constraint of volume conservation can be used to eliminate slope in favor of H , U , and U_x ,

$$S = -HU_x/U, \quad (9)$$

assuming constant river width and flow variations in the along-stream direction (U_x is non-zero). Under steady or slowly varying conditions the kinematic and dynamic relations can be combined to

$$gH^2U_x = C_dU^3, \quad (10)$$

and it is evident that H^2/C_d could be estimated from snapshots of U distributed along a river (so that U_x would also be estimated). If only the time-variable part of the water level, η , were measured, with H_0 being the steady part, $H = H_0 + \eta$, then it is useful to rewrite (10) as

$$H_0^2 + 2H_0\eta + \eta^2 = C_dU^3/U_x. \quad (11)$$

There are thus now 3 unknown fields, H_0 , C_d , and U to be determined from η measurements. Although H_0 and C_d are assumed steady, it is evident that determination of these fields and U would require additional information in the form of hypothesized priors or additional measurements.

The results of Honnorat et al. (2010) illustrate the changing nature of the topographic estimation problem across regimes B, C, and D. They found that velocity measurements were not adequate to identify topographic features at scales smaller than about $10H$, which is approximately equal to L_2 , within regime B (assuming $C_d = 5 \times 10^{-3}$). In this regime they observed a three-dimensional flow feature, a recirculation cell, associated with the topographic feature (a weir). In contrast, the flow measurements, together with upstream and downstream elevation boundary data, and the variational smoothness penalties, were adequate to allow estimation of both drag and depth fields at larger scales within regimes C and D.

Figure 1 also provides a useful guide for developing models of spatial statistics of topography and bottom drag in spatially resolved models. For example, in steady systems at scales, $L < L_3$, the flow is nearly independent of friction. This would suggest that prior covariance models for the friction coefficient, C_d , ought to impose a correlation length comparable to or larger than L_3 . If this is not done, then features of the C_d field at scales smaller than L_3 are likely to be spurious and reflect ill-conditioning of the inverse. In systems governed by wave dynamics one would expect L_4 would provide the minimum useful scale at which C_d can be identified. Similarly, spatial covariance models for the topography ought to use a correlation scale of L_2 or larger for steady flow, and L_1 or larger for the systems governed by wave dynamics. Since $L_1 < L_2$ for small Froude number flows, it is hypothesized that the finest spatial resolution topographic estimates will be obtained using time resolved flows governed by wave dynamics.

4 Summary

Identification of underwater topography from remotely sensed or indirect measurements of environmental fluid flows is an important problem with applications in hydrology, coastal and littoral zone oceanography, and ocean and river modeling. Partly because of the breadth of applications, the literature concerned with inverse methods for bottom topography mapping is found within a range of journals, making it a challenge to identify commonalities and relationships among many approaches. This article has surveyed developments over approximately the last decade in order to identify common themes and enable potential synergies amongst researchers working in different domains and application areas.

The methods used to estimate water depth or bottom topography have ranged from non-parametric models utilizing Bayes' rule to propagate and infer the probability distribution of water depth at a few sites (Plant and Holland 2011), to spatially resolved mapping using estimators that are optimal only in the case of Gaussian statistics (e.g., Honnorat et al. 2009; Zaron et al. 2011; Wilson and Özkan-Haller 2012). Provided the data are sufficiently dense and accurate researchers have found it feasible to estimate and map water depth or underwater topography; although, the specific criteria for data density varies depending on the application. Within hydrology there has been an emphasis on determination of along-channel profiles, sometimes at spatial resolutions of 10's to 100's of kilometers, where it has been computationally feasible to employ ensemble methods to estimate channel depth and frictional parameters for subsequent use in flow forecasting. For studies of nearshore morphology and some riverine applications there has been an emphasis on spatially-resolved mapping, for which both ensemble and variational methods have proved useful.

Fundamental questions of observability of the bottom topography from particular measurement types, and related questions of stability and conditioning, have been studied in the context of particular applications (e.g., Wilson and Özkan-Haller 2012; Landon et al. 2014). Dynamical models used consist of the shallow water or Saint-

Venant equations, with either Chézy or Manning frictional representations, and, in nearshore studies, a phase-averaged wave model coupled through wave radiation stresses and/or dissipation sub-model to the shallow water dynamics. A diagram indicating regimes characterized by space and time scales expressed in terms of Froude number, frictional Reynolds number, aspect ratio, and phase speed of non-dispersive shallow water waves has been proposed in order to reason about applications based on shallow water dynamics, the Saint-Venant equations, or their simplifications.

The treatment of three-dimensional effects, in the form of vertical shear or baroclinic pressure gradients, has not been explicitly incorporated in the bottom mapping efforts. Implicitly the vertical shear has been taken into account by modifying measurement operators so that measurements of near-surface current by remote sensing are modified to correspond to the vertical mean velocity. Baroclinic effects, though recognized as potentially significant, have not been incorporated into the inverse models, but they have been used in estuarine models in which topography is optimized as part of the model calibration process (Cea and French 2012). Bottom topography mapping in the presence of strongly three-dimensional dynamics is likely to be of increasing attention as established methods are refined in the nearshore and estuarine environments.

Acknowledgements Support for this work was provided by NASA, Ocean Surface Topography Science Team Grant #NNX13AH06G, which is gratefully acknowledged.

References

- Biancamaria S, Coauthors, (2010) Preliminary characterization of SWOT hydrology error budget and global capabilities. *Sel Topics Appl Earth Observ Remote Sens IEEE J* 3(1):6–19
- Bjerklie DM, Dingman SL, Bolster CH (2005) Comparison of constitutive flow resistance equations based on the Manning and Chezy equations applied to natural rivers. *Water Resour Res* 41(W11):502
- Blumberg AF, Georgas N (2008) Quantifying uncertainty in estuarine and coastal ocean circulation modeling. *J Hydraul Eng* 134(4):403–415
- Cea L, French JR (2012) Bathymetric error estimation for the calibration and validation of estuarine hydrodynamic models. *East Coast Shelf Sci* 100:124–132
- Chow VT (1959) Open-Channel hydraulics. McGraw-Hill, New York 680 p
- Dunne S, Entekhabi D (2005) An ensemble-based reanalysis approach to land data assimilation. *Water Resour Res* 41:W02013
- Evensen G (2006) Data assimilation: the ensemble Kalman filter. Springer, Berlin 280 p
- Falcao AP, Mazzolari A, Goncalves AB, Araujo MAV, Trigo-Teixeira A (2013) Influence of elevation modelling on hydrodynamic simulations of a tidally-dominated estuary. *J Hydrol* 497:152–164
- Garambois P-A, Monnier J (2015) Inference of effective river properties from remotely sensed observations of water surface. *Adv Water Resour* 79:103–120
- Gill AE (1982) Atmosphere-ocean dynamics. Academic Press, 662 pp
- Gleason CJ, Smith LC (2014) Toward global mapping of river discharge using satellite images and at-many-stations hydraulic geometry. *Proc Nat Acad Sci* 111(13):4788–4791
- Hamill TM, Whitaker JS, Snyder C (2001) Distance-dependent filtering of background error covariance estimates in an ensemble Kalman filter. *Mon Weather Rev* 129(11):2776–2790

- Hinze JO (1975) Turbulence, 2nd edn. McGraw-Hill Publishing, New York 790 pp
- Holland K, Puleo J, Kooney T (2001) Quantification of swash flows using video-based particle image velocimetry. *Coast Eng* 44:65–77
- Holman R, Haller MC (2013) Remote sensing of the nearshore. *Ann Rev Mar Sci* 5:95–113
- Holman R, Plant N, Holland T (2013) cBathy: a robust algorithm for estimating nearshore bathymetry. *J Geophys Res* 118(5):2595–2609
- Holman RA, Holland KT, Lalejini DM, Spansel SD (2011) Surf zone characterization from unmanned aerial vehicle imagery. *Ocean Dyn* 61(11):1927–1935
- Holthuijsen L (1983) Stereophotography of ocean waves. *Appl Ocean Res* 5(4):204–209
- Honorat M, Monnier J, LeDimet F-X (2009) Lagrangian data assimilation for river hydraulics simulations. *Comput Vis Sci* 12(5):235–246
- Honorat M, Monnier J, Riviere N, Huot E, LeDimet FX (2010) Identification of equivalent topography in an open channel flow using Lagrangian data assimilation. *Comput Vis Sci* 13(3):111–119
- Kurapov AL, Özkan-Haller HT (2013) Bathymetry correction using an adjoint component of a coupled nearshore wave-circulation model: Tests with synthetic velocity data. *J Geophys Res* 118(9):4673–4688
- Landon KC, Wilson GW, Özkan-Haller HT, MacMahan JH (2014) Bathymetry estimation using drifter-based velocity measurements on the Kootenai River, Idaho. *J Atm Ocean Tech* 31:503–514
- Luettich RA, Westerink JJ, Scheffner NW (1991) ADCIRC: an advanced three-dimensional circulation model for shelves, coasts and estuaries, report 1: Theory and methodology of ADCIRC-2DDI and ADCIRC-3DL. Tech. Rep. Dredging Research Program, DRP-92-6, Department of the Army
- Pavelsky TM, Durand MT, Andreadis KM, Beighley RE, Paiva RC, Allen GH, Miller ZF (2014) Assessing the potential global extent of SWOT river discharge observations. *J Hydrol* 519, Part B:1516–1525
- Plant NG, Edwards KL, Kaihatu JM, Veeramony J, Hsu L, Holland KT (2009) The effect of bathymetric filtering on nearshore process model results. *Coast Eng* 56(4):484–493
- Plant NG, Holland KG (2011) Prediction and assimilation of surf-zone processes using a Bayesian network: Part I: Forward models. *Coast Eng* 58(1):119–130
- Plant NG, Holland KT, Haller M (2008) Ocean wavenumber estimation from wave-resolving time series imagery. *IEEE Trans Geosci Remote Sens* 46:2644–2658
- Rodriguez E (2015) SWOT science requirements document. Tech. Rep., Jet Propulsion Laboratory
- Roux H, Dartus D (2005) Parameter identification using optimization techniques in open-channel inverse problems. *J Hydraul Res* 43(3):311–320
- Shchepetkin AF, McWilliams JC (2005) The regional ocean modeling system (ROMS): a split-explicit, free-surface, topography-following-coordinate oceanic model. *Ocean Model* 9:347–404
- Simeonov JA, Holland KT, Calantoni J, Anderson SP (2013) Calibrating discharge, bed friction, and datum bias in hydraulic models using water level and surface current observations. *Water Resour Res* 49(12):8026–8038
- Smith WH, Sandwell DT (1994) Bathymetric prediction from dense satellite altimetry and sparse shipboard bathymetry. *J Geophys Res* 99(21):803–821, 824
- Thornton EB, Guza RT (1982) Energy saturation and phase speeds measured on a natural beach. *J Geophys Res* 87:9499–9508
- van Dongeren A, Plant N, Cohen A, Roelvink D, Haller MC, Catalan P (2008) Beach Wizard: Nearshore bathymetry estimation through assimilation of model computations and remote observations. *Coast Eng* 55(12):1016–1027
- Wilson GW, Özkan-Haller HT (2012) Ensemble-based data assimilation for estimation of river depths. *J Atmos Ocean Tech* 29(10):1558–1568
- Wilson GW, Özkan-Haller HT, Holman RA (2010) Data assimilation and bathymetric inversion in a two-dimensional horizontal surf zone model. *J Geophys Res* 115:C12057

- Wilson GW, Özkan-Haller HT, Holman RA, Haller MC, Honegger DA, Chickadel CC (2014) Surf zone bathymetry and circulation predictions via data assimilation of remote sensing observations. *J Geophys Res* 119(3):1993–2016
- Yoon Y, Durand M, Merry CJ, Clark EA, Andreadis KA, Alsdorf DE (2012) Estimating river bathymetry from data assimilation of synthetic SWOT measurements. *J Hydrol* 464–465:363–375
- Zaron ED, Pradal M, Miller PD, Blumberg AF, Georgas N, Li W, Cornuelle JM (2011) Bottom topography mapping via nonlinear data assimilation. *J Atmos Ocean Tech* 28:1606–1623
- Zou X, Navon IM, LeDimet FX (1992) An optimal nudging data assimilation scheme using parameter estimation. *Quart J Royal Met Soc* 118:1163–1186

The Impact of Doppler Wind Lidar Measurements on High-Impact Weather Forecasting: Regional OSSE and Data Assimilation Studies

Zhaoxia Pu, Lei Zhang, Shixuan Zhang, Bruce Gentry,
David Emmitt, Belay Demoz and Robert Atlas

Abstract Wind profiles are essential for operational weather forecasting on all scales and at all latitudes. However, tropospheric winds are the number one unmet measurement objective for improving weather forecasts. In recent years, ground-based and airborne Doppler wind lidar (DWL) wind profiles have been used in field programs and various applications to obtain the necessary wind measurements. These measurements offer the opportunity to examine the impact of wind profiles on numerical weather prediction (NWP). In addition, satellite-based DWL missions are also being planned. Observing System Simulation Experiments (OSSEs) have been conducted to evaluate the impact of future space-based satellite global wind measurements on NWP. While many previous studies have emphasized global NWP systems, in this chapter we provide an overview and summary of recent studies with both data assimilation and OSSEs to demonstrate the value of DWL wind measurements in improving severe weather system forecasts in regional NWP, especially for systems with large societal impacts due to the damage they may cause (e.g., high-impact weather systems). Specifically, we give an overview

Contributed to Springer Book by Seon K. Park and Liang Xu (Eds.).

“Data Assimilation for Atmospheric, Oceanic and Hydrologic Applications, Volume III”
February 2016.

Z. Pu (✉) · L. Zhang · S. Zhang

Department of Atmospheric Sciences, University of Utah, 135 S 1460 E,
Rm. 819, Salt Lake City UT 84112, USA
e-mail: Zhaoxia.Pu@utah.edu

B. Gentry

NASA Goddard Space Flight Center, Greenbelt, Maryland, USA

D. Emmitt

Simpson Weather Associates, Charlottesville, VA, USA

B. Demoz

University of Maryland, Baltimore County, Baltimore, MD, USA

R. Atlas

NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, FL, USA

of previous studies that have examined the impacts of ground-based and airborne DWL on the numerical predictions of mesoscale convective systems and hurricanes. The regional OSSE concept is introduced. Recent results with regional OSSEs using the mesoscale community Weather Research and Forecasting (WRF) model and the NCEP Hurricane WRF (HWRF) model are presented. The potential configuration (e.g., resolution vs. accuracy) for future satellite-based DWL is evaluated. It is found that fairly good forecast impacts can be obtained from high-resolution observations with larger errors compared with accurate observations at a coarser resolution. Finally, the relative impact of ocean-surface wind measurements and 3-dimensional profiles is compared. The advantages of 3-D wind measurements are evident.

1 Introduction

The proper specification and analysis of tropospheric winds is an important prerequisite for accurate weather forecasting and climate study. The World Meteorological Organization (WMO 1996) determined that global wind profiles are “essential for operational weather forecasting on all scales and at all latitudes.” This is because the wind field plays a unique dynamical role in forcing the mass field to adjust to it at all scales in the tropics, and at small scales in the extratropics (Baker et al. 1995; Baker et al. 2014). According to the National Research Council (NRC 2007), “more accurate, more reliable, and longer-term weather forecasts, driven by fundamentally improved tropospheric wind observations from space, would have a direct and measurable societal and economic impact. Tropospheric winds are the number one unmet measurement objective for improving weather forecasts.”

During the last decades, progress has been made in developing and planning global wind measurements. Many scientist’s effort have been devoted to exploring the new space Doppler lidar wind-measuring missions (e.g., Baker et al. 2014). It is expected that future Doppler wind lidars (DWL; Stoffelen et al. 2005; Riishojgaard et al. 2012; Baker et al. 2014) will provide much denser wind profile observations than the currently available rawinsonde networks. Recent observing system simulation experiments (OSSEs) have proved that the assimilation of Doppler wind lidar data will result in improved numerical weather forecasts (e.g., Atlas et al. 1985, 2003, 2015a, b; Atlas 1997; Atlas and Emmitt 2008; Masutani et al. 2006; Masutani et al. 2010; Kalnay and Liu 2007; Pu et al. 2009; Zhang and Pu 2010; Riishojgaard et al. 2012). Meanwhile, the European Space Agency (ESA)’s Earth Explorer Atmospheric Dynamic Mission (ADM-Aeolus) is under preparation for launch, with goals to provide global observations of single wind component profiles from space to improve the quality of weather forecasts, and to advance understanding of atmospheric dynamics and climate processes.

In addition, ground-based wind lidar (Gentry et al. 2000; Demoz et al. 2006) has been used in recent field programs and many other applications, such as wind energy-related sciences (e.g., Pichugina et al. 2012). Airborne Doppler wind lidar

(Gentry et al. 2010; Weissmann et al. 2005 and Weissmann et al. 2012; Emmitt et al. 2011) has also been employed during field programs. For instance, an airborne Doppler wind lidar was onboard the Naval Research Laboratory's P3 during the US Office of Naval Research (ONR)'s Tropical Cyclone Structure 2008 (TCS-08) field program. Wind measurements collected from ground-based and airborne wind lidar have proven to be useful for improving precipitation and tropical cyclone forecasting (e.g., Pu et al. 2010; Weissmann et al. 2012; Zhang and Pu 2011).

In parallel to the above mentioned efforts, the United States National Aeronautics and Space Administration (NASA) has classified tropospheric wind profiling as high-priority science and has invested in developing wind profiling instruments through its Instrument Incubator Program (IIP). In addition to space-based wind lidar measurements, a high-altitude airborne system flown on an Unmanned Aerial Vehicle (UAV) or other advanced platform is of great interest for studying mesoscale atmospheric systems. For instance, a DWL instrument called the Tropospheric Wind Lidar Technology Experiment (TWiLiTE) has been used for several recent missions.

Studies have also been conducted to demonstrate the impacts of Doppler wind lidar measurements on weather forecasting (e.g., Weissmann and Cardinali 2007; Pu et al. 2010) and the impact of potential configurations of future lidar measurements on tropical cyclone forecasts (e.g., Atlas and Emmitt 2008; Atlas et al. 2015a, b; Pu et al. 2009; Zhang and Pu 2010; Riishojgaard et al. 2012).

Considering that the most challenging problem for modern NWP is to predict weather systems that have strong social and economic impact (for instance, hurricanes, winter storms, mesoscale severe convective systems, and other severe systems that can cause life and property damage), there is good reason to anticipate that future observing systems will be helpful for improving the forecast of these high-impact weather events. Therefore, while many previous studies have emphasized global NWP systems, in this chapter we provide an overview and summary of recent studies with both data assimilation and OSSEs to demonstrate the value of DWL wind measurements in improving regional NWP forecasts of severe weather systems that have large societal impacts (e.g., high-impact weather systems). Specifically, the results of previous studies on the impact of ground-based and airborne DWL on numerical prediction of mesoscale convective systems and hurricanes are overviewed. Recent regional OSSE results with an advanced research version of the Weather Research and Forecasting (WRF ARW) model (Skamarock 2008) and the NCEP Hurricane WRF (HWRF) model (Tallapragada et al. 2014, Atlas et al. 2015c) are presented. The potential requirements for future satellite-based DWL are also discussed.

The chapter is organized as follows: Sect. 2 briefly overviews the results from previous studies that demonstrate the impact of ground-based and airborne wind profiles on high-impact weather forecasts. Section 3 introduces the regional OSSE concept and also summarizes case studies with the WRF ARW model. Section 4

presents OSSE results with the NCEP HWRF and a Gridpoint Statistical Interpolation (GSI) data assimilation system. A summary and concluding remarks are provided in Sect. 5.

2 Overview of the Impact of Ground-Based and Airborne DWL Wind Profiles on High-Impact Weather Forecasts

2.1 *Ground-Based DWL Wind Profiles*

Ground-based DWL devices are commonly portable (e.g., mobile) platforms that measure wind profiles. They have been used in many field programs and have also become popular in recent years with the growth of the wind-energy industry.

Among these ground-based DWL devices, the Goddard Lidar Observatory for Winds (GLOW) is a mobile direct detection Doppler lidar system (Gentry and Chen 2003; Fig. 1a). GLOW uses an optical interferometric technique to measure the Doppler shift of the laser signal backscattered by air molecules. The lidar operates

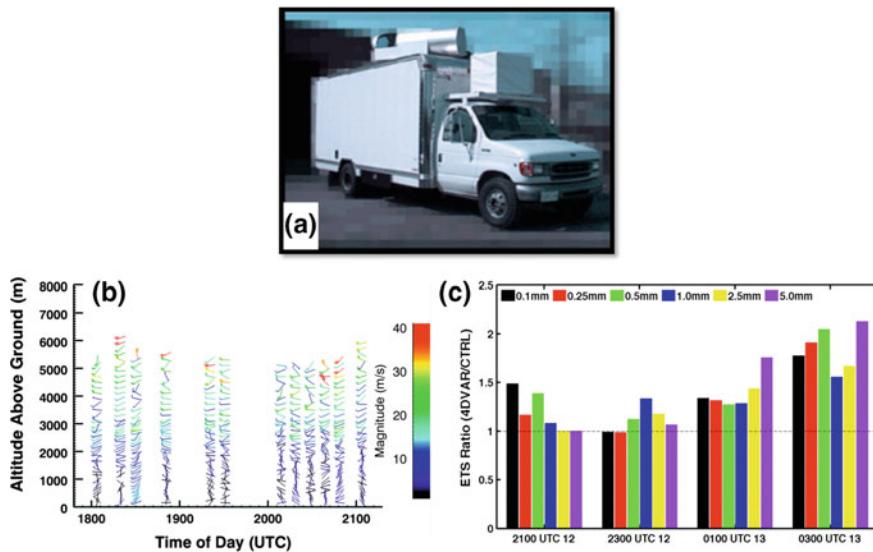


Fig. 1 **a** The photo of GLOW ground-based mobile platform. **b** Time series of GLOW wind profiles at Homestead profiling site (36.558°N , 100.606°W) from 1800 to 2100 UTC Jun 2002. Colors represent the different magnitudes of wind speeds. **c** The ratio of equitable threat scores (ETS) for 1-h accumulated precipitation between control (without assimilation of Lidar wind profiles) and 4DVAR (with assimilation of Lidar wind profiles) experiments with the threshold of 0.1, 0.25, 0.5, 1.0, 2.5, and 5.0 mm at 2100 UTC 12 Jun 2002, 2300 UTC 12 Jun 2002, 0100 UTC 13 Jun 2002, and 0300 UTC 13 Jun 2002. If the ETS ratio greater than 1, 4DVAR outperforms control. (**b** and **c** are from Zhang and Pu 2011)

at a wavelength of 355 nm and is designed to profile winds in clear air from the surface up to the lower stratosphere. In May and June of 2002, GLOW was deployed during IHOP_2002 (International H₂O Project) to collect a continuous time series of wind speed and direction from the surface up to the tropopause and to characterize the flow and dynamics in and above the boundary layer. GLOW was located at the Homestead profiling site (36.5588 N, 100.6068 W) in Oklahoma, USA. In addition, several other lidars, radars, and passive instruments were operated from the Homestead site and provided a unique cluster of observations in the IHOP_2002 field experiments. During IHOP_2002, over 240 h of wind profile measurements from 34 days of operation were collected with GLOW. Data were subjected to quality control and preprocessing, and yielded two types of data products (wind speed, wind direction, and u and v wind components): one in 30-min time intervals and the other in 10-min time intervals, both with 100-m vertical resolution for altitudes below 3 km and 200-m vertical resolution for altitudes above 3 km. Vertical wind profiles are available from the surface up to about 7 km.

To take advantage of the high temporal resolution, the winds with 10-min intervals were used in a data assimilation experiment to examine the impact of the data on numerical simulations of the initiation and evolution of a mesoscale convective system from the Kansas and Oklahoma border to the Texas Panhandle, observed 12–13 June 2002. Specifically, wind profile observations obtained from GLOW were assimilated into the WRF ARW model (Skamarock et al. 2008) using its four-dimensional variational data assimilation (4DVAR) system (Huang et al. 2009). Detailed studies and results are documented in Zhang and Pu (2011).

Figure 1b shows the time series of GLOW wind profiles from 1800 to 2100 UTC 12 June 2002 assimilated by 4DVAR in this study. Numerical experiments indicate that the assimilation of these GLOW wind profiles with high temporal and vertical resolution has a significant influence on the numerical simulation of convective initiation and evolution. Besides the wind fields, the simulation of the structure of the moisture fields associated with the convective system is also improved. Data assimilation also results in more accurate prediction of the location and timing of convective initiations; as a consequence, the skill of quantitative precipitation forecasting (QPF) is greatly enhanced (Fig. 1c). See details in Zhang and Pu (2011).

2.2 *Airborne DWL Wind Profiles*

During the THORPEX Pacific Asian Regional Campaign (TPARC) and ONR TCS-08 field experiments in 2008, an airborne DWL was onboard the U.S. Naval Research Laboratory's P3 research flight. It was the first time the DWL was used for a tropical cyclone mission. With the ability to sample wind profiles at 50 m

resolution vertically and 2 km horizontally, the airborne DWL provided high-resolution wind profiles for tropical cyclone studies. Typhoon Nuri (2008) over the western Pacific Ocean was the first tropical system ever sampled by the airborne DWL. The mission occurred around Nuri when it was still a tropical disturbance, from 2330 UTC 16 August to 0200 UTC 17 August 2008. Nuri was designated a tropical depression at 1200 UTC 17 August 2008 by the Joint Typhoon Warning Center (JTWC). The Japan Meteorological Agency named Nuri as a tropical storm the next day (18 August 2008), and it reached typhoon status late on 18 August 2008. A data assimilation experiment was conducted to examine the impact of the assimilation of DWL observations on the numerical simulation of the formation and development of Typhoon Nuri (see details in Pu et al. 2010).

Figure 2 shows the sample measurements at 1500 m height along the flight track for DWL wind profiles during a three-hour interval (2330 UTC 16 August to 0200 UTC 17 August 2008). Most of the profiles extended from near the surface to a height of 2000 m. In order to assess the quality of the DWL data, the DWL wind profiles were compared with the dropsonde data collected on the same flight. Results showed that the DWL observations agreed well with the dropsonde winds. The correlation between the two observations is nearly 98 % (Pu et al. 2010).

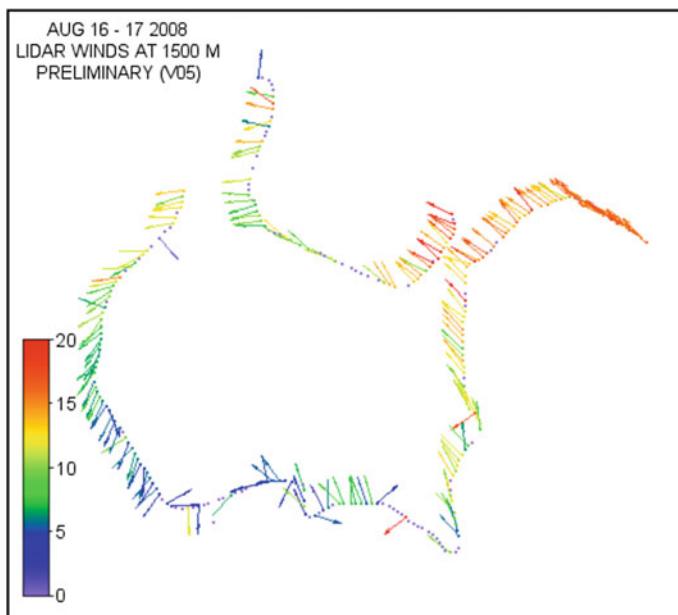
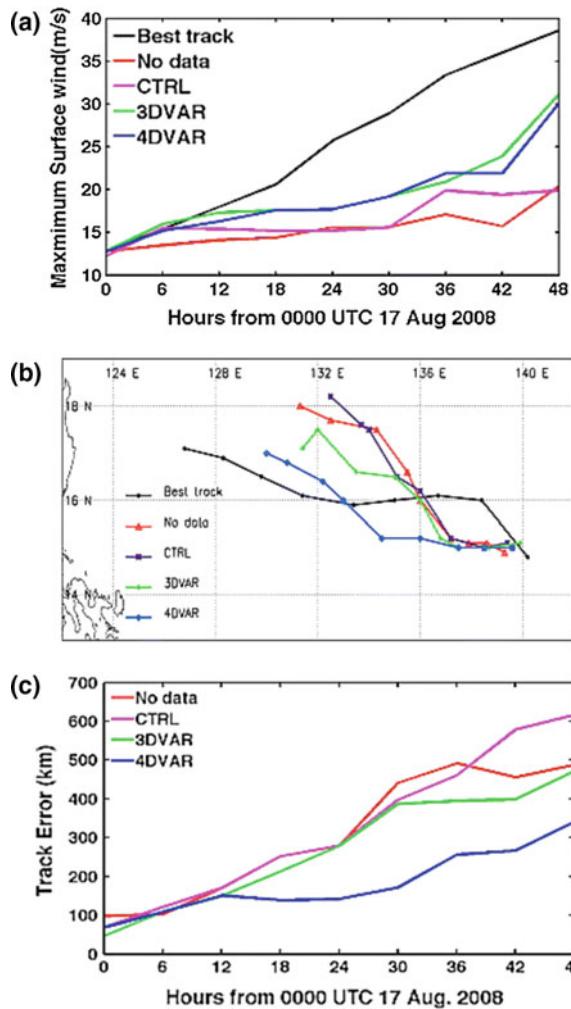


Fig. 2 TPARC/TCS08 Doppler wind lidar (DWL) observational locations along the NRL P-3 flight track with sample lidar wind measurements at 1500 m height level during 2330 UTC 16 August to 0200 UTC 17 August 2008 around Nuri. Colors denote wind speeds

With the WRF ARW model and its three-dimensional variational (3DVAR, Barker et al. 2004) and four-dimensional variational (4DVAR, Huang et al. 2009) data assimilation systems, numerical experiments demonstrate that the DWL data have a positive impact on numerical simulations of Typhoon Nuri in terms of its formation, track, and intensity. Compared with the three-dimensional variational method, the four-dimensional variational data assimilation system is deemed more promising for assimilating DWL data (Fig. 3) as it leads to better analysis and forecast. Detailed results can be found in Pu et al. (2010).

Fig. 3 **a** The maximum surface wind, **b** Nuri's track and **c** track errors from 0000UTC 17 August 2008 to 0000UTC 19 August 2008. The forecasts with (green curves for 3DVAR and blue curves for 4DVAR) and without (“no data” in red curve and “CTRL” in purple curve) assimilation of DWL winds are compared with the JTWC best track data (black curves in Figs. 5a, b). DWL data are assimilated for the period of 0000 UTC–0200 UTC 17 August 2008 in both the 3DVAR and 4DVAR experiments. Conventional observations and dropsondes were assimilated in “CTRL”. (From Pu et al. 2010)



3 The Impact of Satellite-Based Wind Profiles on Hurricane Forecasts: Results from OSSEs with the WRF ARW Model

3.1 Brief Overview of the Regional OSSE Concept and Early Studies

Besides the applications of ground-based and airborne DWL, there have been proposals to use DWL to measure three-dimensional wind profiles globally with polar-orbiting satellites. For instance, the European Space Agency has taken a step forward in planning an ADM-Aeolus space mission. In the United States, a space-based wind lidar science working group has been actively making progress for many years. Many such efforts have been documented in a recent paper by Baker et al. (2014), published in the *Bulletin of the American Meteorological Society*.

Assessing the impact of potential satellite-based DWL wind profiles on numerical prediction is an important step for planned and potential future missions. Efforts have been made with OSSEs to not only examine the impact of planned and potential DWL data on weather forecasting, but also to determine the minimum requirements of wind measurements regarding resolution, distribution, and expected errors in order to ensure improved forecasting. Several previous studies (Atlas et al. 1985; Atlas 1997; Stoffelen et al. 2006; Atlas and Emmitt 2008; Riishojgaard et al. 2012) emphasized the impact of DWL measurements on NWP with global models. Positive impacts were generally found in all studies. Nevertheless, only a very few studies (e.g., Zhang and Pu 2010; Nolan et al. 2013; Atlas et al. 2015a, b) have focused on OSSEs with regional models.

Compared with global OSSEs, regional OSSEs are justified by the need of high-resolution models to realistically resolve mesoscale severe weather systems, because many global models and available nature runs (such as those produced by European Centre for Medium-Range Weather Forecasts (ECMWF)) cannot resolve the detailed structure of mesoscale severe weather systems such as hurricanes.

For instance, in order to support community needs for OSSEs in DWL and other new instrumentation, the ECMWF produced global nature runs using a spectral prediction model in July 2006. There were two nature runs with different resolutions: one was at T511 spectral truncation (about 40 km horizontal resolution; T511 nature run or T511 NR hereafter) with 91 vertical levels and 3-hour frequency output from 1200 UTC 1 May 2005 to 0000 UTC 1 June 2006. The other was a higher-resolution simulation at T799 spectral truncation (about 25 km horizontal resolution; T799 NR hereafter) with 91 vertical levels and hourly output from 27 September 2005 to 1 November 2005. Reale et al. (2007) examined these nature runs and indicated that the datasets produced reasonable Atlantic hurricanes in terms of hurricane track.

Pu et al. (2009) also examined the same nature run datasets and commented that the ECMWF nature runs were sufficiently accurate in describing tropical cyclone tracks and intensity at an intermediate model resolution. However, they were not adequate in representing tropical cyclone inner-core structures. For example, Fig. 4 compares 3-h accumulated hurricane precipitation structures (shaded contour), wind vectors and sea level pressures at 0000 UTC 2 October 2005 from the ECMWF T511 NR with these downscaled from the WRF model numerical simulations at 9 km and 3 km horizontal resolution grid spacings. It shows that only high-resolution simulations from the WRF model at 9 km and 3 km grid spacings can better resolve the tropical cyclone inner-core structure. Thus, it is necessary to generate regional nature runs for regional verification purposes if mesoscale structures such as hurricane inner-core structures are important.

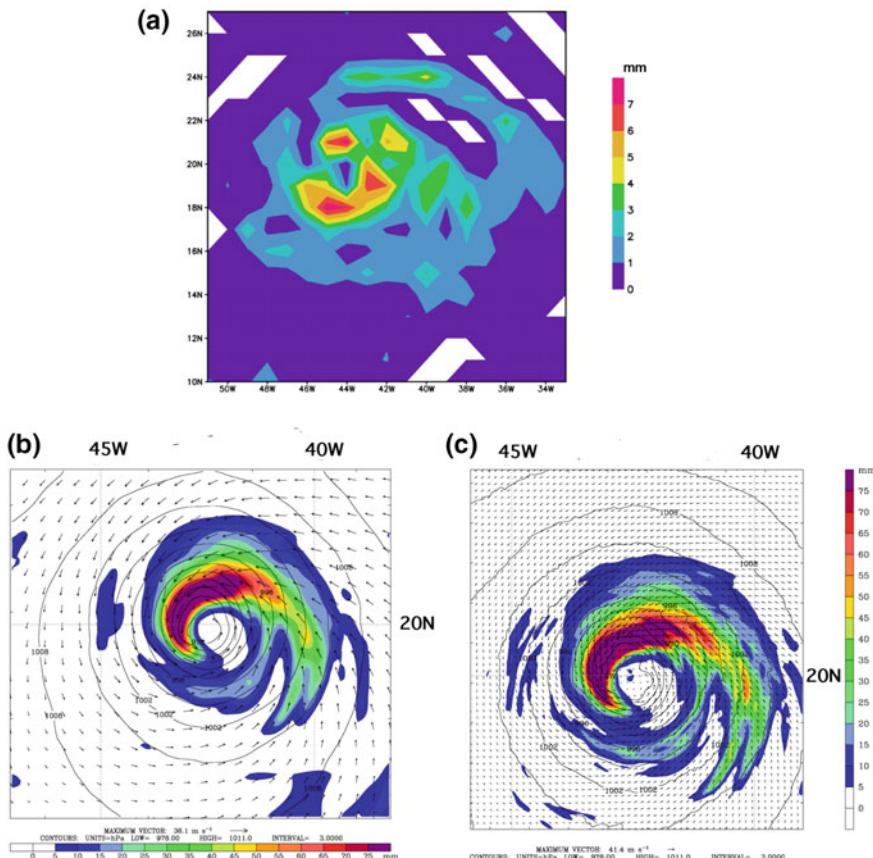


Fig. 4 Comparison of hurricane 3-h precipitation (shaded contour) structures, at 0000 UTC 2 October 2005 from **a** ECMWF T511 nature run and **b** WRF model numerical simulations at 9 km resolution and **c** 3 km resolution. In **b** and **c** wind vectors and sea level pressure are also shown

Therefore, the configuration and procedure of a regional OSSE can be summarized in a flow chart presented in Fig. 5. The study by Zhang and Pu (2010) presented the first regional OSSE for satellite-based DWL observations following the procedure in Fig. 5. The details are as follows:

- (1) Generate a nature run that represents reasonable hurricane structure and intensity

The WRF ARW model was nested inside the ECMWF nature run to generate a set of regional nature runs. The model was initialized using the T799 NR and then integrated forward for 78 h starting at 0000UTC 30 September 2005. The horizontal grid spacings were 27 km, 9 km, and 3 km for the three-level nested domains, respectively. The model physics parameterizations included: the Lin microphysics scheme, the Mellor-Yamada-Janjic planetary boundary layer model (MYJ), the Betts-Miller-Janjic cumulus parameterization scheme, rapid radiative transfer model (RRTM) longwave, and the Dudhia shortwave radiation model (See details in Skamarock 2008).

- (2) Obtain the simulated “observations”

The DWL was assumed to be aboard a given polar-orbiting satellite. The wind measurements were available only twice daily over the same region. Considering the influence of clouds, two configurations of observation sampling (with and without cloud contamination) were simulated at 0600UTC and 1800UTC 01 October 2005, respectively. Wind observations were available from near the surface up to 18km. When the effect of clouds was taken into account, wind profiles were

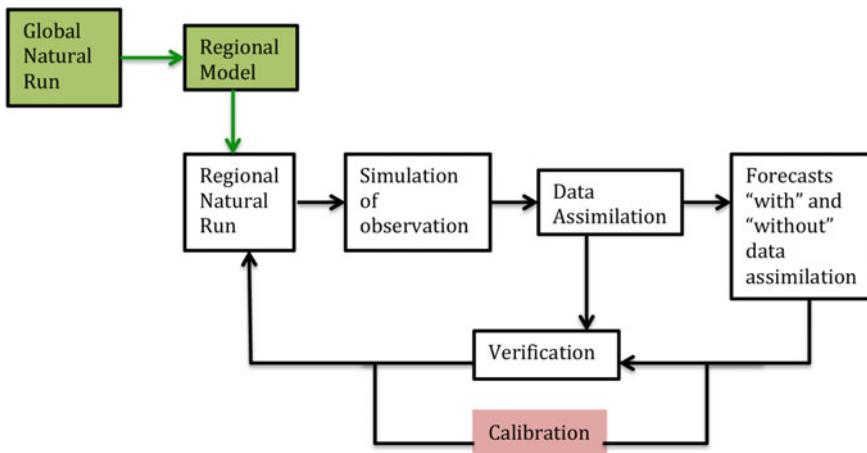


Fig. 5 A diagram of regional OSSE procedure, following a procedure of global OSSE by Atlas (1985) with modification

not available in areas with cloud contamination. The observations used for data assimilation experiments were generated by interpolating the “truth” field (regional nature run) both horizontally and vertically from the “nature run” model grids onto the simulated observational locations and by superimposing random noise. Typical values for the standard deviation of DWL wind errors were 2 ms^{-1} below 2 km and 3 ms^{-1} above 2 km. No bias was assumed for the simulated DWL wind error.

(3) Use a control run to generate a reference field

The control run was a 48-hour free forecast. The initial conditions were obtained by interpolating the ECMWF coarser-resolution T511 NR into the WRF model domains. The boundary conditions were provided by the T799 NR. The model domains were set within the domains of the regional nature run but were smaller in size. In order to take into account the common model errors in OSSEs and also to avoid ideal “twin” experiments, model physical options that were deployed in the control run were different from those used in the regional nature run. Thus, the physical options included: the WRF Single-Moment 6-class microphysics scheme (WSM-6), the Yonsei University planetary boundary layer model (YSU PBL), and the Grell-Devenyi ensemble cumulus parameterization scheme. Other model parameters were the same as in the regional nature run.

(4) Conduct data assimilation experiments

Two data assimilation experiments were conducted with different observational sampling strategies to investigate the potential impact of the simulated DWL wind profiles on tropical cyclone track and intensity forecasts. The WRF 3DVAR system was used to assimilate the DWL wind profiles. Corresponding to the two configurations of the simulated observations, two data assimilation experiments were performed: the first was an ideal experiment that did not consider cloud influence. The other was a more realistic experiment in which observations contaminated by clouds were eliminated. The model domain configuration and physics options for both of these two experiments were the same as those used in the control run. Cycled data assimilation was performed and subsequent forecasts were generated.

(5) Verification

Analysis and forecasting results from Step 4 were compared against the control run and high-resolution regional nature run generated in Step 1. The impacts of the data on track and intensity forecasts were evaluated.

Results from Zhang and Pu (2010) demonstrate the positive impacts of potential satellite DWL data on tropical cyclone track and intensity forecasts. Although the study represents only an early effort with regional OSSEs, the steps and outcomes from this paper provide a regional OSSE concept, with the steps shown in Fig. 5, except that calibration has not been done since the paper only presents a case study.

3.2 The Impact of Resolution and Errors in DWL Wind Measurements on the Numerical Prediction of a Tropical Cyclone

Following the early results of Zhang and Pu (2010) as mentioned above, an additional OSSE study was conducted with the WRF ARW model to evaluate the impact of the resolution and error of DWL data on the numerical prediction of a tropical cyclone.

In this case, OSSEs were conducted in an idealized hurricane case that had track and intensity changes (in terms of the time series of sea level pressure and maximum surface winds) similar to those of Hurricane Bill (2009) during 00 UTC 17 August to 00 UTC 21 August 2009 (Fig. 6). The nature run was generated by the WRF ARW model, with the physical parameterization options including YSU PBL, Thompson microphysics, KF cumulus, RRTM longwave, and the Dudhia shortwave radiation model. The available GTS observation types and observations for

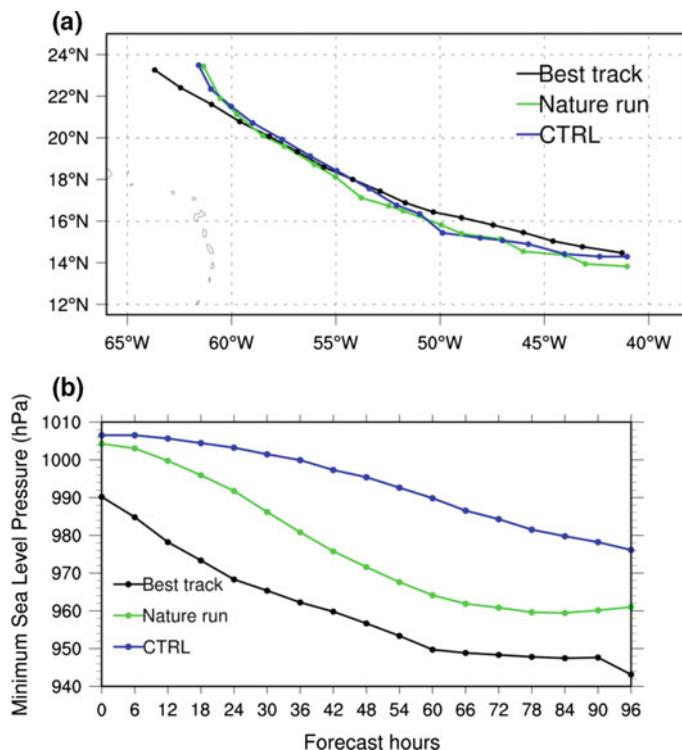


Fig. 6 Time series of **a** track and **b** minimum sea level pressure of simulated Hurricane Bill (green and blue curves), compared with the National Hurricane Center (NHC) best track data (black) during 00 UTC 17 August to 00 UTC 21 August 2009. In OSSE, two simulations were denoted as “Nature run” and control run (CTRL)

Hurricane Bill were used as a reference to generate similar types of simulated GTS data from the nature run. The control run used WRF ARW but with the MYJ PBL scheme, and the WSM6 microphysical scheme. A three-level nested domain was used with horizontal resolutions of 27 km, 9 km, and 3 km, and 41 vertical sigma levels. The model was initialized by the NCEP global forecast system (GFS) final (FNL) analysis at 12 UTC 16 August for the nature run (which was more similar to Hurricane Bill) and 0000 UTC 17 August for the control run. The different initial times and physical parameterization options in the control and nature runs were meant to represent the initial conditions and model errors to some extent. The control run was initialized by an analysis that assimilated GTS data with WRF 3DVAR.

Wind measurements were made by the polar-orbiting satellite, with data samples twice a day in the same regional domain. Figure 7 reveals the location of the measurements in the outermost model domain at 0600 UTC 17 August and 1800 UTC 17 August. Three different horizontal resolutions of wind measurements, 60 km, 180 km, and 360 km, were assumed. The vertical resolution of the data was

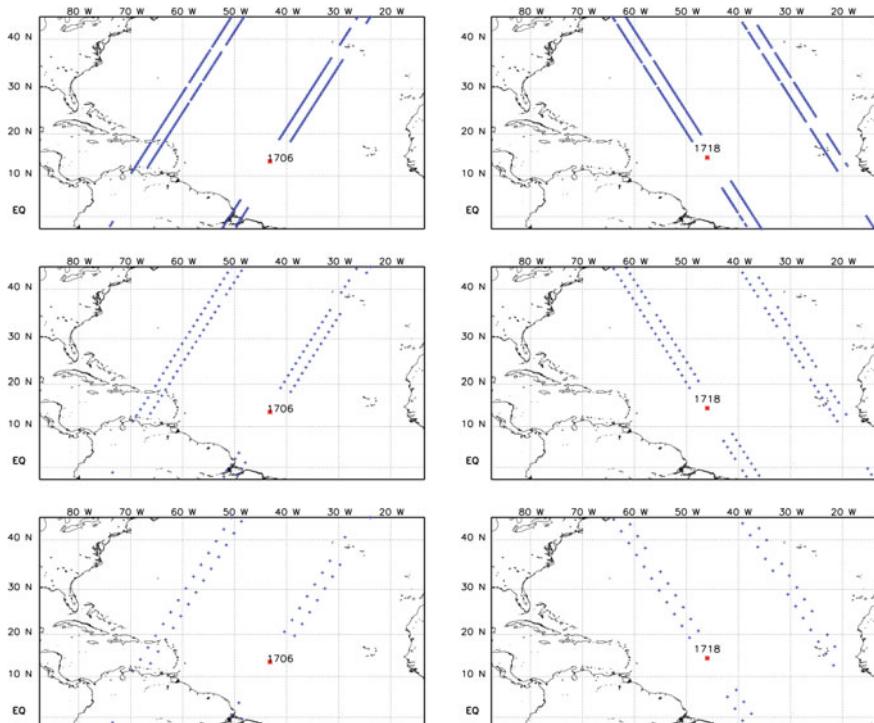


Fig. 7 Simulated DWL observation locations at 0600 UTC (left panels) and 1800 UTC (right panels) 17 August 2009. The resolution of measurements in top, middle and bottom panels are 60 km, 180 km, and 360 km, respectively. The red dot in each panel denotes the position of the hurricane center

assumed to be 250 m below 2 km and 1 km from 2 km up to 18 km. The cloud contaminations were considered for the DWL data availability (in terms of coverage). Similar to the earlier experiments, typical values for the standard deviation of DWL wind errors were 2 ms^{-1} below 2 km and 3 ms^{-1} above 2 km. No bias was assumed for the simulated DWL wind error.

Following the steps in Fig. 5, the following OSSEs were conducted.

(1) Impact of the resolution of DWL wind measurements on hurricane forecasts

Three data assimilation experiments were performed from 0000 UTC to 1800 UTC 17 August to assimilate GTS data and DWL wind measurements, and forecasts were then made until 0000 UTC 21 August. Figure 8 shows the track and track errors from different data assimilation experiments, compared against the control run and the nature run. It is obvious that the assimilation of satellite-based DWL data results in positive impacts on the hurricane track and intensity forecasts. Specifically, observations at the higher resolution are more beneficial to track forecasts, as the track errors produced by the experiment that assimilated data at 60 km are smaller than those generated by experiments that assimilated data at 180 km and 360 km. The errors in intensity (as revealed by minimum sea level pressure) confirm this conclusion (Fig. 8), while assimilation of wind measurements at higher resolution results in better intensity forecasts. Moreover, Fig. 9 compares the accumulated 3-h rainfall forecasts at 1200 UTC 19 August. It is clear that the assimilation of DWL

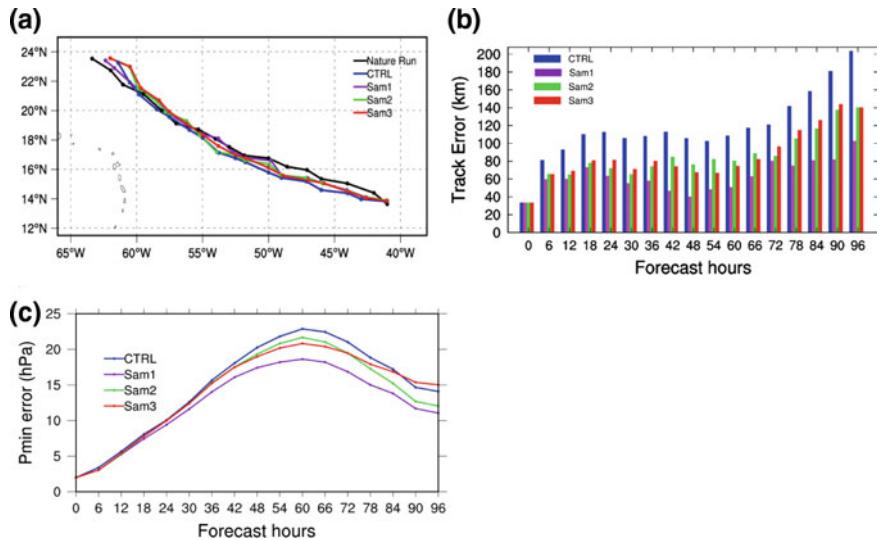


Fig. 8 Time series of **a** track, **b** track errors, **c** forecast errors in minimum sea level pressure during 00 UTC 17 August to 00 UTC 21 August 2009. The control experiment (CTRL, without assimilation of DWL observations) and simulations with assimilation of DWL observations at 60 km (Sam1), 180 km (Sam2) and 360 km (Sam3) resolution are compared against the nature run track and intensity

wind data results in better forecasts of vortex structure. Specifically, higher-resolution data (e.g., at 60 km) leads to a more realistic forecast of hurricane inner-core rainfall structure (Fig. 9c compared with Fig. 9a, b, d, e).

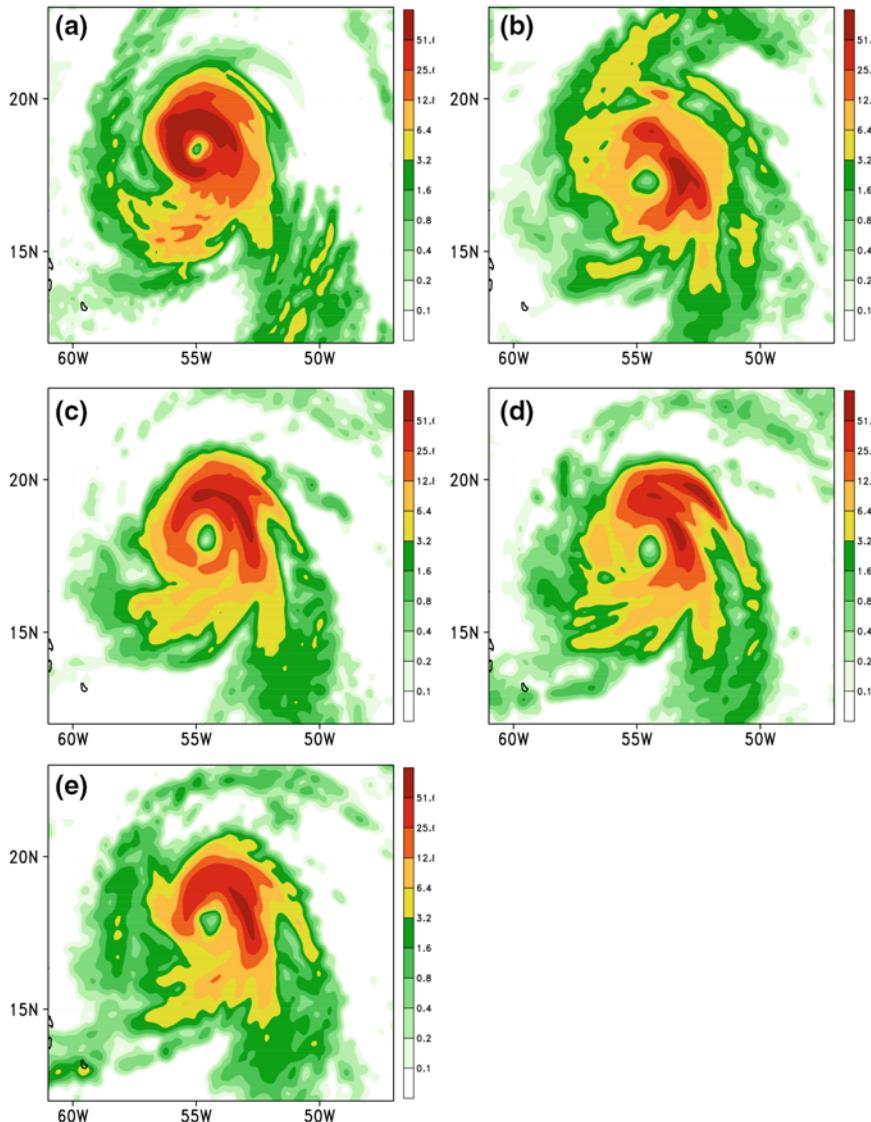


Fig. 9 Accumulated 3-h rainfall (unit: mm) at 1200 UTC 19 August from **a** Nature run, **b** CTRL, **c** Sam1, **d** Sam2, **e** Sam3. The control experiment (CTRL, without assimilation of DWL observations) and simulations with assimilation of DWL observations at 60 km (Sam1), 180 km (Sam2) and 360 km (Sam3) resolution

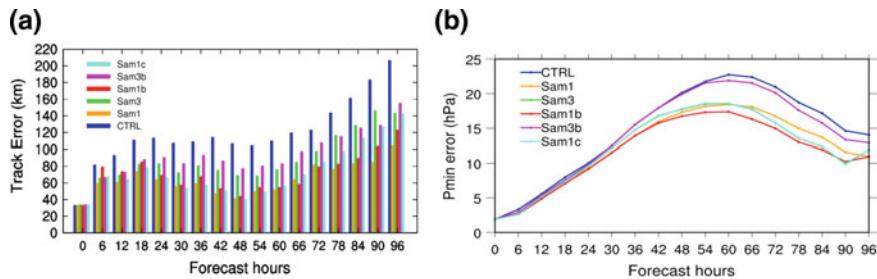


Fig. 10 Time series of **a** track error, **b** forecast errors in minimum center sea level pressure during 00 UTC 17 August to 00 UTC 21 August 2009. The control experiment (CTRL, without assimilation of DWL observations) and simulations with assimilation of DWL observations at 60 km (Sam1), 180 km (Sam2) and 360 km (Sam3) resolution are compared against the nature run track and intensity. The Sam1b and Sam1c are the same as Sam1, except for the observation error is increased at 50 % (Sam1b) and doubled (Sam1c), respectively. The experiment Sam3b is the same as Sam3, except for the observation error is increased at 50 %

- (2) The influence of measurement errors at different resolutions on hurricane forecasts

An additional set of OSSEs was conducted to test the influence of errors in DWL wind measurements on the analysis and forecasting of hurricanes. Three additional experiments were performed with the assimilation of DWL measurements with different errors: (1) data at 60 km resolution but with the standard deviation of the errors increased by 50 % at each vertical level; (2) data at 360 km resolution but with the standard deviation of the errors increased by 50 % at each vertical level; (3) data at 60 km resolution but with the standard deviation of the errors doubled at each vertical level. The data assimilation results compared with the nature run and the control run, as well as the experiments that were described in the previous section.

From Fig. 10, it is apparent that the measurements at higher resolution (e.g., 60 km) are more beneficial to forecast improvements (data assimilation versus control) than those at a coarser resolution. All experiments that assimilated higher-resolution data outperform those that assimilated coarser-resolution data regardless of the magnitude of the observation errors. *Compared with relatively accurate but coarser spatial resolution measurements, high spatial resolution measurements with modest random errors result in better analysis and forecasts of hurricane track and intensity.*

4 Recent OSSE Results with the HWRF Model: 3-D Wind Profiles Versus Ocean-Surface Winds

Following the maturity of global OSSEs (Atlas 1997) with DWL wind measurements (Stoffelen et al. 2006; Atlas and Emmitt 2008; Riishojgaard et al. 2012; Atlas et al. 2015a,b) and recent attempts (Pu et al. 2009; Zhang and Pu 2010) with

regional OSSEs, regional OSSEs have become more accepted as a tool to assess the data impacts from planned future instruments. In a recent study, Nolan et al. (2013) generated a set of regional nature runs with the WRF ARW model to support the community needs in regional OSSEs. In particular, this set of regional nature runs was created by high-resolution WRF ARW simulations (27 km/9 km/3 km/1 km horizontal resolution in a four-level nested WRF ARW domain) using the ECMWF global nature run (T511 NR) as the initial and boundary conditions for a hurricane case during 12 UTC 28 July to 1200 UTC 11 August. This nature run has now been used to support an upcoming satellite mission (planned to launch in October 2016), the Cyclone Global Navigation Satellite System (CYGNSS), which aims to improve tropical weather analysis and prediction.

Specifically, CYGNSS uses a constellation of eight small satellites carried into orbit on a single launch vehicle. In orbit, eight micro-satellite observatories receive both direct and reflected signals from Global Positioning System (GPS) satellites. The direct signals pinpoint CYGNSS observatory positions, while the reflected signals respond to ocean surface roughness, from which wind speed is retrieved. Because of the availability of wind speed information from CYGNSS, OSSEs were performed to assess the potential impact of CYGNSS surface wind speed data on tropical cyclone forecasts. The NCEP HWRF model Version 3.6 (Tallapragada et al. 2014; Atlas et al. 2015c) and a Gridpoint Statistical Interpolation (GSI) data assimilation system (Wu et al. 2002), representing advanced hurricane operational forecasting and data assimilation systems, were used for the OSSEs. The HWRF model was developed at the Environmental Modeling Center (EMC) at the National Centers for Environmental Prediction (NCEP) in collaboration with the NOAA Hurricane Research Division (HRD) and other partners. It has provided real-time tropical cyclone forecasts to the National Hurricane Center (NHC) for the Atlantic and eastern North Pacific basins since it became operational at NCEP in the 2007 hurricane season. A vortex initialization scheme (Liu et al. 2011) and the NCEP GSI data assimilation system were implemented with HWRF to provide initial conditions for HWRF. The purpose of the vortex initialization is to locate the vortex in its observed location. In order to do this, a position correction is done first, followed by an intensity correction process with adjustments to the moisture and thermodynamics fields. After that, data assimilation is performed using GSI with available conventional, radar, and satellite data. In this study, we used a version of HWRF model and GSI data assimilation system that are most close to the NCEP operational version of HWRF forecasting system in May 2015, but excluded a vortex relocation and intensity correction scheme in the OSSE in order to make the impact of assimilation of CYGNSS data on hurricane forecasting more clear.

The OSSEs for CYGNSS-simulated observations were conducted for multiple cases. A sample case shown in this chapter is chosen for a rapid intensification period of 00 UTC 01 August to 00 UTC 04 August 2005 from a nature run by Nolan et al. (2013). As shown in Fig. 11a-c, the CYGNSS data are available at 12 UTC, 15 UTC, and 18 UTC 01 August with good coverage. The first set of experiments was conducted to assimilate CYGNSS surface wind data into the HWRF model using GSI. For the control run, the initial and boundary conditions for the HWRF model

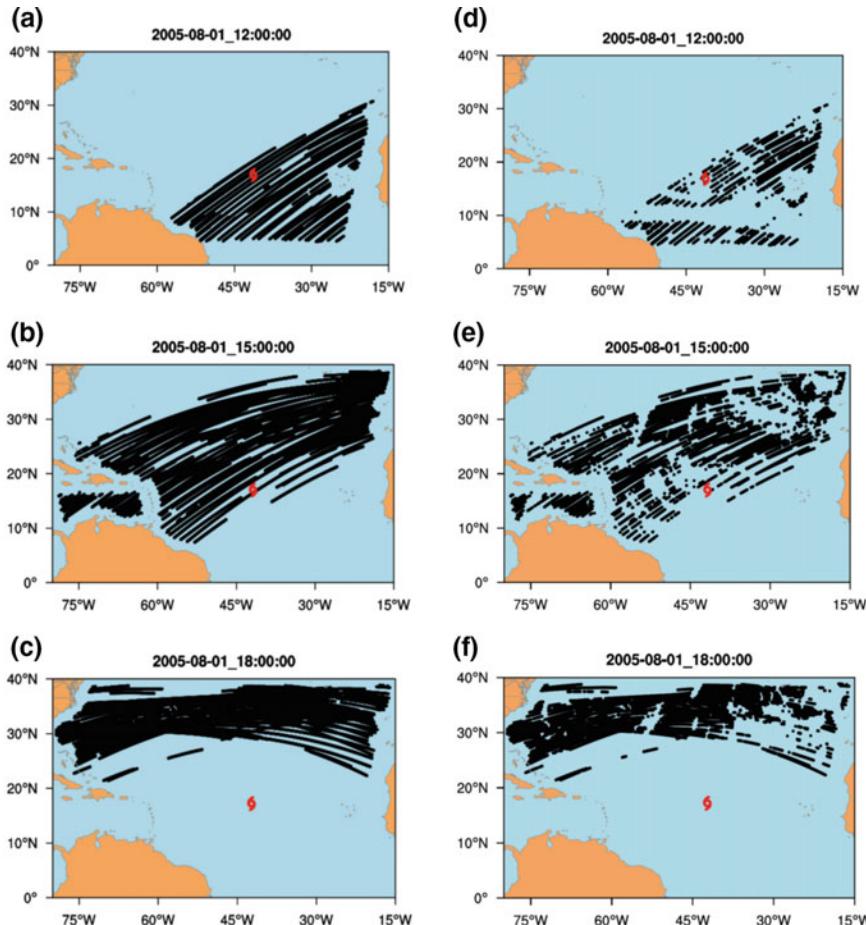


Fig. 11 The coverage and location of the CYGNSS observations at **a** 1200 UTC, **b** 1500 UTC, **c** 1800 UTC 01 August 2005. The center position of the hurricane was marked by a red sign. **d**, **e** and **f** are the same as **a**, **b**, **c** but for the locations of DWL wind profiles. Note that the lidar data were not available over the cloudy areas in **(d)**, **(e)** and **(f)**

were provided by the ECMWF T511 NR. For the data assimilation experiment, cycled data assimilation was performed at 12 UTC, 15 UTC, and 18 UTC, followed by a 72 h HWRF forecast. Results (Fig. 12) show that CYGNSS surface wind data have a positive impact on hurricane track and intensity forecasts, although the positive impact for the track is mostly in short-range forecasts (most times in the first 2 days). Considering the total amount of surface data available against the large degrees of freedom with the 3-D structure of the hurricane, this impact is significant.

In addition, in order to further evaluate the value of potential satellite-based 3-D DWL wind measurements, and also to compare the relative impact of CYGNSS surface wind and 3-D wind measurements, we use the same satellite swath of

CYGNSS but assume the measurements are 3-D wind profiles (extending from the surface to a height of \sim 18 km, with a vertical resolution at 250 m below 2 km height and 1 km above 2 km height) and also account the cloud effects for the DWL measurements (see data coverage in Fig. 11 d–f). Similar to the OSSEs with

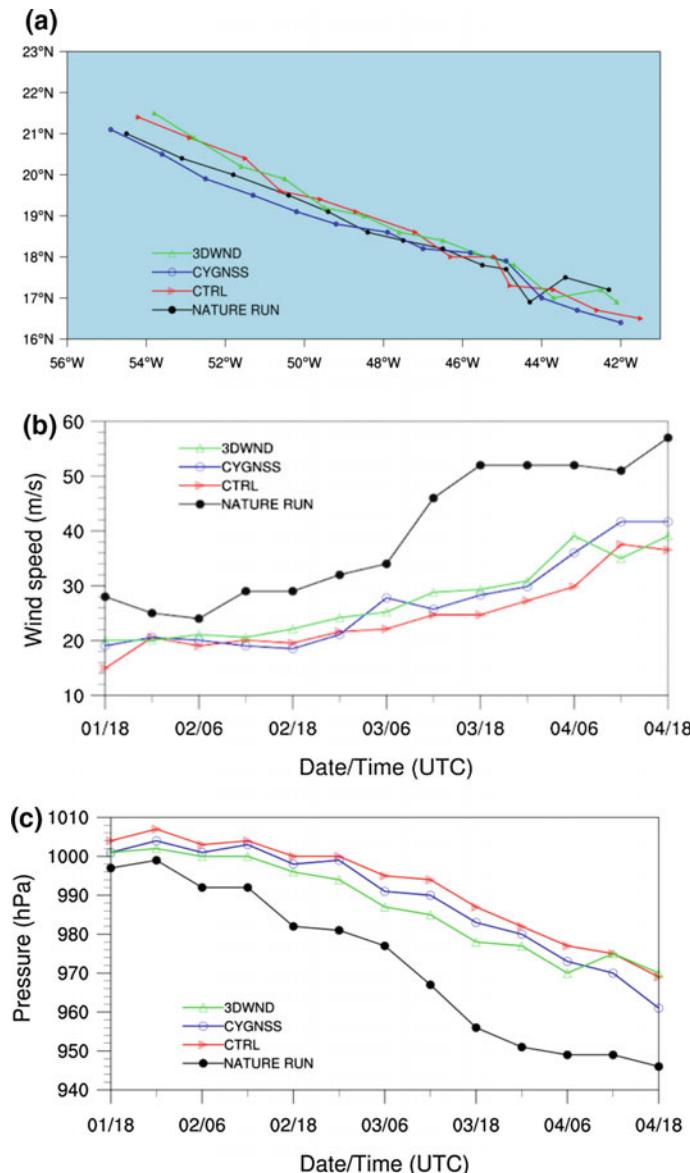


Fig. 12 Time series of **a** track and **b** minimum center sea level pressure of the hurricane from nature run, control (CTRL, red curve), and the experiments with assimilation of CYGNSS ocean surface wind (CYGNSS, blue curve) and 3D wind measurements (3D winds, green curve)

CYGNSS surface wind data, cycled data assimilation experiments were performed for the 3-D wind data. Results (Fig. 12) indicate that assimilation of the 3-D winds results in significant impacts on analysis and forecasts of the hurricane in terms of both track and intensity. The improvement to the forecasts is much more obvious and larger than in the experiments that only assimilated surface wind data. Moreover, assimilating 3-D winds ensures a reasonable vortex inner-core structure in the analyses and forecasts. Figures 13 and 14 reveal that the assimilation of 3-D winds results in better vortex structure in terms of surface wind and hurricane warm core. Specifically, assimilation of CYGNSS ocean surface wind improves the distribution and intensity of surface wind in analysis and forecast. Assimilation of 3-D wind results in even better representation of location and magnitude of the maximum surface wind (Fig. 13). In addition, assimilation of CYGNSS ocean surface wind has influence on the temperature field but it only helps resolving the realistic magnitude of warm-core in the low level of atmosphere, while assimilation of 3-D wind leads to significant changes in the temperature field in both low and upper levels and also results in a much better warm-core structure of the hurricane that is compatible with the nature run (Fig. 14).

Additional OSSE experiments were also performed for the period during the hurricane mature stage between 00 UTC 8 August and 00 UTC 11 August. A similar conclusion was obtained in terms of the impact of CYGNSS data and 3-D wind data on the analysis and forecast of the hurricane (details not shown).

5 Summary and Concluding Remarks

While many previous studies have emphasized OSSEs with DWL wind measurements using global NWP systems, in this chapter we gave an overview of our recent studies with regional NWP models in both real data assimilation experiments and OSSEs to demonstrate the value of DWL wind measurements in improving the forecast of severe weather systems (e.g., high-impact weather systems). Specifically, the impact of ground-based and airborne DWL on the numerical prediction of mesoscale convective systems and hurricanes was evaluated. The positive impacts of the data on the analysis and forecasts of high-impact weather systems have been shown. The regional OSSEs with both the WRF ARW and HWRF models also demonstrated potential positive impacts of satellite-based DWL measurements. It is found that more beneficial forecast impacts can be obtained from high spatial resolution observations with larger random errors compared with more precise observations obtained at a coarser resolution. The relative impact of satellite-based ocean-surface wind measurements and 3-dimensional profiles is compared, and the advantage of 3-D DWL wind measurements is evident.

Despite the positive impacts of DWL data on high-impact weather systems demonstrated in this chapter, the OSSEs presented in this study are mostly “quick OSSEs.” In several cases we did not completely use the conventional and satellite data that have already been available in the current operational system (although at

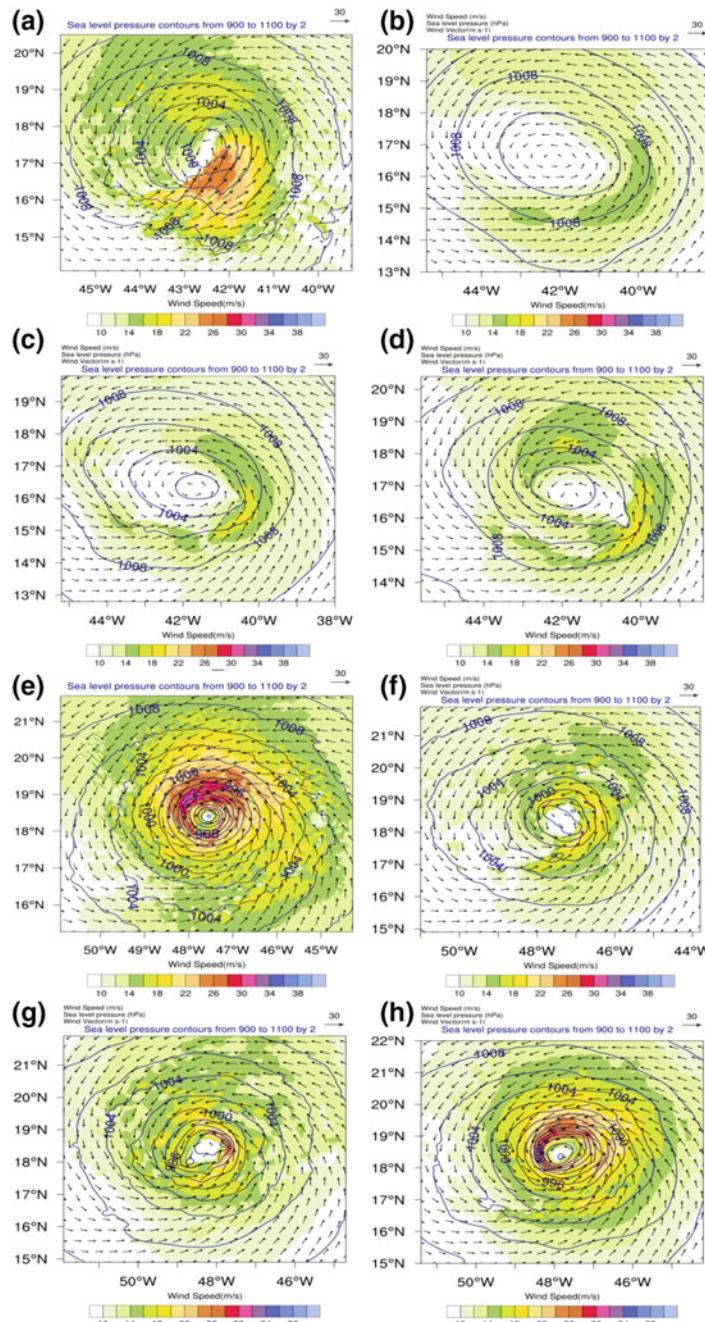


Fig. 13 Surface wind vectors and speeds (shaded contour) and sea-level pressure (contour) at 18 UTC 01 August 2005 from **a** Nature Run, **b** Control, **c** the experiment with assimilation of CYGNSS ocean surface winds and **d** the experiment with assimilation of 3-D winds. **e–h** are the same as **(a)–(d)** except for the 36 h forecast at 06 UTC 03 August 2005

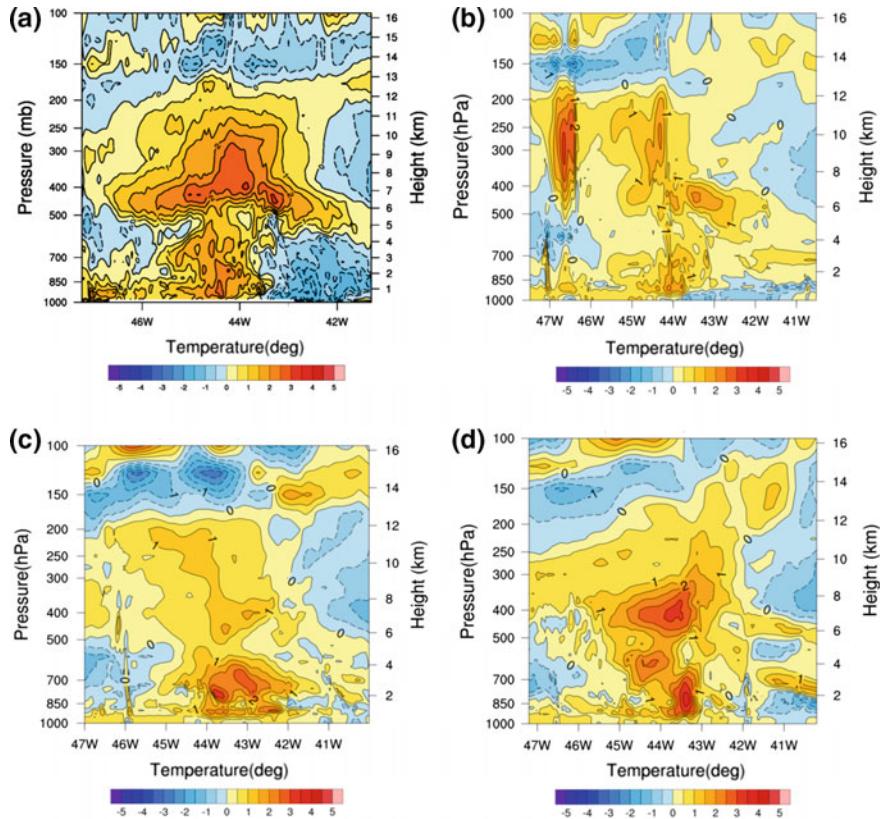


Fig. 14 West-East cross section of temperature ($^{\circ}\text{C}$) through the hurricane center at 06 UTC 02 August 2005 from **a** Nature Run, **b** Control, **c** the experiment with assimilation of CYGNSS ocean surface winds and **d** the experiment with assimilation of 3-D winds

the same time data are commonly sparse over the ocean). Therefore, a more careful calibration process (see Fig. 5) should be conducted within the OSSE framework in future studies. Nevertheless, all experiments consistently demonstrate the value of DWL measurements and prove that they are necessary in improving forecasts of high-impact weather systems. We anticipate more measurements and applications with DWL wind-profiling measurements in future research and operations, on all platforms, including ground-based, airborne and space-borne instruments.

Acknowledgements This study was supported by NASA Lidar Wind Science and Weather Programs. Some early results were also partially supported by U. S. Office of Naval Research and National Science Foundation. The computer resources from University of Utah's Center for High Performing Computer, NASA's High-End computing and NCAR Yellowstone computer are greatly appreciated.

The review comments from an anonymous reviewer were helpful for improving the manuscript.

References

- Atlas R, Kalnay E, Baker WE, Susskind J, Reuter D, Halem M (1985) Simulation studies of the impact of future observing systems on weather prediction. In: Proceedings of 7th AMS conference on numerical weather prediction, 17–20 (1985) Montreal, Quebec, Canada
- Atlas R (1997) Atmospheric observations and experiments to assess their usefulness in data assimilation. *J Meteor Soc Japan* 75:111–130
- Atlas R, Emmitt GD (2008) Review of observing system simulation experiments to evaluate the potential impact of lidar winds on numerical weather prediction. *ILRC24* 2:726–729. ISBN: 978-0-615-21489-4
- Atlas R, Emmitt GD, Terry J, Brin E, Ardizzone J, Jusem JC, Bungato D (2003) Recent observing system simulation experiments at the NASA DAO. AMS preprint volume for the seventh symposium on integrated observing system, 9–13 Feb 2003, Long Beach, CA
- Atlas R, Hoffman RH, Ma Z, Emmitt GD, Wood SA, Greco S, Tucker S, Bucci L, Annane B, Murillo S (2015a) Observing system simulation experiments (OSSEs) to evaluate the potential impact of an optical autocovariance wind lidar (OAWL) on numerical weather prediction. *J Atmos Ocean Technol* 32:1593–1613
- Atlas R, Bucci L, Annane B, Hoffman R, Murillo S (2015b) Observing system simulation experiments to assess the potential impact of new observing systems on hurricane forecasting. *Marine Technol Soc J* 49:140–148
- Atlas R, Tallapragada V, Gopalakrishnan S (2015c) Advances in tropical cyclone intensity forecasts. *Marine Technol Soc J* 49:149–160
- Barker DM, Huang W, Guo Y-R, Bourgeois AJ, Xiao Q (2004) A Three-dimensional variational (3DVAR) data assimilation system for use with MM5: implementation and initial results. *Mon Weather Rev* 132:897–914
- Baker WE, Coauthors (1995) Lidar-measured winds from space: a key component for weather and climate prediction. *Bull Am Meteor Soc* 76:869–888
- Baker WE, Atlas R, Cardinali C, Clement A, Emmitt GD, Gentry BM, Hardesty RM, Michael EK, Kavaya J, Langland R, Ma Z, Masutani M, McCarty W, Pierce RB, Pu Z, Riishojgaard LP, Ryan J, Tucker A, Weissmann M, Yoe JG (2014) Lidar-measured wind profiles: the missing link in the global observing system. *Bull Am Meteor Soc* 95:543–564
- Demoz B, Flamant C, Weckwerth T, Whiteman D, Evans K, Fabry F, Girolamo PD, Miller D, Geerts B, Brown W, Schwemmer G, Gentry B, Feltz W, Wang Z (2006) The dryline on 22 May 2002 during IHOP_2002: convective scale measurements at the profiling site. *Mon Weather Rev* 134:294–310
- Emmitt GD, Pu Z, Godwind K, Greco S (2011) Airborne Doppler Wind lidar data impacts on tropical cyclone track and intensity forecasting: the data processing, interpretation and assimilation. In: 15th symposium on integrated observing and assimilation systems for the atmosphere, oceans and land surface (IOAS-AOLS). 91st AMS annual meeting, 23–27 Jan 2011, Seattle, WA
- Gentry BM, Chen H, Li SX (2000) Wind measurements with 355-nm molecular Doppler lidar. *Opt Lett* 25:1231–1233
- Gentry BM, Chen H (2003) Tropospheric wind measurements obtained with the Goddard Lidar Observatory for Winds (GLOW): validation and performance.. In: Singh UN (ed) Lidar remote sensing for industry and environment monitoring II. International Society for Optical Engineering, SPIE Proceedings, vol 4484, pp 74–81
- Gentry B, McGill M, Schwemmer G, Hardesty M, Brewer A, Wilkerson T, Atlas R, Sirota M, Lindemann S, Hovis F (2010) The Tropospheric Wind Lidar Technology Experiment (TWiLiTE): status and future plans. Presented at the working group on space-based lidar winds, 24–26 Aug 2010, Bar Harbor, ME
- Huang XY et al (2009) Four-dimensional variational data assimilation for WRF: formulation and preliminary results. *Mon Weather Rev* 137:299–314

- Kalnay E, Liu J (2007) Adaptive observation strategies for lidar observations. In: AMS symposium on integrated observing systems, 13–17 Jan 2007, San Antonio, TX (also see similar presentation at <http://space.hsv.usra.edu/LWG/index.html>)
- Liu Q, Lord S, Tallapragada V (2011) Vortex initialization of the atmospheric model in HWRF. Hurricane Tutorial, Development Testbed Center. NCAR, Boulder, CO. http://www.dtcenter.org/HurrWRF/users/docs/presentations/tutorial2011/Lecture_2001_HWRF_Vortex.pdf
- Masutani M, Woollen JS, Lord SJ, Kleespies TJ, Emmitt GD, Sun H, Wood SA, Greco S, Terry J, Treadon R, Campana KA (2006) Observing system simulation at NCEP, NCEP office Note 451
- Masutani M, Coauthors (2010) Observing system simulation experiments at the National Centers for Environmental Prediction. *J Geophys Res* 115:D07101
- National Research Council (NRC) (2007) Earth science and applications from space: national imperatives for the next decade and beyond. The Academies Press, Washington DC, pp 2–14
- Nolan DS, Atlas R, Bhatia KT, Bucci LR (2013) Development and validation of hurricane nature run using the joint OSSE nature run and the WRF model. *J Adv Model Earth Syst* 5. doi:[10.1029/jame.20031](https://doi.org/10.1029/jame.20031)
- Pichugina YL, Banta RM, Brewer WA, Sandberg SP, Hardesty RM (2012) Doppler lidar-based wind-profile measurement system for offshore wind-energy and other marine boundary layer applications. *J Appl Meteor Climatol* 51:327–349
- Pu Z, Gentry B, Demoz B (2009) Potential impact of lidar wind measurements on high-impact weather forecasting: a regional OSSEs study. Preprints, 13th conference on integrated observing systems for atmosphere, ocean, and land surface (IOAS-AOLS), Phoenix, AZ. Am Meteor Soc 13.5. https://ams.confex.com/ams/89annual/techprogram/paper_150417.htm
- Pu Z, Zhang L, Emmitt GD (2010) Impact of airborne Doppler wind lidar profiles on numerical simulations of a tropical cyclone. *Geophys Res Lett* 37:L05801
- Reale O, Terry J, Masutani M, Andersson E, Riishojaard LP, Jusem JC (2007) Preliminary evaluation of the European Centre for Medium-Range Weather Forecasts' (ECMWF) Nature Run over the tropical Atlantic and African monsoon region. *Geophys Res Lett* 34:L22810. doi:[10.1029/L031640](https://doi.org/10.1029/L031640)
- Riishojaard LP, Ma Z, Masutani M, Woollen JS, Emmitt GD, Wood SA, Greco S (2012) Observation system simulation experiments for a global wind observing sounder. *Geophys Res Lett* 39:L17805. doi:[10.1029/2012GL051814](https://doi.org/10.1029/2012GL051814)
- Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker M, Duda KG, Huang XY, Wang W, Powers JG (2008) A description of the advanced research WRF Version 3. NCAR Tech. Note, NCAR/TN-475 + STR, 113 pp
- Stoffelen A, Pailleux J, Kallen E, Vaughan JM, Isaksen L, Flamant P, Wergen W, Andersson E, Schyberg H, Culoma A, Meynart R, Endemann M, Ingmann P (2005) The atmospheric dynamics mission for global wind field measurement. *Bull Am Meteorol Soc* 86:73–87
- Stoffelen A, Marseille GJ, Bouttier F, Vasiljevic D, Hann SD, Cardinal C (2006) ADM-Aeolus Doppler wind lidar observing system simulation experiment. *Q J R Meteorol Soc* 132:1927–1947
- Tallapragada V, Bernardet L, Biswas M, Gopalakrishnan S, Kwon Y, Liu Q, Marchok T, Sheinin D, Tong M, Trahan S, Tuleya S, Yablonsky R, Zhang X (2014) Hurricane Weather Research and Forecasting (HWRF) Model: 2014 scientific documentation. HWRF Development Testbed Center Tech. Rep. 99 pp http://www.dtcenter.org/HurrWRF/users/docs/scientific_documents/HWRFv3.6a_ScientificDoc.pdf
- Weissmann M, Busen R, Dörnbrack A, Rahm S, Reitebuch O (2005) Targeted observations with an airborne wind lidar. *J Atmos Oceanic Technol* 22:1706–1719
- Weissmann M, Cardinali C (2007) Impact of airborne Doppler lidar observations on ECMWF forecasts. *Q J R Meteorol Soc* 133:107–116
- Weissmann M, Langland RH, Cardinali C, Pauley PM, Rahm S (2012) Influence of airborne Doppler wind lidar profiles on ECMWF and NOGAPS forecasts. *Q J R Meteorol Soc* 138:118–130

- World Meteorological Organization (1996) Guide to meteorological instruments and methods of observation, 6th edn. WMO-No. 8, pp I.12–31, I.13–1
- Wu W-S, Purser J, Parrish D (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon Weather Rev* 130:2905–2916
- Zhang L, Pu Z (2010) An Observing System Simulation Experiment (OSSE) to assess the impact of Doppler wind lidar (DWL) measurements on the numerical simulation of a tropical cyclone. *Adv Meteorol Article ID 743863*, 14 pp. doi:[10.1155/2010/743863](https://doi.org/10.1155/2010/743863)
- Zhang L, Pu Z (2011) Four-dimensional assimilation of multi-time wind profiles over a single station and numerical simulation of a mesoscale convective system observed during IHOP_2002. *Mon Weather Rev* 139:3369–3388

A Three-Dimensional Variational Radar Data Assimilation Scheme Developed for Convective Scale NWP

Jidong Gao

Abstract A three-dimensional variational data assimilation scheme (3DVAR) has been developed for Convective scale NWP. In the scheme, a cost function is defined by a background term, an observation term, and a weak constraint term. The function is minimized through a limited memory, quasi-Newton conjugate-gradient algorithm. The background error covariance matrix, though simple, is modeled by a recursive filter. Furthermore, the square root of this matrix is used to precondition the minimization problem. In its original development, only radar radial velocity data could be assimilated. Recent developments for 3DVAR include the use of a model-derived diagnostic pressure equation constraint (DPEC) as a weak constraint, and the capability to assimilate reflectivity directly in the 3DVAR framework. The original radial-velocity-only 3DVAR method is applied to assimilate radial velocity observations considering beam broadening and earth curvature for an idealized supercell storm case, and real supercell storm cases. It is shown that the horizontal circulations, both within and around the storms, as well as the strong updraft and the associated downdraft, are well analyzed. The results also indicate that the method is quite insensitive to the effect of beam broadening, but very sensitive to the effect of earth curvature. So in the real data case studies, the effect of earth curvature is considered while beam broadening is not. Based on this 3DVAR framework, a real-time, weather-adaptive analysis system has been developed for the NOAA Warn-on-Forecast (WoF) project to incorporate all available radar observations within a moveable analysis domain. The system performed very well within the NOAA Hazardous Weather Testbed Experimental Warning Program during preliminary testing in recent years when many severe weather events were successfully detected and analyzed. The impact of DPEC on radar data assimilation is examined primarily in the context of storm forecasts. It is found that the experiments using DPEC generally predict higher low-level vertical vorticity near the time of observed tornados than the experiments not using DPEC. Finally, the impact of assimilating both radar reflectivity and radial velocity data with an

J. Gao (✉)

NOAA/National Severe Storms Laboratory, National Weather Center, 120, David, L. Boren Blvd., Norman, OK 73072, USA
e-mail: jidong.gao@noaa.gov

intermittent 3DVAR system is explored using an idealized thunderstorm case. It is found that by assimilating reflectivity data using simple hydrometer classification while also assimilating radial velocity data, the model can reconstruct the supercell thunderstorm quickly and the quality of analyses are improved compared to two other experiments without reflectivity and hydrometer classification. This paper represents the author's research efforts in radar data assimilation for convective scale NWP during the past several years.

1 Introduction

The assimilation of radial velocity data into storm-scale NWP models is relatively easy due to the linear relationship between radial velocity and three components of the wind field, which are prognostic variables in NWP models. Three-dimensional variational data assimilation methods provide relatively simple and efficient tools for this purpose. The variational formulation for NWP was introduced first by the pioneering work of Sasaki (1955), and was further reformulated in 1970 (Sasaki 1970a, b, c). Sasaki was also the first scientist to use the variational method to retrieve two-dimensional wind fields from single Doppler radar radial velocity observations (Sasaki et al. 1989). Since then, numerous variational methods including adjoint techniques have been proposed to initialize NWP models (Lewis and Derber 1985; LeDimet and Talagrand 1986; Talagrand and Courtier 1987; Thacker and Long 1988; Daley 1991; Sun and Crook 1997; Rabier et al. 2000; Xiao et al. 2005 Lewis et al. 2006) and to do wind analyses (Sun et al. 1991; Xu and Qiu 1994, 1995; Xu et al. 2001; Qiu and Xu 1996; Sun and Crook 2001; Gao et al. 1999, 2002, 2004).

Among the above-mentioned studies, Gao et al. (1999) proposed a variational approach to do wind analysis in Cartesian coordinates. This approach allowed flexible use of radar data in combination with other information (e.g., soundings, or VAD profiles, etc.) as well as the use of the mass continuity constraint and a smoothness constraint through a definition of a cost function. Their formulation used the theoretical principle described in Sasaki (1970a, b, c). In particular, it applied the anelastic mass conservation equation as a “weak constraint” (terminology introduced by Sasaki 1970a), allowing the severe error accumulation in the vertical velocity to be reduced because explicit integration of the anelastic continuity equation is avoided. The method performs well in both idealized OSSE and real data cases. However, there exist some difficulties in specifying the optimal weighting for each constraint and in determining how far one radial wind observation should spread to nearby model grid points. In another more standard 3DVAR development, the background error matrix is modeled by a recursive filter, and the square root of the matrix is used for preconditioning. Using the recursive filter is a simple and efficient way to spread the effect of each radar observation to the analyzed grid points (Wu et al. 2002). The aim of the preconditioning procedure is to decrease the number of iterations in the minimization process for obtaining the

optimal solution of the analysis. The weight assigned to each constraint is specified according to its assumed error deviation, so that each weighting has a physical meaning, instead of being chosen somewhat arbitrarily based on experience and trial experiments, as done in Gao et al. (1999).

How to best assimilate reflectivity data in addition to radial velocity data in variational methods remains an open question. One of the greatest difficulties in reflectivity assimilation is the uncertainty in the reflectivity forward operators which link the model hydrometeor variables with radar observed reflectivity. These uncertainties occur because of the complex features of numerical model microphysical schemes. Another difficulty is that the reflectivity forward model is highly nonlinear, which often leads to violation of some basic assumptions used in data assimilation methods, such as assuming Gaussian error distributions (Daley 1991).

Because of the above difficulty, a cloud analysis scheme named the Local Analysis and Prediction System (LAPS) that analyzes hydrometer variables and adjusts in-cloud temperatures was developed by Albers et al. (1996). Zhang et al. (1998) modified the scheme to make it more suitable for convective scale thunderstorms. Using this modified cloud analysis method, Hu et al. (2006a, b), and Schenkman et al. (2011) have shown reasonable success in simulating and forecasting convective storms, including tornadoes and supercells, when radial velocity data were also assimilated using the 3DVAR method (Gao et al. 2004). Hu et al. (2006a) indicated that reflectivity data play a larger role in the assimilation than radial velocity data for some cases. The adjusted in-cloud temperature has also been used to initialize an NWP model with a digital filter technique as in Weygandt and Benjamin (2007). However, this cloud analysis scheme by necessity includes a number of empirical relationships between the hydrometer variables. Many uncertainties exist in these parameter settings and these uncertainties may limit its value for storm-scale data assimilation and NWP as discussed by Gao et al. (2009b).

Sun and Crook (1997) were the first to realize the importance of assimilating both reflectivity and radial velocity in a more quantitative way in their four-dimensional variational data assimilation (4DVAR) scheme. They compared several ways to assimilate reflectivity data using Observing System Simulation Experiments (OSSEs). They found that the best way to assimilate reflectivity data was to transform it into equivalent rainwater mixing ratio, and then assimilate the rainwater into the cloud-scale model. However, their experiments neglected ice microphysics. They also tried to assimilate reflectivity directly into their model, but with mixed results, because directly assimilating reflectivity increases the nonlinearity of the 4DVAR problem. Sun and Crook (1997) also showed that the computational cost and strong nonlinearities with model microphysics, including ice microphysics, often cause difficulties in the 4DVAR assimilation of radar data. The ensemble Kalman filter (EnKF) is another advanced method for directly assimilating radar reflectivity and radial velocity data into model initial conditions

(Zhang et al. 2004; Tong and Xue 2005; Dowell et al. 2011; Yussouf et al. 2013 and many others). Assimilating both reflectivity and radial velocity data into a storm-scale NWP model can generally improve the quality of the analysis and forecast, as seen in both idealized case and real data cases. However, several problems remain. Dowell et al. (2011) found that when a reflectivity observation is assimilated, bias errors in the model fields associated with reflectivity (rain, snow, and hail–graupel) can be projected onto other model variables through the ensemble covariances, leading, for example, to temperature analyses being very sensitive to ensemble spread and the characterization of low-level cold pools being unreliable when obtained through reflectivity-data assimilation.

Considering the advantages and shortcomings based on the above studies, we believe that the direct assimilation of both radar reflectivity and radial velocity data in a three-dimensional variational (3DVAR) framework for convective scale models has considerable merits, in terms of computational efficiency and fidelity of results. For example, 3DVAR is computationally efficient compared to 4DVAR and EnKF (Gao et al. 1999, 2004, 2009a; Ge and Gao 2007; Ge et al. 2010; Hu et al. 2006a, b; Xiao et al. 2005). In Xiao et al. (2005), a 3DVAR radar reflectivity data assimilation scheme was developed with the model total water mixing ratio used as a control variable. A warm-rain process, its linear version, and its adjoint were incorporated into the 3DVAR system to partition the moisture and hydrometeor increments. Using the onshore Doppler radar data from Jindo, South Korea, they showed the positive impacts of assimilating radar reflectivity data on short-range quantitative precipitation forecasts. However, similar to Sun and Crook (1997), the ice microphysical variables, such as ice, snow, and hail, were not included as control variables. This simplification limited the value of this approach when applied to deep convective storms.

In Gao and Stensrud (2012), ice microphysical variables (i.e., snow and hail mixing ratios) were included as control variables for the cost function defined in a 3DVAR method. A forward operator for reflectivity was developed by using a background temperature field from a NWP model as guidance for the automatic classification of hydrometeor types. It was the first time a hydrometeor classification was used in combination with storm-scale variational data assimilation. While previous research for classification of hydrometeor types was largely based on radar observations only (Zrnic et al. 2001), the inclusion of environmental information improves the hydrometeor classification at the surface (Elmore 2011).

This review paper reports on some research results for convective scale radar data assimilation related to the 3DVAR framework developed mainly by the author over the past several years. In Sect. 2, a brief description of the 3DVAR method with both radial velocity and the reflectivity forward operators is provided. Several case studies about these forward operators and the use of model derived diagnostic pressure equation constraint (DPEC) as a weak constraint in the data assimilation scheme are reported in Sect. 3. Section 4 contains a summary and future works.

2 Description of Data Assimilation Method

(a) The Definition of the Cost Function

The 3DVAR system is designed and developed especially for radar data assimilation at the convective scale (Gao et al. 1999, 2002, 2004, 2013; Hu et al. 2006a, b; Stensrud and Gao 2010; Ge et al. 2010, 2012). The method applies weak constraints that are suitable for convective storms in a different manner than that developed for large-scale applications. A 3DVAR system starts from a first guess, or background, that is often provided by a forecast model, and adjusts the first guess fields as observations are assimilated. The resulting analysis is a blend of the model background fields and the observations. The process is influenced by assumed constraints and is determined by minimizing a cost function using numerical techniques. The resulting analysis can be used to initialize NWP models. The 3DVAR system described here is currently designed to initialize the Advanced Regional Prediction System (ARPS) and the Weather Research and Forecasting (WRF) models.

Following Gao et al. (2004), the standard cost function of 3DVAR can be written as,

$$J(\mathbf{x}) = J_B + J_O + J_C = \frac{1}{2} (\mathbf{x} - \mathbf{x}^b)^T \mathbf{B}^{-1} (\mathbf{x} - \mathbf{x}^b) + \frac{1}{2} [H(\mathbf{x}) - \mathbf{y}^o]^T \mathbf{R}^{-1} [H(\mathbf{x}) - \mathbf{y}^o] + J_C(\mathbf{x}), \quad (1)$$

where the first term on the right hand side measures the departure of the analysis vector, \mathbf{x} , from the background, \mathbf{x}^b , weighted by the inverse of the background error covariance matrix \mathbf{B} . In an earlier version of 3DVAR system (Gao et al. 2004; Hu et al. 2006a, b), the analysis vector \mathbf{x} contains the three wind components (u , v , and w), potential temperature (θ), pressure (p) and water vapor mixing ratio (q_v). Recently, the hydrometeor-related model variables, including the mixing ratios for rain water (q_r), snow (q_s) and hail (q_h), are added to the analysis vector for the purpose of assimilating reflectivity in variational framework. The second term on the right hand side is observation term, which measures the departure of the analysis from the observation vector, \mathbf{y}^o . In this study, \mathbf{y}^o includes both radar radial velocity and reflectivity. The forward model, $H(\mathbf{x})$, for both quantities is defined later. The observation term is weighted by the inverse of observation error covariance matrix \mathbf{R} that includes both instrument and representativeness errors. Term $J_C(\mathbf{x})$ in Eq. (1) represents the dynamic or equation constraints and it will be discussed in subsection 2(c).

(b) Radial Velocity Forward Operator

In the second term, radar radial velocity data is part of the observation vector \mathbf{y}^o in Eq. (1). The radar forward observation operator for radial velocity which including the effect of earth curvature is written as follows in Doviak and Zrnic (1993) as

$$v_r = \frac{dh}{dr} w + \frac{ds}{dr} (u \sin \phi + v \cos \phi), \quad (2)$$

where v_r is the projected radial velocity, r is the slant range (ray path distance), h is the height above the curving earth's surface, s is the distance along the earth's surface, and ϕ is the radar azimuth angle.

If beam broadening is also considered in the radar forward observation operator in the vertical direction, the observation operator for mapping v_r derived from (2) with multiple vertical model levels onto elevation angles (Rihan et al. 2008) is formulated as:

$$v_{r,e} = (\sum G v_r \Delta z) / (\sum G \Delta z) \quad (3)$$

where $v_{r,e}$ is the radial velocity on an elevation angle, Δz is the vertical model grid spacing. G describes the two-way power gain distribution within the radar beam and is formulated as $G = e^{-4 \ln 4 \alpha^2 \beta^2}$ (Wood and Brown 1997) with α as the distance from the center of the radar beam in radians and β as the one degree beam width. The summation is over vertical model grid points enclosed by the half-power beam lobe.

When only radial velocity is used, the observation error variance is usually set to be between 1 m s^{-1} , and 2 m s^{-1} which is the typical radar instrumental error (Doviak et al. 1976; Miller and Sun 2003), we choose the former in this study. The second term in (1) for radial velocity can be written as

$$J_{OV_r} = \frac{1}{2} [H(v_{r,e}) - v_r^{ob}] \mathbf{R}_{v_r}^{-1} [H(v_{r,e}) - v_r^{ob}]. \quad (4)$$

The $H(v_{r,e})$ defines a process which transfers model variable u , v , w into the radial velocity v_r using the forward operator defined by (2), and/or (3), and a tri-linear interpolation operator used to interpolate the radial velocity v_r from model grid points to radar observation locations. However, different from the 3DVAR developed at NCAR for WRF model (Barker et al. 2004) designed mainly for synoptic and mesoscale data assimilation, our 3DVAR can use multiple analysis passes that have different spatial influence scales. The use of multiple passes is found to be advantageous for analyzing convective storms (Hu et al. 2006b; Gao et al. 2009a; Schenkman et al. 2011).

(c) Reflectivity Forward Operator

As discussed before, in early versions of the 3DVAR program, reflectivity data are not assimilated directly. Instead, it is used in a cloud analysis scheme to adjust the hydrometeor variables and in-cloud temperature and moisture fields (Hu et al. 2006a, b). Gao and Stensrud (2012) present a method for directly assimilating reflectivity by adding an observation term and defining a reflectivity observation operator. In this approach, a modified forward operator for radar reflectivity is developed which uses a background temperature field from a numerical weather prediction model for hydrometeor classification.

The forward model for equivalent radar reflectivity factor for simulated storms is obtained by summing the contributions from three hydrometeor mixing ratios—rain, snow and hail—using the following formulation (Lin et al. 1983; Gilmore et al. 2004; Dowell et al. 2011)

$$Z_e = Z(q_r) + Z(q_s) + Z(q_h). \quad (5)$$

The rain component of the reflectivity is calculated, based on Smith et al. (1975), using

$$Z(q_r) = 3.63 \times 10^9 (\rho q_r)^{1.75}, \quad (6)$$

where ρ is atmospheric density. If the temperature is cooler than 0 °C, then the component of the reflectivity from dry snow is calculated using

$$Z(q_s) = 9.80 \times 10^8 (\rho q_s)^{1.75}. \quad (7)$$

For wet snow, which occurs at temperatures warmer than 0 °C, the reflectivity is calculated using

$$Z(q_s) = 4.26 \times 10^{11} (\rho q_s)^{1.75}. \quad (8)$$

For hail, the reflectivity formulation based on default Lin et al. (1983) and Gilmore et al. (2004) is used, such that

$$Z(q_h) = 4.33 \times 10^{10} (\rho q_h)^{1.75} \quad (9)$$

The assimilation of reflectivity observations using Eqs. (5)–(9) into the numerical model to obtain hydrometeor variables, such as rain water (q_r), snow (q_s) and hail mixing ratios (q_h), is less than ideal since the reflectivity factor used here is a function of all three hydrometeor variables. This leads to the solution being possibly underdetermined. For example, it is possible to obtain a non-zero snow water mixing ratio in the low levels of the model where only rain water are expected because of the very warm temperatures at these levels. To solve this problem, the forward reflectivity operator Eq. (5) can be modified to use information from the model background such that,

$$Z_e = \begin{cases} Z(q_r) & T_b > 5^\circ C \\ Z(q_s) + Z(q_h) & T_b < -5^\circ C \\ \alpha Z(q_r) + (1 - \alpha)[Z(q_s) + Z(q_h)] & -5^\circ C < T_b < 5^\circ C \end{cases} \quad (10)$$

where α varies linearly between 0 at $T_b = -5$ °C and 1 at $T_b = 5$ °C, and T_b is the background temperature from a NWP model. Our experiments indicate that the equivalent reflectivity factor Z_e calculated using Eqs. (5) and (10) are quite similar in term of the precipitation patterns and reflectivity values produced. But in (10),

the a priori partitioning of the hydrometeor variables is made so that when (10) is used to assimilate reflectivity observations, the model background temperature can provide guidance about how much of the correction should occur in the rain water variable q_r and how much correction should occur in the snow variable q_s and hail variable q_h .

The last step for computing reflectivity is to convert equivalent reflectivity factor to the customary logarithmic scale using

$$Z_{dB} = 10 \log_{10} Z_e \quad (11)$$

where the units of Z_{dB} are dBZ. The logarithmic scale is convenient for use in observed precipitation and it also is convenient for specifying error variances in dB (Doviak and Zrnic 1993).

The second term in (1) for reflectivity can be written as,

$$J_{OZ} = \frac{1}{2} [H(Z_{dB}) - Z_{dB}^{ob}] \mathbf{R}_Z^{-1} [H(Z_{dB}) - Z_{dB}^{ob}]. \quad (12)$$

Similar to (4), the $H(Z_{dB})$ in (12) defines a process which transfers model variable q_r, q_s, q_h into reflectivity Z_{dB} using the forward operator defined from (5) to (11), and a tri-linear interpolation operator used to interpolate reflectivity from model grid points to radar observation locations. The quality control for both radial velocity and reflectivity also includes buddy checking, velocity dealiasing for radar data and removal of anomalous propagation returns and ground clutter.

(d) Weak Constraint J_c

In the 3DVAR, cross-correlations among state variables are not included in the background error covariance and certain balance between analysis variables is realized by incorporating weak constraints in the cost function as the J_c term in (1). Currently J_c includes two terms as defined in the following,

$$J_c = J_{MC} + J_{DP}. \quad (13)$$

A constraint is imposed on the analyzed wind components based on the anelastic mass continuity equation, such that

$$J_{MC} = \frac{1}{2} \lambda_c (\partial \bar{\rho} u / \partial x + \partial \bar{\rho} v / \partial y + \partial \bar{\rho} w / \partial z)^2, \quad (14)$$

where $\bar{\rho}$ is the mean air density at a given horizontal level, and the weighting coefficient, λ_c , controls the relative importance of this penalty term in the cost function. The value of $\lambda_c = 5.0 \times 10^{-4}$ is used for the current application. This value determines the relative importance of mass continuity equation constraint and its optimal value is determined through many numerical experiments in a

trial-and-error fashion (Gao et al. 1999; Sun and Crook 2001). Gao et al. (1999, 2004) found that this constraint is effective in producing suitable analyses of vertical velocity, as it builds up the relationship among the three wind components. As pointed out in Gao et al. (1999), using the mass continuity equation as a weak instead of a strong constraint avoids the error accumulation associated with the explicit vertical integration of the mass continuity equation as often used in conventional dual-Doppler wind synthesis schemes. Thus, thunderstorm updrafts can be more accurately analyzed and the analysis is less sensitive to the lower and upper boundary conditions than occurs when the mass continuity equation is used as a “strong constraint” (terminology introduced by Sasaki 1970a). The use of the weak mass continuity constraint links the three components of wind field by the 3DVAR method in response to the assimilation of the radial velocity observations.

The second term J_{DP} is the diagnostic pressure equation constraint (DPEC) term defined as follows (Ge et al. 2012),

$$J_{DP} = P(\mathbf{x})^T \mathbf{A}_P^{-1}(\mathbf{x}) \quad (15)$$

$$P \equiv \nabla \cdot \vec{E} \equiv -\nabla^2 p' - \nabla \cdot (\bar{\rho} \vec{V} \cdot \nabla \vec{V}) + g \frac{\partial}{\partial z} \left(\bar{\rho} \left[\frac{\theta'}{\theta} - \frac{p'}{\bar{\rho} c_s^2} + \frac{q_v'}{\varepsilon + \bar{q}_v} - \frac{q_v' + q_{liquid+ice}}{1 + \bar{q}_v} \right] \right) + \nabla \cdot \vec{C} + \nabla \cdot \vec{D}, \quad (16)$$

where,

$$\vec{E} = \frac{\partial(\bar{\rho} \vec{V})}{\partial t} = \vec{i} \frac{\partial(\bar{\rho} u)}{\partial t} + \vec{j} \frac{\partial(\bar{\rho} v)}{\partial t} + \vec{k} \frac{\partial(\bar{\rho} w)}{\partial t}, \quad (17)$$

$$\vec{V} = \hat{u} \hat{i} + \hat{j} \hat{v} + \hat{k} \hat{w}, \quad (18)$$

$$\vec{C} = \hat{i}(\bar{\rho} fv - \bar{\rho} \tilde{f} w) + \hat{j}(\bar{\rho} fu) + \hat{k}(\bar{\rho} \tilde{f} u), \quad (19)$$

$$\vec{D} = \hat{i} D_u + \hat{j} D_v + \hat{k} D_w. \quad (20)$$

The vector \vec{E} is the forcing term of the vector Euclidian momentum equation. The $q_{liquid+ice}$ includes hydrometeor mixing ratios. The \hat{i} , \hat{j} and \hat{k} are unit vectors in the x, y and z directions. The overbar represents base state and the primed variables are perturbations from a base state, c_s is the acoustic wave speed, and ε is the ratio of the gas constants for dry air and water vapor. The Coriolis coefficients are $f = 2\Omega \sin(\phi)$ and $\tilde{f} = 2\Omega \cos(\phi)$, where Ω is the angular velocity of the earth and ϕ is latitude. The terms, D_u , D_v and D_w contain the subgrid scale turbulence and computational mixing terms.

Equation (16) is derived by applying the divergence operator to the three momentum equations of the ARPS model (Xue et al. 2000, 2003):

$$\bar{\rho} \frac{\partial u}{\partial t} = -\bar{\rho} \vec{V} \bullet \nabla u - \frac{\partial p'}{\partial x} + (\bar{\rho} f v - \bar{\rho} \tilde{f} w) + D_u, \quad (21)$$

$$\bar{\rho} \frac{\partial v}{\partial t} = -\bar{\rho} \vec{V} \bullet \nabla v - \frac{\partial p'}{\partial y} + \bar{\rho} f u + D_v, \quad (22)$$

$$\begin{aligned} \bar{\rho} \frac{\partial w}{\partial t} &= -\bar{\rho} \vec{V} \bullet \nabla w - \frac{\partial p'}{\partial z} \\ &+ \bar{\rho} g \left[\frac{\theta'}{\theta} - \frac{p'}{\bar{\rho} c_s^2} + \frac{q_v'}{\varepsilon + \bar{q}_v} - \frac{q_v' + q_{liquid+ice}}{1 + \bar{q}_v} \right] + \bar{\rho} \tilde{f} u + D_w \end{aligned} \quad (23)$$

The \mathbf{A}_P in Eq. (15) is the error covariance matrix associated with the DPEC constraint, which is assumed to be diagonal with empirically defined constant diagonal elements as the variances. The inverse diagonal matrix is called the weighting coefficient and determines the relative importance of the DPEC constraint and its optimal value can be determined through numerical experiments, similar to the way to determine weighting coefficient for mass continuity constraint. Usually the constraint terms with their weights should be similar orders of magnitude as other terms in J for them to be effective.

(e) Variable Transfer and Precondition

To effectively precondition the minimization problem, we follow Courtier et al. (1994) and Courtier (1997) and define an alternative control variable \mathbf{v} , such that $\mathbf{Cv} = \sqrt{\mathbf{B}}\mathbf{v} = (\mathbf{x} - \mathbf{x}^b)$. This allows the cost function (1) to be changed into an incremental form, such that

$$J_{inc}(\mathbf{v}) = \frac{1}{2} \mathbf{v}^T \mathbf{v} + \frac{1}{2} (\mathbf{H}\mathbf{Cv} - \mathbf{d})^T \mathbf{R}^{-1} (\mathbf{H}\mathbf{Cv} - \mathbf{d}) + J_c(\mathbf{v}) \quad (24)$$

where \mathbf{H} is the linearized version of H and $\mathbf{d} \equiv \mathbf{y}^o - H(\mathbf{x}^b)$. The gradient and Hessian of J_{inc} can also be derived, with the former obtained by differentiating (24) with respect to \mathbf{v} , yielding,

$$\nabla J_{inc} = (\mathbf{I} + \mathbf{C}^T \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{C}) \mathbf{v} - \mathbf{C}^T \mathbf{H}^T \mathbf{R}^{-1} \mathbf{d} + \nabla J_c(\mathbf{v}) \quad (25)$$

where \mathbf{I} is the identity matrix. The Hessian then follows as

$$\nabla^2 J_{inc} = \mathbf{I} + \mathbf{C}^T \mathbf{H}^T \mathbf{R}^{-1} \mathbf{H} \mathbf{C} + \nabla^2 J_c(\mathbf{v}). \quad (26)$$

From (26), it is clear that the preconditioning prevents the smallest eigenvalue from becoming close to zero. This potentially could improve convergence of minimization algorithms and allows the variational problem to be solved more efficiently.

The matrix \mathbf{C} defined in (24) can be broken down as

$$\mathbf{C} = \mathbf{D}\mathbf{F}\mathbf{L} \quad (27)$$

where \mathbf{D} is a diagonal matrix consisting of the standard deviation of background errors and \mathbf{L} is a scaling factor. The matrix \mathbf{F} is the square root of a matrix with diagonal elements equal to one and off diagonal elements equal to the background error correlation coefficients. However, the matrix \mathbf{F} is too large to be used directly in the minimization calculations. Instead, it is modeled by a spatial recursive filter (Purser et al. 2003), which to first order is defined by

$$\begin{aligned} Y_i &= \alpha Y_{i-1} + (1 - \alpha) X_i & \text{for } i = 1, \dots, n \\ Z_i &= \alpha Z_{i+1} + (1 - \alpha) Y_i & \text{for } i = n, \dots, 1 \end{aligned} \quad (28)$$

where X_i is the initial value at grid point i , Y_i is the value after filtering for $i = 1$ to n , Z_i is the initial value after one pass of the filter in each direction and α is the filter coefficient given by the following formulation (Lorenc 1992),

$$\begin{aligned} \alpha &= 1 + E - \sqrt{E(E+2)} \\ E &= 2N\Delta^2/(4L^2) \end{aligned} \quad (29)$$

where L is the horizontal correlation scale, Δ is the horizontal grid spacing, and N is the number of filter passes to be applied. This is a first-order recursive filter, applied in both directions to ensure zero phase change. Multi-pass filters (N greater than unity) are built up by repeated application of (28). In this study, two passes of the recursive filter are used ($N = 2$). Xie et al. (2005, 2011) proved theoretically that the multiple passes approach with recursive filter is superior to the conventional single pass 3DVAR method.

3 Several Examples of Idealized and Real Data Case Studies

(a) Idealized Case for Effect of Beam Broadening

In this section, we evaluate the impact of beam broadening and earth curvature on 3DVAR data assimilation system using simulated data derived from ARPS model runs using Eqs. (2) and (3). Only mass continuity Eq. (14) is used as a weak constraint and model derived weak constraint (15)–(23) is not used here. The ARPS model is used in a 3D cloud model mode. The 20 May 1977 Del City, Oklahoma tornadic supercell storm is used to conduct several series of experiments. This storm has been thoroughly studied by multiple Doppler analysis and numerical simulation (Ray et al. 1981; Klemp et al. 1981; Klemp and Rotunno 1983).

The model is configured as the following: $67 \times 67 \times 35$ grid points and $1 \text{ km} \times 1 \text{ km} \times 0.5 \text{ km}$ grid intervals for the x, y, and z directions, respectively, so as to establish a physical domain of $64 \times 64 \times 16 \text{ km}$. The simulation starts with a modified sounding (as in Klemp et al. 1981) which favors the development of a supercell thunderstorm. The thermal bubble has a 4 K perturbation, and is centered at $x = 48 \text{ km}$, $y = 16 \text{ km}$ and $z = 1.5 \text{ km}$ with the lower-left corner of the domain as the origin. The radius of the bubble is 10 km in the x and y directions and 1.5 km in the z direction. The three-category ice microphysical scheme of Lin et al. (1983) is used together with a 1.5-order turbulent kinetic energy subgrid parameterization. Open boundary conditions are used for the lateral boundaries and rigid wall conditions for the top and bottom boundaries. An upper-level Rayleigh damping layer is also included to inhibit wave reflection from the top of the model.

The simulation runs for 2 h. The initial convective cell strengthens over the first 20 min and begins to split into two cells at around 1 h. To keep the right-moving storm near the center of the model domain, a mean storm speed ($U = 3 \text{ m s}^{-1}$, $V = 14 \text{ m s}^{-1}$) is subtracted from the sounding. At about 2 h into the simulation, the right mover is still near the center of the domain as expected and the left mover is located at the northwest corner. Figure 2a and Fig. 3a show horizontal and vertical cross sections of simulated wind, vertical velocity at 2 h respectively (vertical cross section is plotted through line A-B in Fig. 2a). A strong rotating updraft (with maximum vertical velocity exceeding 29 m s^{-1}) and associated low-level downdraft are evident near the center of the domain. The updraft tilts eastward in the upper part of the troposphere. The evolution of the simulated storm is qualitatively similar to that described by Klemp and Wilhelmson (1978).

The pseudo radar radial observations from two Doppler radars are obtained by sampling this simulated storm at 2 h using radar forward operators expressed in Eq. (2). The simulated data are obtained from the simulated wind field fixed at $t = 2 \text{ h}$, as a function of various radar locations. Of the two radars, one is put at $x = 33 \text{ km}$ relative to the origin of model domain (lower left corner), while its y coordinate is varied in increments of 10 km from $y = -190 \text{ km}$ to $y = 10 \text{ km}$. A second radar is set at position $y = 25 \text{ km}$ while its x coordinate is varied from $x = 0 \text{ km}$ to $x = -200 \text{ km}$ in intervals of 10 km. In this way, we are able to test the impact of the beam broadening and the earth curvature as a function of distance from the center of the storm ranging from about 20 km to 220 km.

The elapsed times for the radars to obtain the volume scans are neglected, and thus we assume that the radial wind observations are simultaneous. For simplicity, the two radars will cover the entire horizontal physical grids (i.e. $64 \times 64 \text{ km}$) which assumes that the radars sweep almost continuously in horizontal direction. The elevation angles are $0.5^\circ, 0.9^\circ, 1.3^\circ, 2.4^\circ, 3.1^\circ, 4.0^\circ, 5.1^\circ, 6.4^\circ, 7.5^\circ, 8.7^\circ, 10.0^\circ, 12.0^\circ, 16.7^\circ, 19.5^\circ$ (same as the WSR-88D convective precipitation volume coverage pattern, VCP 11). The simulated data are only specified in precipitation regions (where reflectivity is greater than zero dBZ). In order to simulate the radar measurement statistical error, 1 m s^{-1} random error (white noise) is added to the radial velocities in the pseudo observation data.

Table 1 List of data analysis/assimilation experiments

Name ^a	Radar distance	Description
CNTL1_xxx	20 km ~ 220 km at an interval of 10 km (xxx is the radar distance in km)	analyses at $t = 2$ h (21 experiments)
NoBB1_xxx		
NoCV1_xxx		

^aCNTL means both the effects of beam broadening and earth curvature are considered

NoBB means the effect of beam broadening are neglected

NoCV means the effect of earth curvature are neglected

Corresponding to the radial wind observations, three categories, 21 experiments each category, of data analysis experiments (see Table 1, which lists all experiments) will be conducted at $t = 2$ h with varied surface ranges between radar location and storm center. In the first category of experiments, both the effect of beam broadening and the effects of earth curvature are considered using the radar forward observation operator as defined in Eq. (2). They will be referred as CNTL1 experiments (label 1 means at single time level). In the second category of experiments, the effect of beam broadening is not considered and Eq. (3) will be replaced with a simple tri-linear interpolation scheme. It will be referred as NoBB1 experiments. In the third category of experiments, the effect of earth curvature will not be considered and Eq. (2) will be replaced with the commonly used Cartesian radar forward operator (Gao et al. 1999). It will be referred as NoCV1 experiments. The distance between the storm and the radar varys from 20 km to 220 km at an interval of 10 km for both radars. So each individual experiment will be referred by its category name followed by the distance in km, as described above, e.g. CNTL1_60, NoBB1_60, NoCV1_60, etc.

To compare the accuracy of the analysis from different experiments, the RMS error statistics of the horizontal wind components (V_h) and scalar model variables (s) between the experiments and the truth simulation run are computed. The computation of the RMS error statistics is only done over model grid points where the reflectivity (estimated from the local hydrometeor mixing ratios) of the simulation run is greater than 5 dBZ. As stated above, the purpose of these experiments is to test the impact of beam broadening and earth curvature on 3DVAR wind analysis. The variations of RMS errors for NoBB1 and NoCV1 are plotted in Fig. 1 along with that for CNTL1. The horizontal section at $z = 3.5$ km AGL and the vertical cross section at $y = 22.5$ km of wind fields for the truth simulation, CNTL1_60, NoBB1_60, NoCV1_60 and CNTL1_150, NoBB1_150, NoCV1_150 are plotted in Figs. 2, 3, 4 and 5.

We first discuss the impact of beam broadening. The RMS error of the horizontal winds and the vertical velocities plotted as a function of the distance for both CNTL1 (solid lines) and NoBB1 (dashed lines) experiments are shown in Fig. 1. It is found that the RMS error differences for both horizontal winds and vertical velocities between these 21 CNTL1 experiments and their corresponding NoBB1 experiments gradually increase as the distance between the storm center and radar locations increase. These differences are less than 0.35 m s^{-1} for horizontal winds

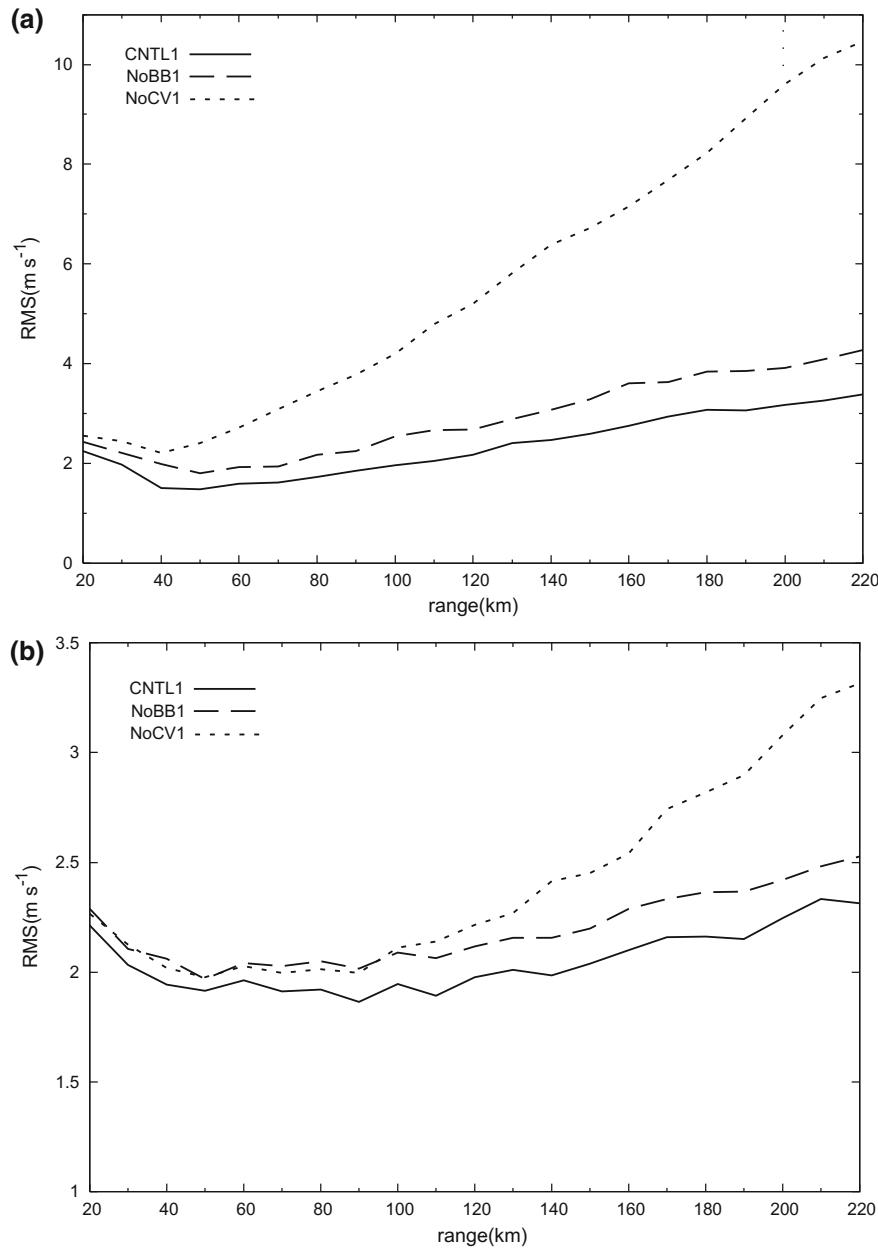


Fig. 1 The variation of RMS errors with the distance between the center of the storm and radar locations, for **a** horizontal wind, and **b** vertical velocity. The *solid lines* are for CNTL1 experiments, the *dashed lines* are for the NoBB1 experiments, the *dotted lines* are for the NoCV1 experiments (Courtesy of the American Meteorology Society)

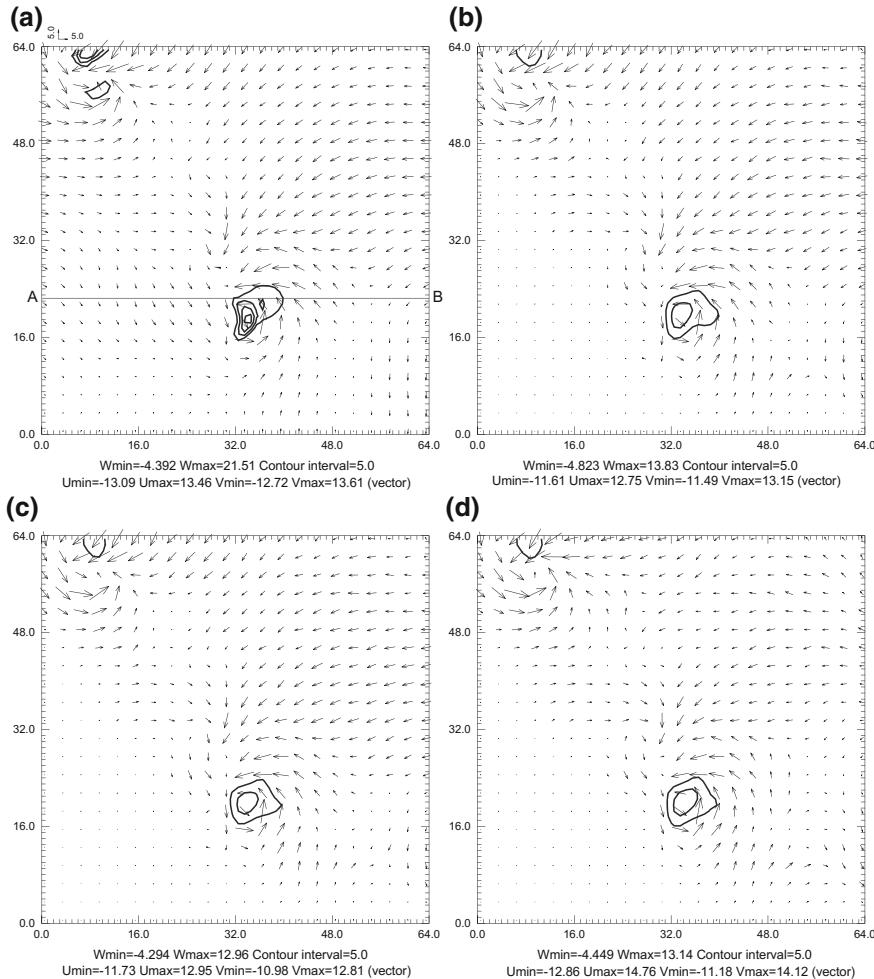


Fig. 2 Perturbation horizontal winds (vectors, m s^{-1}) and vertical velocity w (contours, m s^{-1}) at $t = 120$ min and 3.5 km AGL for **a** truth simulation; **b** CNTL1_60; **c** NoBB1_60; **d** NoCV1_60. The w contour starting from 5 m s^{-1} with an interval of 5 m s^{-1} (Courtesy of the American Meteorology Society)

and less than 0.1 m s^{-1} for vertical velocities within the range of 60 km. Beyond 60 km, the differences for horizontal winds become more noticeable as the range increases, reaching over 1 m s^{-1} at the range of 220 km, while the difference for vertical velocity shows little change. This means that additional error due to the neglect of beam broadening are gradually introduced in NoBB1 experiments but the maximum error is no more than the statistical error in the observations.

The variation in the RMS errors for horizontal winds and vertical velocities as a function of distance for experiment NoCV1 is also plotted in Fig. 1 in dotted lines.

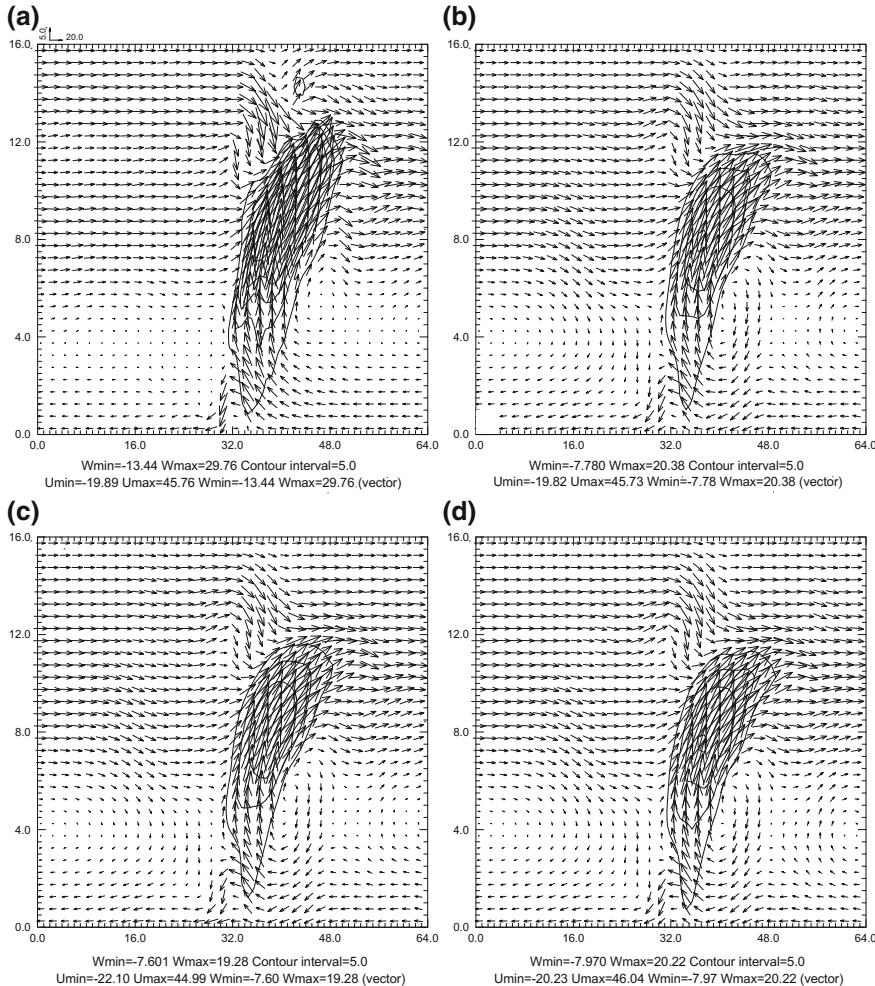


Fig. 3 Total u-w wind vectors and vertical velocity (contours) of the 20 May 1977 supercell storm at $t = 120$ min and $y = 22.5$ km (along the line A-B in Fig. 3a) for **a** truth simulation; **b** CNTL1_60; **c** NoBB1_60; **d** NoCV1_60 (Courtesy of the American Meteorology Society)

It is easily identified that the neglecting of the earth curvature can lead to very large RMS errors in the analysis of horizontal winds, especially beyond 60 km. It exhibits an additional 7.1 m s^{-1} RMS error of horizontal winds compared to CNTL1 experiment at the range of 220 km (Fig. 1a). The RMS error differences for vertical velocities between CNTL1 and NoCV1 experiments are evident when the surface range is over 150 km (Fig. 1b). So in the sense of the evolution of RMS errors, we can conclude that overlooking the earth curvature has a much greater negative impact on variational wind analysis than the neglect of beam broadening.

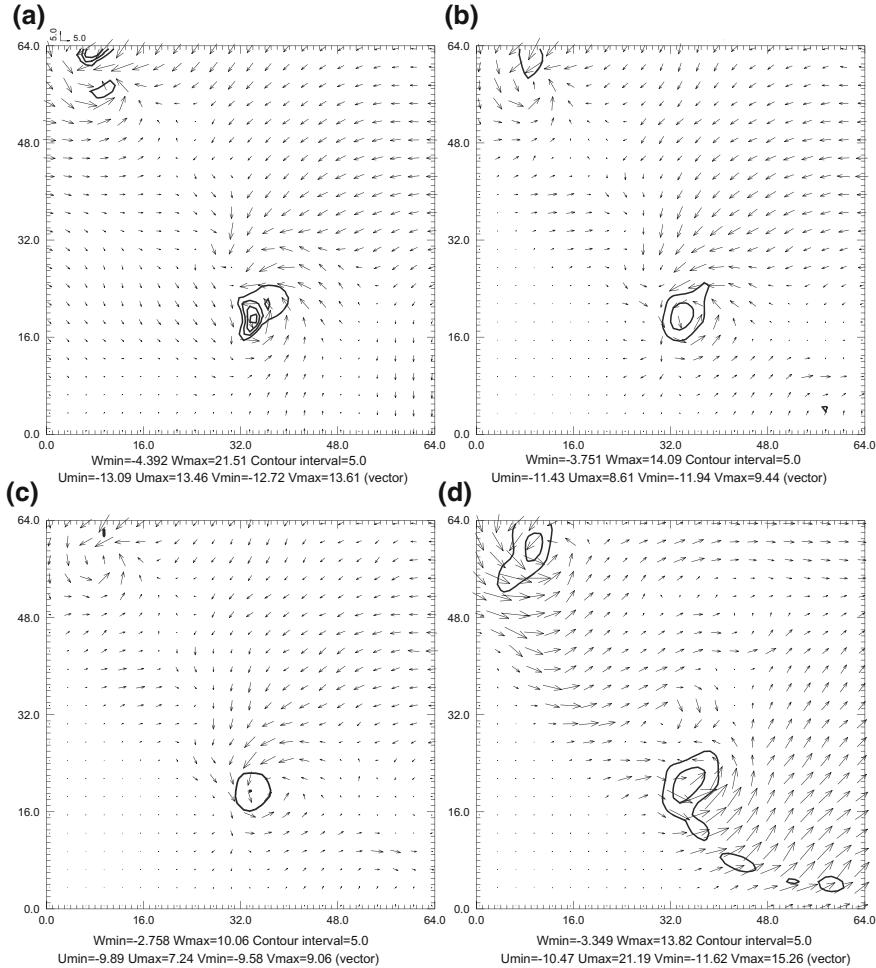


Fig. 4 Same as Fig. 3, but for **a** truth simulation; **b** CNTL1_150; **c** NoBB1_150; **d** NoCV1_150 (Courtesy of the American Meteorology Society)

As the RMS statistics suggest, the differences in the 3-D wind fields among all three categories of experiments CNTL1, NoBB1 and NoCV1 should be very small when the distance between the storm and radars is less than 60 km. Figure 2 and Fig. 3 confirm this conclusion. Figure 2 shows that the horizontal wind and vertical velocity fields at 3.5 km AGL for the truth simulation, CNTL1_60, NoBB1_60, and NoCV1_60 for the case where the radar is 60 km from the storm. Though the 3DVAR analysis is not perfect, the horizontal cyclonic rotation associated with the right and left movers are clearly evident in all three experiments (Fig. 2b-d). They are all pretty close to the truth simulation (Fig. 2a). The analyzed maximum vertical velocities (Fig. 3b-d) for all three categories of experiments are generally several

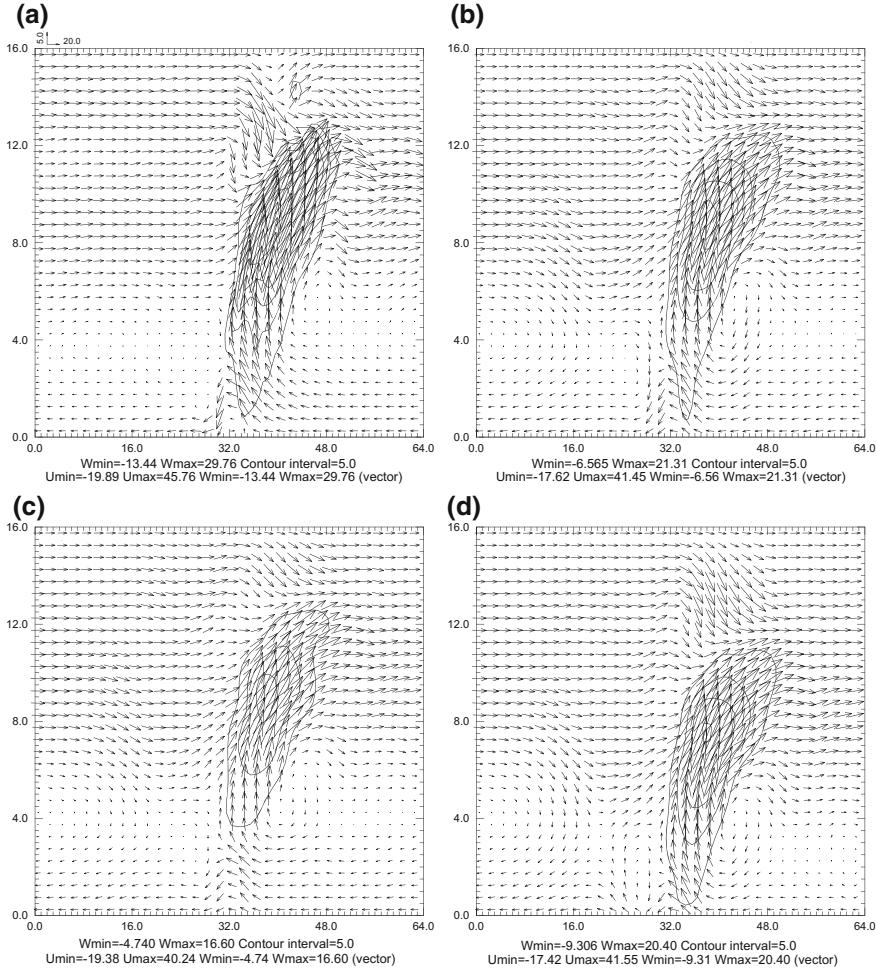


Fig. 5 Same as Fig. 4, but for **a** truth simulation; **b** CNTL1_150; **c** NoBB1_150; **d** NoCV1_150 (Courtesy of the American Meteorology Society)

meters per second weaker than the truth simulation, but the pattern is nearly the same for all three experiments. So the error from neglecting both beam broadening and earth curvature at this range is pretty small.

When the distance between the storm and radar location is 150 km or greater, the differences among these experiments become larger and can no longer be ignored. As an example, horizontal cross sections at $z = 3.5$ km and vertical cross sections are plotted as in Figs. 4 and 5 for the surface range of 150 km (the truth simulation is re-plotted for ease of comparison). It is clearly evident that the rotation signature near the center of the storm in Fig. 4b for CNTL1_150 is stronger than that in Fig. 4c for NoBB1_150. Also Fig. 5b shows a much stronger and deeper rotation

updraft than Fig. 5c. The maximum vertical velocity in Fig. 5b is 21.31 m s^{-1} , much closer to the simulation result (as shown in Fig. 5a) than that in Fig. 5c which is only 16.60 m s^{-1} . Apparently, CNTL1_150 experiment does a better job for the wind analysis than NoBB1_150 in which no effect of beam broadening is considered.

For experiment NoCV1_150 in which the influence of the earth's curvature is not considered, Fig. 4d shows that the perturbation horizontal winds are unexpectedly strong and quite noisy. The signatures of cyclonic rotation within each of the cells are not so well analyzed. Although the strength of the major updraft in Fig. 5d is well captured, just as in Fig. 5b of CNTL1_150, the updraft in Fig. 5d is incorrectly positioned in the vertical direction, about 1 km below than that in Fig. 5a. All these distorted features are evidently caused by the neglect of the effect of the earth curvature in the radar forward observation operator. It should be noted that the wind analysis generally becomes worse even in CNTL1_150 km experiment because of the poorer resolution in the data at that distance.

In summary, the impacts of both the beam broadening and earth curvature are dependent on the surface range between the center of the storm and the radar location. It appears that within a range of 60 km, both the impacts of beam broadening and earth curvature can be neglected. As the distance increases beyond 60 km, more and more additional errors are introduced into the wind analysis from both earth curvature and beam broadening effects. Specifically, the neglect of the earth curvature exhibits much more negative impact than the neglect of the beam broadening. When the distance to the storm exceeds 150 km, overlooking the earth curvature and the beam broadening both bring much more obvious negative impact on the 3-dimensional wind analysis. So the simplified radial velocity forward operators and a simple tri-linear interpolation used in Gao et al. (1999) and many other studies are not recommended when the distance to the storm is greater than 60 km. To reduce computational cost, the effect of earth curvature (Eq. 2) should be considered, while the effect of beam broadening (Eq. 3) can be ignored.

(b) Realtime 3DVAR Analysis

Based on the above 3DVAR framework, Gao et al. (2013) developed a real-time, weather-adaptive analysis system for the NOAA Warn-on-Forecast (WoF) project to incorporate all available radar observations within a moveable analysis domain. This radar-based analysis system is shown to have the potential to provide improved information for making severe weather warning decisions. Some key features of the system include: (1) incorporating radar observations from multiple WSR-88Ds with NCEP forecast products as a background state, (2) the ability to automatically detect and analyze severe local hazardous weather events at 1 km horizontal resolution every 5 min in real time based on the current weather situation, and (3) the identification of strong mid-level circulations embedded in thunderstorms.

To assess the potential usefulness of the weather-adaptive realtime analysis system to warning operations, it was informally or formally tested and evaluated by forecasters who participated in NOAA Hazardous Weather Tested (HWT) spring

experiments from 2010 to 2013. During this time period, many severe weather events were successfully and automatically identified and analyzed by the 3DVAR system. For all these cases, the storm automatic positioning system performed very well. In general, strong circulations and vertical velocities associated with severe weather events were all successfully analyzed and identified. These analyses not only collocated quite well with synthesized reflectivity fields from multiple radars, but also agreed well with archived Storm Prediction Center (SPC) storm reports, which provides data about severe weather events including tornadoes, hail and strong wind events. The performance of the system, including the automatic storm positioning capability, is demonstrated in a following case as an example. More case studies and evaluations can be found in Gao et al. (2013), Smith et al. (2014), and Calhoun et al. (2014).

The case is a tornadic supercell that occurred on 20–21 April 2010 over the Texas Panhandle. A single supercell was observed over a 4 h period from 2200 UTC 20 April to 0200 UTC 21 April. At least two tornadoes, large hail and strong winds were reported during the lifetime of the storm (Fig. 6a). Radial velocity observations from several nearby WSR-88Ds, including KAMA, KLBB, and KFDX, were incorporated into the 3DVAR analyses (Fig. 6b).

The supercell storm initiated near the New Mexico-Texas state line. During the first two hours, the circulation in the storm was not very strong, but the mesocyclone within the supercell intensified just prior to 0000 UTC 21 April. Near 0004 UTC the first weak tornado was reported, after which the storm continued to develop and grew in intensity. The automated floating 3DVAR domain followed the evolution of this storm very closely. The analyzed horizontal winds, vertical vorticity and the interpolated radar reflectivity at 3 km AGL suggest that the storm was at peak intensity from 0030 UTC to 0100 UTC 21 April (Fig. 7). During this half hour

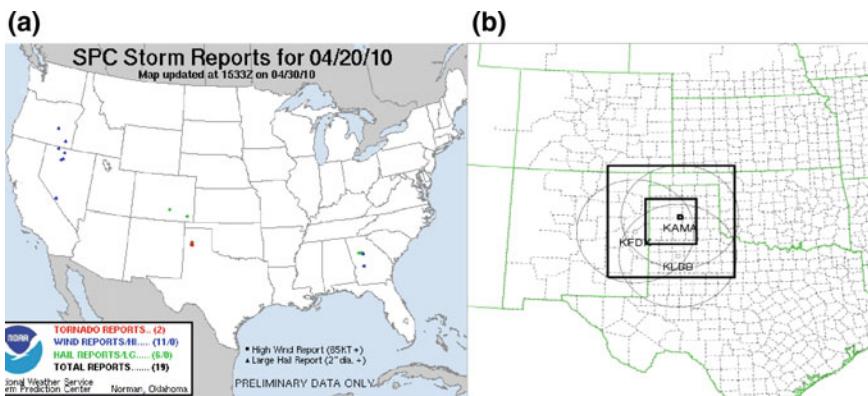


Fig. 6 Texas Panhandle tornadic storm event on April 20, 2010. **a** The storm report from Storm Prediction Center (SPC); and **b** the illustration of 3DVAR analysis domain. The inner domain of 200×200 km is used for 3DVAR analysis. The outer domain of 500×500 km is used to identify 88D radars to be used. WSD-88D radars used are KAMA, KFDX, and KLBB (Courtesy of the American Meteorology Society)

period, a strong closed circulation was evident at 3 km AGL with maximum vertical vorticity above 0.01 s^{-1} at almost all times. The strong mesocyclone (indicated by the vertical vorticity) was approximately collocated with the reflectivity core, and extended through about 10 km in height at both 0045 and 0050 UTC (Fig. 8). A weak echo region (WER) was also evident near the center of the storm below 4 km AGL (Fig. 8a, c) and low-level storm inflow was clearly shown. A second tornado touched down near 0047 UTC (SPC storm reports, Fig. 6a). A gradual occlusion of the hook echo and wind fields occurred from 0035 and 0100 UTC (Fig. 9a–f), and it appears that the second tornado touched down in the middle of this occlusion process (from 0045 to 0055 UTC). During this period, this storm also produced large hail according to SPC storm reports (Fig. 6a).

(c) A Real Data Case using Model Equations as Weak Constraint

The rapid increase in computer power and the development of storm-scale nonhydrostatic models, such as the Weather Research and Forecasting (WRF) model, and the Advanced Regional Prediction System (ARPS, Xue et al. 2000), have made real-time, explicit forecasts of thunderstorms both possible and more and more common in recent years (Kong et al. 2009; Xue et al. 2010; Clark et al. 2011). The operational WSR-88D network of the United States is a valuable and unique source of data for storm-scale numerical weather prediction (NWP), because of its capability to provide observations at high enough spatial and temporal resolutions to resolve convective storms. This situation leads to radar data assimilation playing an important role in providing an accurate storm-scale initial condition for NWP. However, obtaining an accurate storm-scale initial condition remains a significant challenge since these radars only observe radial velocity and reflectivity, neither of which is a variable in the NWP model. Another big challenge is that there is no simple balance constraint which can couple different model variables in a dynamic consistent manner. To overcome this difficulty, Ge et al. (2012) described the development of DPEC (shown in Eqs. 14–22) and testing of it with idealized experiments. DPEC was also applied to a real supercell case but only radial velocity was assimilated there. In the following, DPEC is further applied to one real tornadic supercell thunderstorm case—The 5 May 2007 Greensburg/Kansas (KS) case. In addition to assimilating radial velocity data in the 3DVAR, radar reflectivity data are used in a could analysis (Hu et al. 2006a) for this case. The impact of DPEC on radar data assimilation is examined mainly based on the storm forecasts (Ge et al. 2013).

The 5 May 2007 Greensburg, KS, tornadic thunderstorm complex produced 18 tornadoes in the Dodge City forecast area and 47 tornado reports in Kansas, Nebraska and Missouri. One of them is the strongest tornadoes in recent years. This tornado started moving through Greensburg at 0245 UTC 5 May 2007 (2145 CDT 4 May) and destroyed over 90 % of the town. The tornado damage was rated at EF5—the highest rating on the Enhanced Fujita scale (McCarthy et al. 2007). A detailed description of the supercell that spawned this tornado and its environment setting can be found in Bluestein (2009) and Stensrud and Gao (2010).

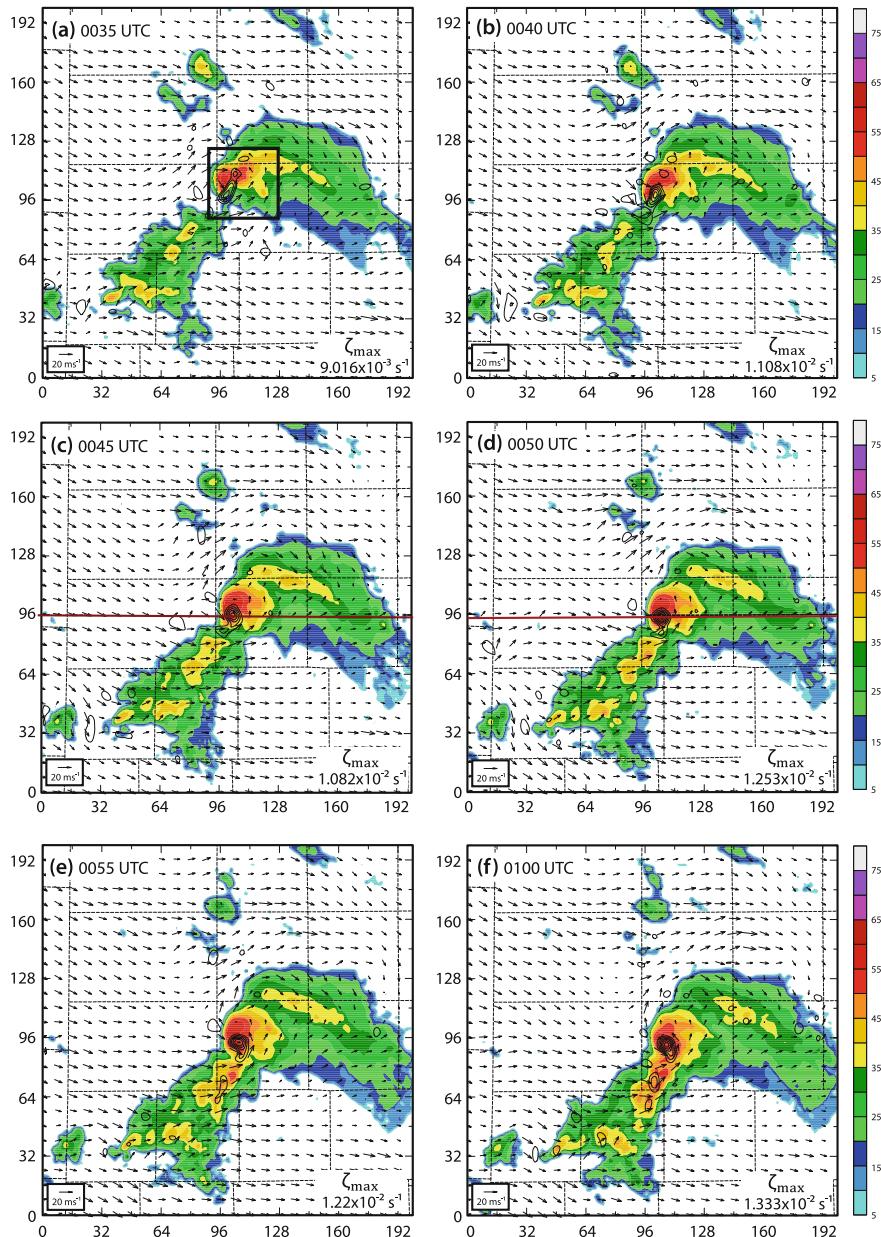


Fig. 7 The synthesized reflectivity (color shaded, dBZ), horizontal wind fields (vectors, ms^{-1}), and vertical vorticity (contour interval, $2 \times 10^{-3} \text{ s}^{-1}$) at 3 km AGL using data from 88D radars shown in Fig. 6b, at **a** 0035 UTC, **b** 0040 UTC, **c** 0045 UTC, **d** 0050 UTC, **e** 0055 UTC, and **f** 0100 UTC, 20-21 Apr 2010 near Umbarger, Texas. Maroon line denotes location of cross-sections in Fig. 8. Black box in **(a)** is the zoom-in area for Fig. 9 (Courtesy of the American Meteorology Society)

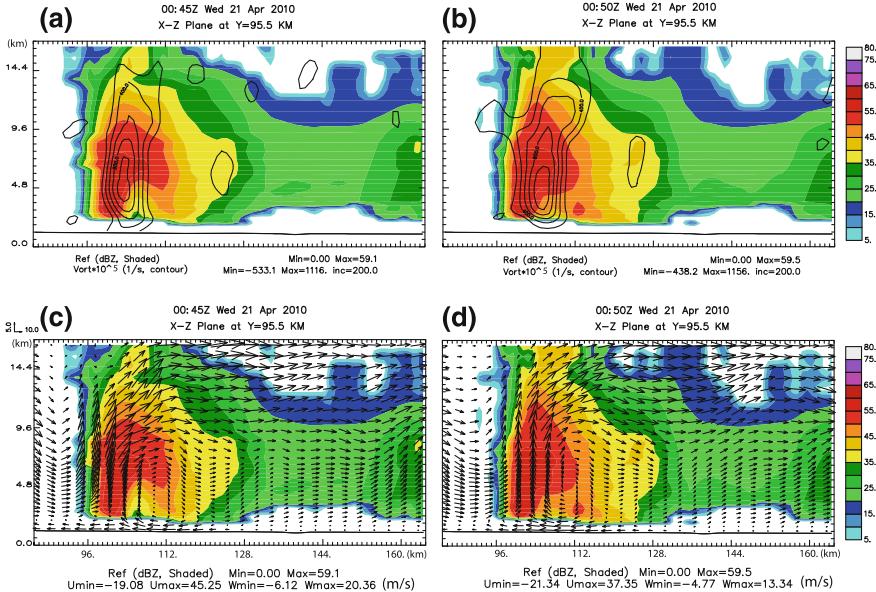


Fig. 8 Vertical vorticity (black contours in **a** and **b**), wind vectors (in **c** and **d**), and reflectivity (color shaded) along a vertical slice indicated in Fig. 7, at 0045 UTC (left panel), and 0050 UTC (right panel), 20 April, 2010 near Umbarger, Texas. Between these two time levels, a tornado touched down (Courtesy of the American Meteorology Society)

For this real data case, we use 3 km grid spacing with 200×200 grid points in the horizontal. The ARPS model domain is shown in Fig. 10. The domain is selected with sufficient coverage to contain the principal features of interest while maintaining some distance between the primary storms and the lateral boundaries. The model uses 47 terrain-following vertical layers, with nonlinear vertical stretching, via a hyperbolic tangent function that yields a spacing of 100 m at the ground and expands to approximately 800 m at the top of the domain. The Lin (1983) three-ice microphysical scheme is used together with a 1.5-order turbulent kinetic energy subgrid parameterization. A wave radiation condition is applied at the top boundary and rigid-wall conditions are applied to the bottom boundary.

The impact of the DPEC will be discussed in terms of the quality of ensuing forecasts instead of the analysis because no truth or high-resolution observation is available for verification of the analysis. Four experiments are conducted for this case (Table 2). The first experiment does not include DPEC in J and will be referred as experiment NoDP1. The second experiment uses DPEC with the DP weighting coefficient of $1.0E8$ and is referred as experiment DP1. The third and fourth experiments are similar to DP1 except the DP weighting coefficients are multiplied and divided by 5 respectively. They are referred as experiments DP1m5 and DP1d5 respectively. In all the above four experiments the mass continuity equation constraint is used with the MC weighting coefficient of $1.0E8$.

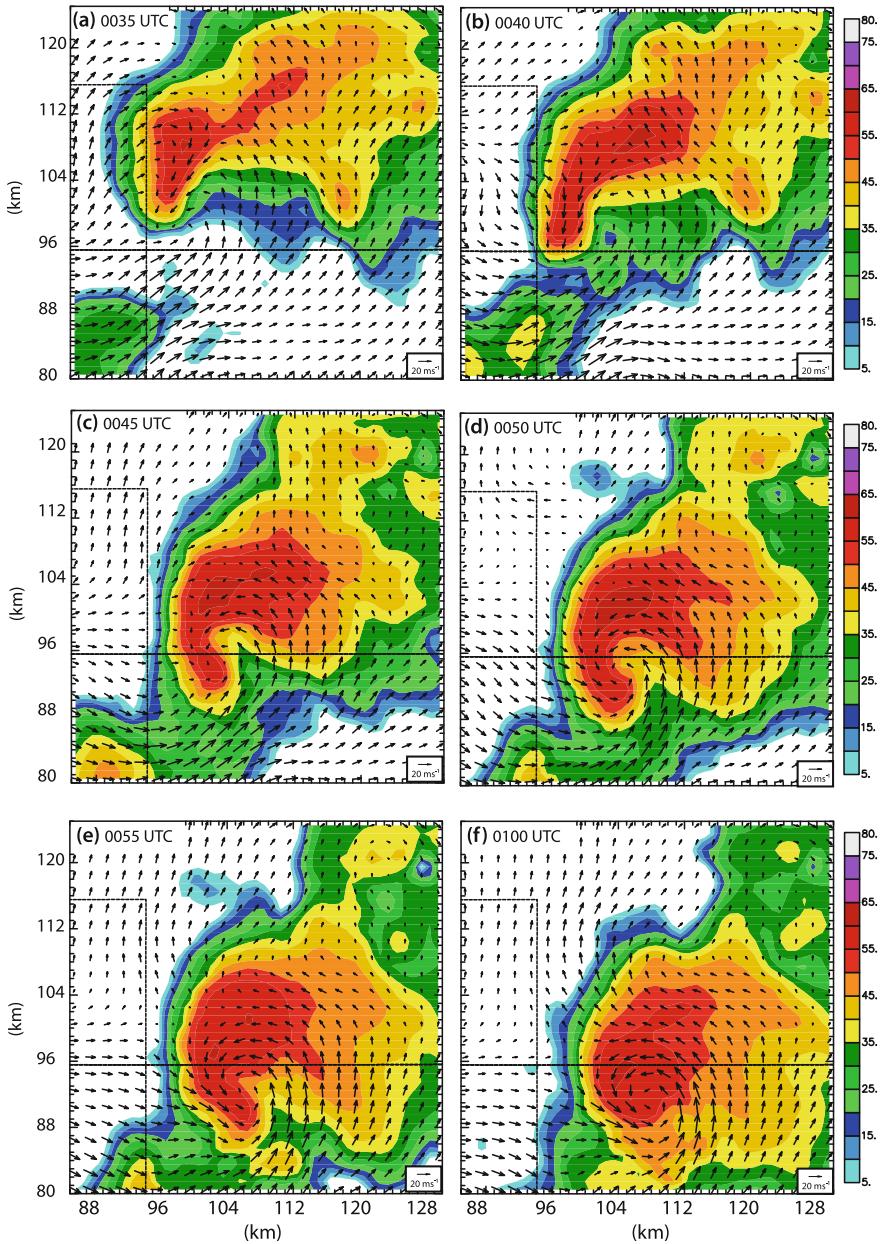


Fig. 9 The synthesized reflectivity (color shaded, dBZ), horizontal wind fields (vectors, ms^{-1}), at 1.5 km AGL roomed in from the box in Fig. 7a, at **a** 0035 UTC, **b** 0040 UTC, **c** 0045 UTC, **d** 0050 UTC, **e** 0055 UTC, and **f** 0100 UTC, 20–21 Apr 2010 near Umbarger, Texas (Courtesy of the American Meteorology Society)

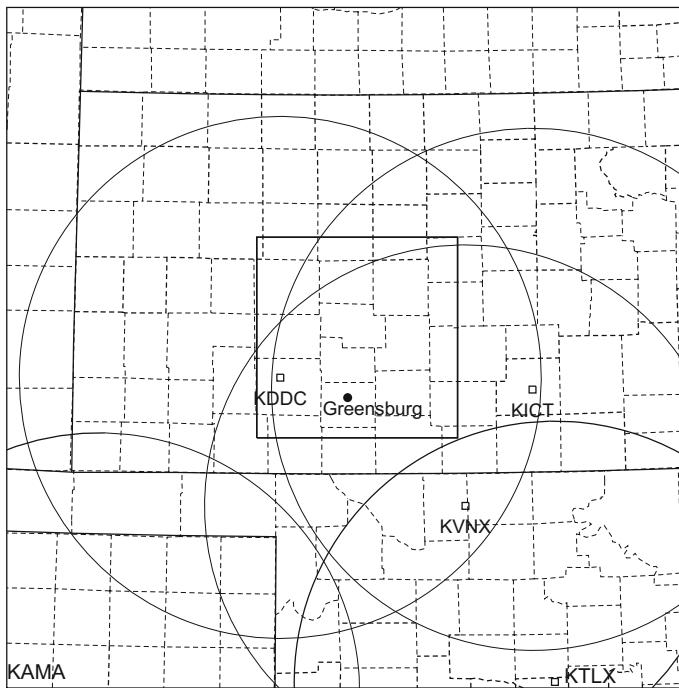


Fig. 10 The model domain with county boundaries for the 5 May 2007 Greensburg tornadic supercell thunderstorm case study. The five radars as well as their 230 km range circles are also shown. The black dot shows the location of the town of Greensburg. The black bold inner box illustrates the domain coverage in Fig. 11 (Courtesy of the Hindawi Publishing Corporation)

Table 2 List of data assimilation experiments (DP stands for “Diagnostic Pressure equation”)

Case name	Experiment name	DP weighting coefficient
5 May 2007 Greensburg case	NODP1	0
	DP1	1E8
	DP1d5	2E7
	DP1m5	5E8

For all the above four experiments, data from five radars at Dodge City, Kansas (KDDC), Vance Air Force Base, Oklahoma (KVNX), Wichita, Kansas (KICT), Oklahoma City, Oklahoma (KTLX), Amarillo, Texas (KAMA) are used (Fig. 10). A quality control procedure is applied before the use of the radar data, which includes clutter removing, velocity dealiasing using SOLOII software from the National Center for Atmospheric Research (NCAR). The initial analysis background and the boundary conditions come from the mean of a mesoscale ensemble assimilation system run at 30 km grid spacing (Stensrud and Gao 2010). While Stensrud and Gao (2010) performed a 3DVAR analysis only at one time level

before the launch of the forecast, the present study performs cycled 3DVAR analyses with a 1-h-long assimilation period before the forecast. A 5 min ARPS forecast follows each analysis, and this process is repeated until the end of the 1-h assimilation period. From the final analysis, a 1-h forecast is launched. So each experiment consisted of a 1-h assimilation period (from 0130-0230 UTC) and a 1-h forecast period (0230-0330 UTC).

We focus on the discussion of the major supercell thunderstorm at the southernmost side that produces the EF-5 tornado hitting the Greensburg area between 0245 UTC \sim 0305 UTC. It bears a hook echo signature at 0230 UTC. As the major storm reaches Greensburg, the hook echo signature becomes less prominent due to reflectivity wrap up. During this period, the radar velocity observations indicate strong cyclonic rotation associated with the strong tornado. The major storm moves gradually towards northeast. After passing the town of Greensburg, a second, EF-3, tornado develops at the end of Greensburg tornado just northeast of the town

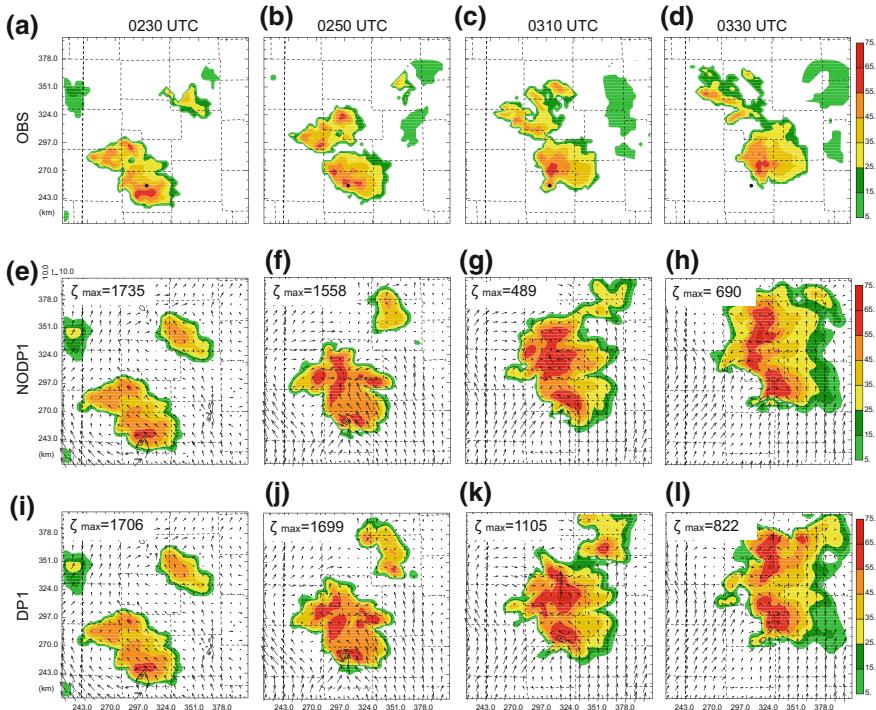


Fig. 11 Observed radar reflectivity mosaic (dBZ) at 2 km MSL from KDDC, KICT, KVN, KAMA, KTLX Doppler radars valid at **a** 0230, **b** 0250, **c** 0310, and **d** 0330 UTC; simulated radar reflectivity (dBZ), horizontal winds, and vertical vorticity (contours starting at 0.005 s^{-1} with an interval of 0.005 s^{-1}) at 2 km MSL from NoDP1 valid at **e** 0230, **f** 0250, **g** 0310, and **h** 0330 UTC; and from DP1 valid at **i** 0230, **j** 0250, **k** 0310, and **l** 0330 UTC. The duration 0230-0330 UTC covers the 1-h forecast period. The black dots in **(a-d)** indicate the location of the town of Greensburg. The maximum vertical vorticity is shown for NoDP1 and DP1 experiments with the unit of 10^{-5} s^{-1} (Courtesy of the Hindawi Publishing Corporation)

(McCarthy et al. 2007). A radar reflectivity mosaic can be created from the aforementioned five WSR-88D radars by interpolating reflectivity data from all radars into model grid points and keeping the largest reflectivity value for each grid point. The reflectivity mosaic is then used for forecast verification. The evolution of the storm as indicated by the radar reflectivity mosaic at 2 km MSL is shown in Fig. 11 from 0230 to 0330 UTC every 20 min. Note that the hook echo is not evident in these figures due to the use of 3 km resolution and a smoothing procedure applied in the mosaic generating process.

To demonstrate the impact of DPEC, We will now investigate these data assimilation experiments ingesting both the radial velocity data and reflectivity data. Figure 11e-l show the reflectivity, horizontal wind vector and vertical vorticity at $z = 2$ km MSL from 0230 UTC to 0330 UTC every 20 min for the NoDP1 and DP1 experiments. It is shown that after 1 h. data assimilation (Fig. 11e, i), the storm is already spun up in terms of the reflectivity pattern. The reflectivity pattern, strength and location agree well with the observation (Fig. 11a). A rotating circulation and a strong vertical vorticity column are collocated at the observed hook-echo region. The major storm then moves gradually towards northeast, which also agrees with the observation. Since 0300UTC, the predicted major storm moves faster than what observed. In spite of this, both NODP1 and DP1 still make reasonable forecasts in terms of the general evolution of the major storm. DP1d5 and DP1m5 produce very similar forecasts as DP1 and are therefore not shown in Fig. 11.

In Fig. 11, in terms of reflectivity pattern, there is no significant difference in the general evolution of the major storm between the NODP1 and DP1 experiments. The computed forecast scores (not presented here) also show little difference, seconding the above finding. However, there is evident difference in the predicted low-level mesocyclone rotation as partly indicated by larger maximum vertical vorticity in Fig. 11j-l than that in Fig. 11f-h. As a further demonstration, Fig. 12

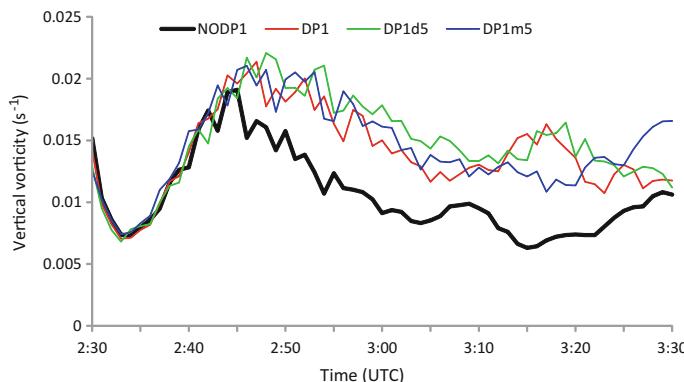


Fig. 12 The time series of maximum vertical vorticities below two kilometers from 0230 UTC to 0330 UTC 5 May 2007 every 1 min. The horizontal axis shows the time in UTC, the vertical axis shows the vertical vorticity value in unit of s^{-1} . The black line is for experiment NoDP1, the red line for experiment DP1, the blue line for DP1m5 and the green line for DP1d5 (Courtesy of the Hindawi Publishing Corporation)

shows the time series of the maximum vertical vorticity below two kilometers every 1 min from 0230 UTC to 0330 UTC for all four experiments. It is illustrated in Fig. 12 that since 0245 UTC and until the end of the forecast, the low level maximum vertical vorticity from the experiments applying DPEC (the red, blue and green lines) is much higher than that from the NODP1 experiment (the black line). Our detailed examinations show that larger low-level vertical vorticity corresponds to a better-defined mesocyclone vortex, which is stronger and deeper. This kind of behavior is very similar to findings in Ge et al. (2012). As an example, Fig. 13 presents the vertical vorticity at the vertical cross section through the center of the major storm at $y = 259.5$ km at 0250 UTC 5 May 2007. It is noticeable that the experiments using DPEC (Fig. 13b-d) predict stronger and deeper rotation columns than the “NODP1” experiment (Fig. 13a). Therefore, it can be concluded that for the experiments here, although the use of DPEC does not evidently improve the

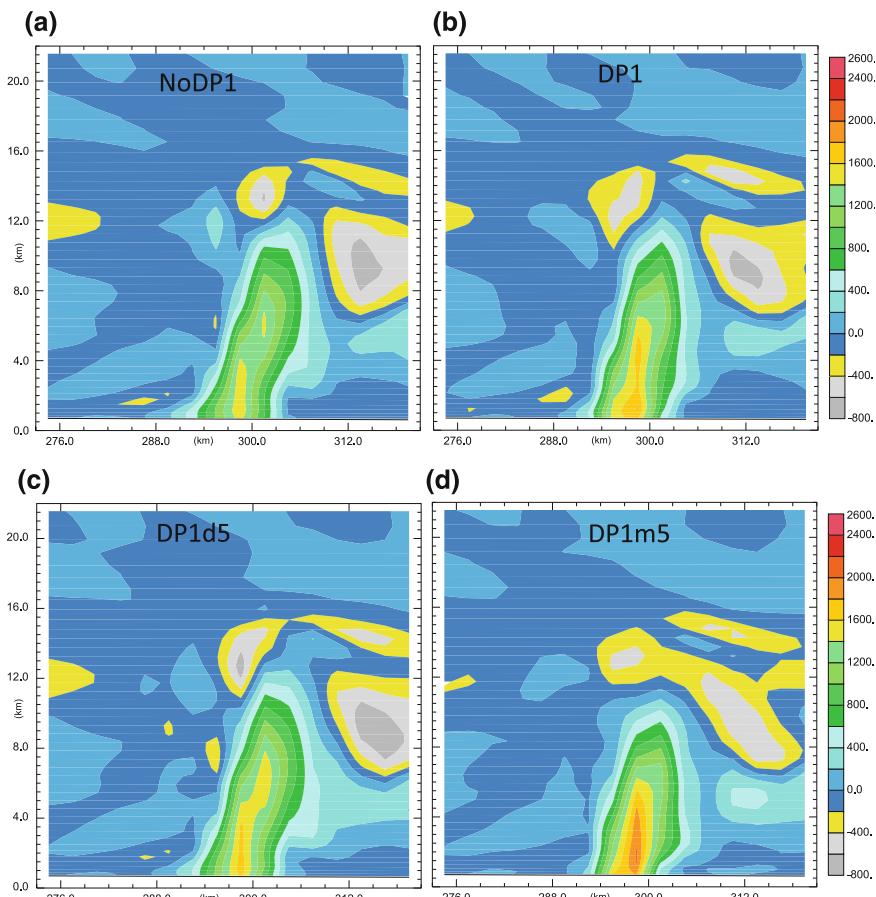


Fig. 13 The vertical vorticity (in unit of 10^{-5} s^{-1}) at the vertical cross section through the center of the major storm at $y = 253.5$ km at 0250 UTC 5 May 2007 for the **a** NODP1, **b** DP1, **c** DP1d5 and **d** DP1m5 (Courtesy of the Hindawi Publishing Corporation)

forecast of the general evolution of the major storm in terms of reflectivity pattern, it does help improve the forecast of the mesocyclone rotation associated with the observed Greensburg tornado. It is found that the experiments using DPEC generally predict higher low-level vertical vorticity than the experiments not using DPEC near the time of observed tornados. Therefore, it is concluded that the use of DPEC improves the forecast of supercell mesocyclone rotation of the major thunderstorm for this case. The experiments using different weighting coefficients generate similar results. This suggests that DPEC is not very sensitive to the weighting coefficients, similar to the study with idealized case (Ge et al. 2012). More research is needed to get more general conclusions with this DPEC constraint.

(d) Assimilating both V_r and Z in Variational Framework

In the last sub-section, a complex cloud analysis package is used to assimilate radar reflectivity data into a storm-scale NWP model. However, the uncertainties associated with the empirical parameters used in the cloud analysis make this method less than optimal. The assimilation of radar reflectivity into a storm-scale NWP model using a variational framework is a potentially more beneficial approach, yet this has not been thoroughly studied for storm-scale data assimilation. Gao and Stensrud (2012) was the first study to include ice related hydrometeors as analysis variables in variational storm-scale data assimilation, and also the first to partition hydrometeor variables using a background temperature field from a storm-scale NWP model (ARPS model in this case). We briefly illustrate this study with the following idealized case, more detail can be found in Gao and Stensrud (2012).

The effectiveness of the assimilation of radar reflectivity data combined with radial velocity in the 3DVAR is evaluated by utilizing synthetic Doppler radar data derived from a truth simulation of a supercell thunderstorm. The ARPS is used to generate the truth simulation and is used in the 3DVAR experiments. Similar to Sect. 3(a), the environment from a well-documented tornadic supercell storm that occurred near Del City, Oklahoma, on 20 May 1977 is used for the experiments (Ray et al. 1981). Different from Sect. 3(a), parameter settings for the ARPS model include $57 \times 57 \times 35$ total grid points with grid spacing $dx = dy = 1$ km in the horizontal and $dz = 500$ m in the vertical.

During the truth simulation, the initial convective cell develops and strengthens over the first 30 min. The strength of the cell then decreases over the next 30 min or so, in association with the cell splitting at around 55 min. The right moving (relative to the storm motion vector which is towards north-northeast) cell tends to dominate the system (Fig. 14a–e). The evolution of the simulated storm is qualitatively similar to that described by Klemp and Wilhelmson (1978).

The three-dimensional wind and hydrometeor fields from this truth run are sampled by two ground-based pseudo-radars, located at the southwest and southeast corners of the computational domain, to obtain synthetic radial velocity and reflectivity observations using forward model Eqs. (2) through (11). Random errors drawn from a normal distribution with zero mean and a standard deviation of

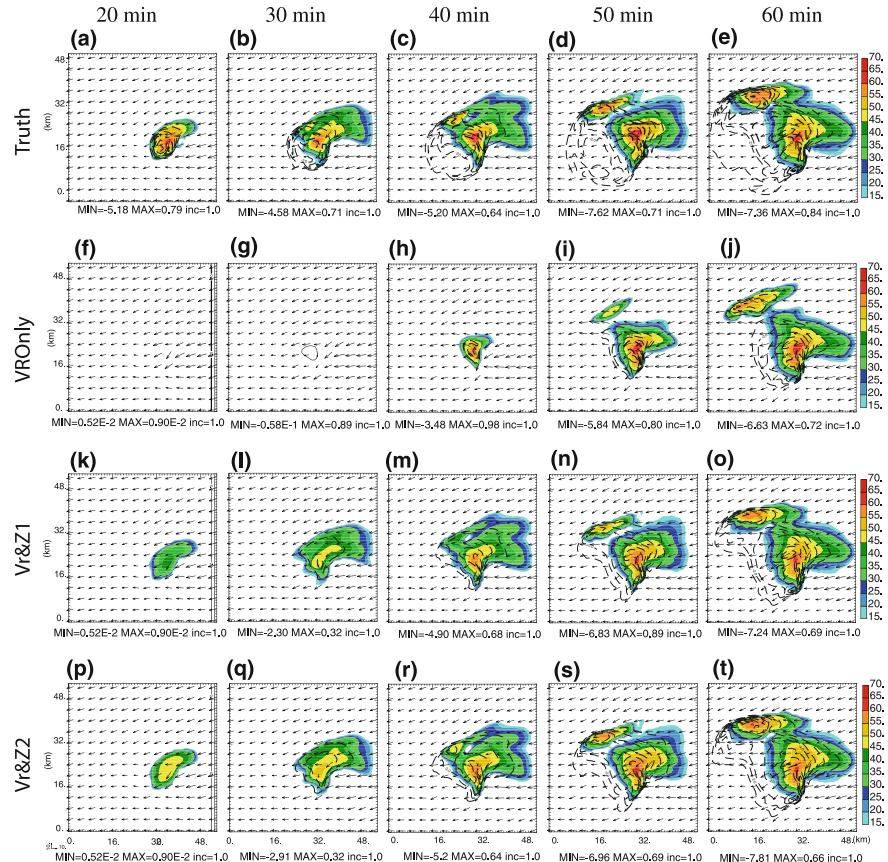


Fig. 14 Horizontal winds (vectors; m s^{-1}), perturbation potential temperature (contours at 1 K intervals) and simulated reflectivity (shaded contours, in dBZ) at 250 m AGL, for the truth simulation (a–e); the cycled 3DVAR analyses with assimilating radial velocity only (f–g), the cycled 3DVAR analysis with assimilating radial velocity and reflectivity without hydrometeor classification (k–o), and the cycled 3DVAR analysis with assimilating radial velocity and reflectivity with hydrometeor classification (p–t). The times shown are 20, 30, 40, 50 and 60 min of model integration time. The assimilation interval is 5 min (Courtesy of the American Meteorology Society)

1 m s^{-1} are added to the synthetic radial velocity data, and random errors with zero mean and a standard deviation of 3 dBZ are added to the synthetic reflectivity data. Since v_r in (3) is sampled directly from the model velocity fields, hydrometeor sedimentation is not involved. As in Snyder and Zhang (2003), the simulated observations are assumed to be available on the grid points and only available where the truth reflectivity is greater than zero in the analysis domain. The elapsed times for the volume scans of the radars are neglected, and thus it is presumed that the radial wind and reflectivity observations throughout the radar volume are sampled simultaneously.

The 3DVAR method is applied using a cycling approach. An initial analysis that does not include any effects of deep convection is used as the background to start the cycle (this is simply the horizontally homogeneous conditions provided by the assumed environmental sounding). Synthetic radar data are then assimilated using the 3DVAR to produce a new analysis and a short 5-min forecast is made starting from this analysis using the ARPS model. This 5-min forecast then becomes the background for the assimilation of the synthetic radar data valid at this new observation time. The 3DVAR analysis and forecast cycles are started at 20 min of the truth run integration time and the radial velocity and/or reflectivity observations are assimilated every 5 min thereafter. The final analysis is done at 60 min into the truth run integration.

Three experiments are performed to determine the performance of reflectivity data assimilation. In the first experiment, only radial velocity data are assimilated. In the second experiment, radial velocity data with forward operator Eq. (2) and reflectivity data with forward operator Eq. (4) are assimilated. In the third experiment, radial velocity data using Eq. (2) and reflectivity data using the newly developed forward operator (10) are assimilated in which the hydrometeor partitioning is based upon the background temperature field. The analysis domain is the same as the truth run domain discussed earlier.

The horizontal winds, potential temperature perturbations and reflectivity at 250 m above ground level for the truth run and all three experiments at 20, 30, 40, 50 and 60 min of model time are shown in Fig. 14. For the first experiment with radial velocity data assimilation only, it can be seen that only the velocity structure around the storm is somewhat captured after 10 min or three analysis cycles (Fig. 14f, g) while the perturbation potential temperature field is very weak. The precipitation, as indicated by reflectivity, has not yet appeared. After two more analysis-forecast cycles, or 20 min into the assimilation ($t = 40$ min), the precipitation has reached to the ground, but it is weaker and covers a much smaller area compared to the truth simulation (cf. Figures 14c, h). After six analysis cycles ($t = 50$ min), the right-moving storm cell is similar to the truth but the left-moving storm is still very weak. Also, the extent of the cold pool is too small on the southwest side of the storm even though its maximum intensity underneath the cells is close to the truth (cf. Fig. 14d, i). Only after 40 min into the assimilation ($t = 60$ min) does the overall storm structures become close to the truth, though the extent of the cold pool is still reduced (cf. Fig. 14e, j). In general, comparisons with the truth simulation indicate that the development of precipitation is significantly delayed in first experiment when only radial velocity is assimilated.

For the second experiment, when the simulated reflectivity data are also assimilated, the development of the storm is much faster. At the beginning of analysis-forecast cycles, the precipitation already is reaching the ground and a wind perturbation is seen, with the patterns being quite similar to, although weaker than, the truth (Fig. 14k). After two more analysis-forecast cycles, the storm cell compares reasonably well with the truth (Fig. 14l) and a weak cold pool also appears, which does not emerge at this time level in the first experiment (Fig. 14g). With two more analysis-forecast cycles ($t = 40$ min) completed, the overall storm structures

begin to closely resemble the truth, though the extent of the cold pool is still smaller than in the truth run (Fig. 14m). After the final analysis cycle, the low level flow and reflectivity patterns, as well as the strength and extent of the cold pool, are in very good agreement with the truth (cf. Fig. 14o, e) although some differences still exist in terms of the shape of cold pool and reflectivity. These results suggest that the assimilation of both radial velocity and reflectivity in the cycled 3DVAR system helps to reduce the spin-up time for developing hydrometeors in the analyses and forecasts.

The third experiment is performed when both radial velocity and modified reflectivity forward operator (10) is used. In Eq. (10), the hydrometeor classification is made though a background temperature field from the ARPS model which is updated as the model is integrated forward in time. Overall, the analysis is noticeably improved compared to that of first two experiments. The precipitation, in term of reflectivity, is immediately stronger from the very beginning of assimilation as compared with the previous two experiments (cf. Fig. 14p, f, k). At 20 min into the assimilation ($t = 40$ min), the cell splitting process has already begun, and the cold pool is stronger and has larger extent than in the other two experiments (cf. Fig. 14r, h, m). By 40 min into the assimilation ($t = 60$ min), the structure of the storm is almost the same as in the truth (cf. Fig. 14t, e). Overall, the analysis converges much faster and the analysis errors are smaller than the other two experiments as expected. These results highlight the value of the partitioning of the hydrometeor types using the background temperature field to the 3DVAR system.

To illustrate why the third experiment with hydrometeor classification is more reasonable, the contours for rain water, snow and hail mixing ratios after the first assimilation at 250 m above ground level are shown in Fig. 15. This time level is chosen because at first cycle of assimilation, background fields for these hydrometeor variables are zero and the correction to these variables during the assimilation is purely from reflectivity observations. For the first experiment without assimilating reflectivity, all hydrometeor variables for rain water, snow and hail are zero, so they are not shown. For the second experiment, it is shown that obtained rain water after the assimilation near the ground level is very weak (Fig. 15d) compared to the truth (Fig. 15a). Because the temperature field is usually quite warm at this level, the expected snow and hail after the assimilation should be zero (Fig. 15b) though it is not true for hails that reach ground (Fig. 15c). However, obtained snow and hail in the second experiment show some quite big values around the storm center (Fig. 15e, f). This is not physically consistent. In the third experiment with hydrometeor classification, we obtain a reasonable pattern for rain water (Fig. 15g) and almost zero values for snow and hail near ground level (Fig. 15h, i). Similarly, in the second experiment, we also found non-zero rain water mixing ratio in the upper levels of the analysis where only snow and hail are expected because of the very cold temperatures at these levels (figures not shown), while such behavior does not appear in the third experiment. So it is obvious that the assimilation of reflectivity using Eq. (10) with hydrometeor classification produces results which is much more physical consistent.

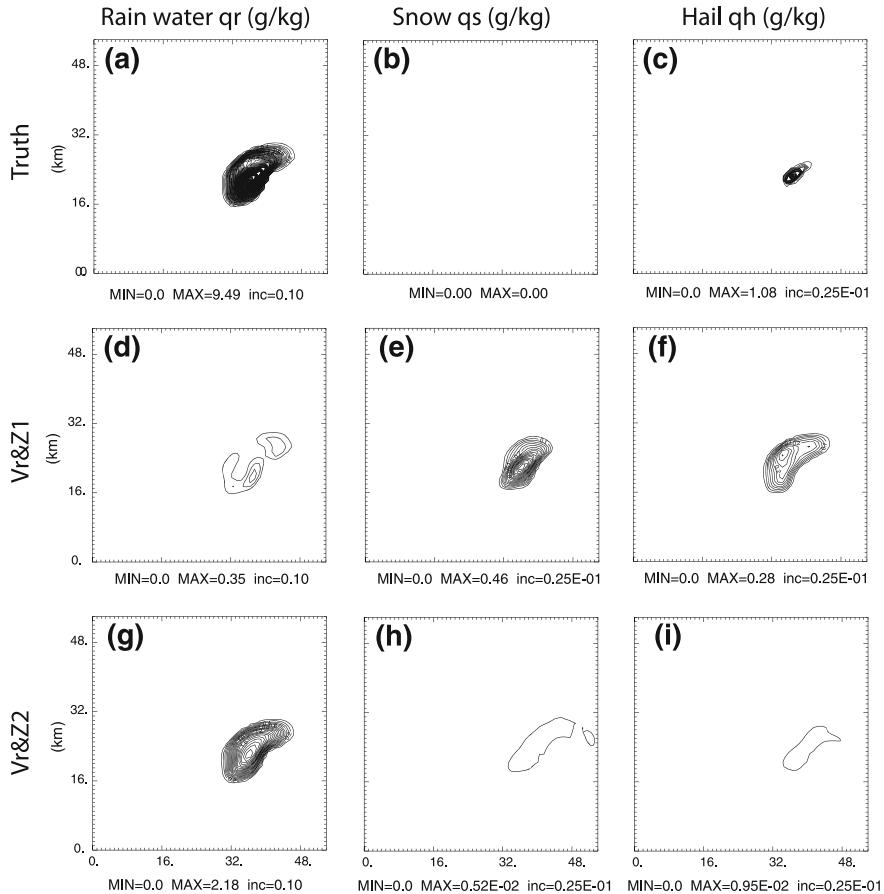


Fig. 15 Contours for rain water mixing ratio (q_r), snow mixing ratio (q_s) and hail mixing ratio (q_h) at 250 m AGL and after the first data assimilation (or 20 min of model integration time), for the truth simulation (a–c); the cycled 3DVAR analysis with assimilating radial velocity and reflectivity without hydrometeor classification (d–f), and the cycled 3DVAR analysis with assimilating radial velocity and reflectivity with hydrometeor classification (g–i) (Courtesy of the American Meteorology Society)

To judge the analyses quantitatively, the root-mean-square (rms) errors of the analyzed fields for all vertical model levels are calculated against the truth. The rms errors are averaged over those grid points where the reflectivity is greater than 10 dBZ in the truth simulation, similar to Snyder and Zhang (2003). The rms errors of vertical velocity w , perturbation potential temperature q' , perturbation pressure p' , water vapor mixing ratio q_v , rain water mixing ratio q_r , and reflectivity Z (derived from the mixing ratios of rain water, snow and hail) in all three experiments generally decrease with the data assimilation cycles starting from the first analysis (Fig. 16). The rms errors decrease much faster for most of model variables (except

w) in the second and third experiments when reflectivity is also assimilated along with radial velocity. This rms error reduction is especially visible at later assimilation cycles. The improvement for w (Fig. 16a) and also for the horizontal wind components (not shown) are not large, since the assimilation of reflectivity has little direct impact on wind field until later times when the temperature and pressure fields are also improved. The rms errors for the third experiment with hydrometer classification are smaller than those for the second experiment, but the rms error values are similar, especially from the very beginning (except for rain water mixing ratio, which has a large decrease in rms error from the beginning in Fig. 16e). However, the comparison of both reflectivity field and perturbation potential temperature against the truth in Fig. 14p, k, a indicates better analyses are produced by the third experiment than by the other two experiments. This is not surprising because the storm cells are small-scale, discrete features and any errors in their structure or position will strongly influence the rms errors. This result simply confirms that multiple measures of accuracy are needed when analyzing results from assimilation methods.

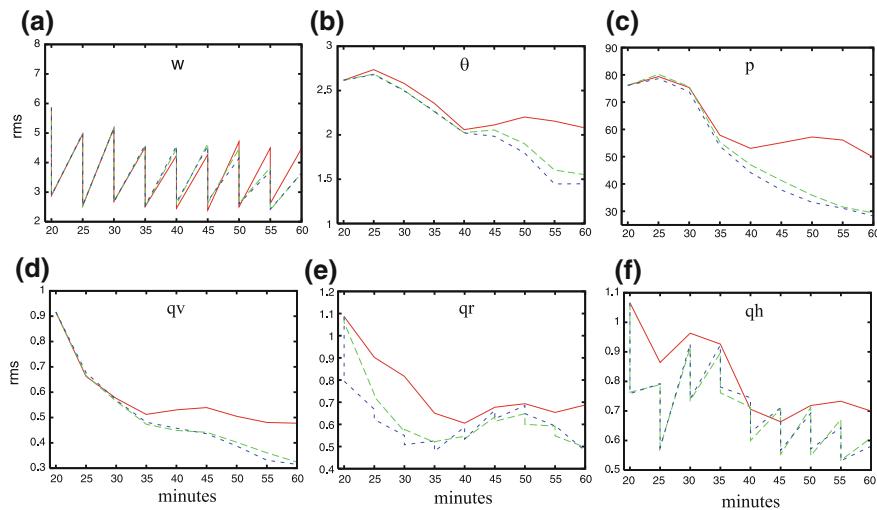


Fig. 16 The rms errors of the cycled 3DVAR analyses and forecast, averaged over points at which the reflectivity is greater than 10 dBZ. **a** for vertical velocity w , **b** for perturbation potential temperature θ' , **c** for perturbation pressure p' , **d** for water vapor mixing ratio q_v , **e** for rain water mixing ratio q_r , and **f** for reflectivity Z . The red line is for the radial velocity data only experiment 1, the green line is for the radial velocity and reflectivity without hydrometer classification experiment 2, and the blue line is for the radial velocity and reflectivity with hydrometer classification experiment 3 (Courtesy of the American Meteorology Society)

4 Summary and Future Work

An innovative three-dimensional variational data assimilation scheme for convective-scale NWP has been developed over the past several years. In this evolved approach, a cost function is defined by a background term, a radar observation term, and a weak constraints term, as in conventional 3DVAR schemes. The background error covariance matrix is modeled by a recursive filter and the square root of this matrix is used to precondition the minimization problem. The ability to assimilate radar radial velocity data was one of the first innovations in this scheme, followed by capabilities for using model equation-derived mass continuity and DPEC as weak constraints, in addition to the ability to assimilate radar reflectivity directly. The impact of some of these features is illustrated in this paper.

First, the radial-velocity innovation is examined independently in a framework designed specifically to evaluate the effects of beam broadening and earth curvature in simulations of an idealized supercell storm. This series of tests demonstrates that the method is very sensitive to the effect of the earth's curvature as the surface range increases, but relatively insensitive to beam broadening. These idealized tests informed our decision to neglect beam-broadening in real data cases while including earth-curvature effects. The 3DVAR framework that emerged from this testing has been applied as part of the NOAA Warn-on-Forecast project as a real-time, weather-adaptive analysis system that incorporates available radar observations within a moveable analysis domain. The system has performed very well during springtime testing in the NOAA HWT for several years, including successful detection and characterization of many severe weather events during simulated severe-weather-warning exercises. Current plans call for expansion of this work, including larger relocatable domains and more year-round testing and collaboration with National Weather Service Forecasters.

The impact of DPEC on radar data assimilation has been examined mainly on the basis of storm simulations. It has been found that the experiments using DPEC generally predict higher low-level vertical vorticity than the experiments not using DPEC near the simulation time to that of observed tornados. However, more case studies are needed to provide more conclusive evidence that DPEC has a positive impact on storm-scale radar data assimilation, and subsequent forecasts.

Finally, the impact of assimilating radar reflectivity data in addition to radial velocity with intermittent 3DVAR analysis-forecast cycles has been examined using an idealized thunderstorm case. A forward operator for reflectivity has been developed using the background temperature field from a NWP model to inform the classification of hydrometeor types. Three preliminary experiments were performed. The first experiment used radial velocity only, the second used both radial velocity and reflectivity without hydrometeor classification, and the third used both radial velocity and reflectivity with the newly developed forward operator, including hydrometeor classification. It was found that by assimilating only radial velocity data, the model can reconstruct the dynamical structures of supercell thunderstorm well within several assimilation cycles, but a spin-up problem delays

the development of precipitation significantly. The spin-up problem is reduced when assimilating reflectivity without hydrometeor classification. The analysis converges faster and the analysis errors are smallest when using the new reflectivity formulation with the classification of hydrometeor types. The cold pool also develops earlier and agrees better with the truth run when using the hydrometeor classification. The overall conclusion is that the assimilation of both radial velocity and reflectivity data in the 3DVAR system significantly reduces the severity of the spin-up problem and has the potential to improve short-range storm-scale analysis and forecasting systems. The partitioning of hydrometeors when assimilating reflectivity can be also used within other advanced data assimilation methods, such as ensemble Kalman filter and 4DVAR techniques (Gao and Stensrud 2012).

The reliability and accuracy of the updated 3DVAR system have been also illustrated by a number of other data assimilation and forecast experiments as well (e.g., Gao et al. 2010; Stensrud and Gao 2010; Xue et al. 2010; Schenkman et al. 2011). Examples of other successful applications of the system include real-time low-level wind analyses (Gao et al. 2010) and initializing high-resolution real-time severe weather forecasts (Brewster et al. 2010) using radar data from the Engineering Research Center for Collaborative Adaptive Sensing of the Atmosphere (CASA) IP1 network supported by National Science Foundation (McLaughlin et al. 2009). The system also has been used since 2008 to provide initial conditions for very high resolution (1–4 km grid spacing) deterministic and ensemble WRF model runs that assimilate WSR-88D radar data over a continental US domain in support of the HWT Spring Forecasting Experiment (SFE; Kain et al. 2010; Xue et al. 2010). A primary focus of the SFE is evaluating the utility of computer models of the atmosphere to improve predictions of hazardous and convective weather events in support of the NOAA Storm Prediction Center operations.

In spite of these advances, the 3DVAR method has its shortcomings. For example, it cannot use observational data from multiple times levels and its background error covariances are static and not flow-dependent. Because of these limitations, many studies have explored the use of more sophisticated assimilation methods, such as EnKF and 4DVAR for convective scale data assimilation and NWP. But none of these methods is universally superior to the others—each alternative has its own distinct advantages and disadvantages (Lorenc 2003; Kalnay 2007).

All things considered, including the accuracy of data assimilation methods and convective NWP models, timely delivery of forecast products, etc., we believe that new innovations in the 3DVAR data assimilation methodology provide a competitive option for convective-scale NWP. Furthermore, additional improvements to 3DVAR are likely to come as we develop balance constraints for the convective scale and optimize the incorporation of ensemble information into this 3DVAR system. The ongoing work continues to derive inspiration from Sasaki (2003), who advocated including a mixture of probabilistic and deterministic constraints in variational methods and Lorenc (2003) who proposed improving the accuracy of analysis by using combining variational and EnKF methods in a hybrid approach.

Many research articles about hybrid methods have been published in recent years, most focusing on synoptic-scale and mesoscale NWP (Buehner 2005; Buehner et al. 2010a, b; Wang et al. 2008a, b, 2013; Barker et al. 2012; Zhang et al. 2013). Gao and Stensrud (2014) showed that hybrid methods show promise for convective-scale NWP as well. Specifically, they demonstrated that incorporation of ensemble-estimated covariance in a variational approach (3DVAR in this case) can significantly improve the accuracy of the assimilation of simulated radar data for a supercell storm. This result holds even when just a few ensemble members are used and the estimated covariance contains severe sampling errors. If this result holds more generally, it may have significant implications for the Warn-on-Forecast (WoF) concept proposed by Stensrud et al. (2009). This concept includes a frequently updated, numerical-model-based, probabilistic convective-scale analysis and forecast system that supports warning operations within NOAA. It is essential that ensemble forecasts are utilized in the WoF concept to produce robust probabilistic forecast guidance, but while relatively small ensembles may be adequate for WoF-type forecasts, much larger ensembles are necessary to create a robust pure-ensemble DA system. Thus, given limited computer resources for the near future, it may not be possible to implement a pure ensemble approach to DA in early-generation WoF systems, but a WoF system that uses a relatively small ensemble for DA, within a hybrid approach, may be feasible with the computer resources that are expected to be available. This gives us confidence that we can implement a unified ensemble-based DA and prediction system for WoF even with the relatively limited computer resources that will be available in the near future.

Acknowledgements This work was supported by NOAA's Warn-on-Forecast project and NSF grants and NSF AGS-1341878. The assistance of Dr. Henry Neeman and the University of Oklahoma Supercomputing Center for Education & Research IT team is gratefully acknowledged.

References

- Albers SC, McGinley JA, Birkenheuer DA, Smart JR (1996) The local analysis and prediction system (LAPS): analysis of clouds, precipitation and temperature. *Weather Forecast* 11:273–287
- Barker DM, Huang W, Guo Y-R, Xiao QN (2004) A three-dimensional (3DVAR) data assimilation system for use with MM5: implementation and initial results. *Mon Weather Rev* 132:897–914
- Barker DM et al (2012) The weather research and forecasting model's community variational/ensemble data assimilation system: WRFDA. *Bull Am Meteor Soc* 93(831–843):2012
- Brewster K, Thomas K, Gao J, Brotzge J, Xue M, Wang Y (2010) A nowcasting system using full physics numerical weather prediction initialized with CASA and NEXRAD radar data. Preprints. In: 25th conference severe local storms, Denver, CO, American Meteor Society, Denver, CO, Paper 9.4
- Buehner M (2005) Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. *Q J R Meteor Soc* 131:1013–1043

- Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B (2010a) Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part I: description and single-observation experiments. *Mon Weather Rev* 138:1550–1566
- Buehner M, Houtekamer PL, Charette C, Mitchell HL, He B (2010b) Intercomparison of variational data assimilation and the ensemble Kalman filter for global deterministic NWP. Part II: one-month experiments with real observations. *Mon Weather Rev* 138:1567–1586
- Calhoun KM, Smith TM, Kingfield DM, Gao J, Stensrud DJ (2014) Forecaster use and evaluation of realtime 3DVAR analyses during Severe Thunderstorm and Tornado warning operations in the hazardous weather Testbed. *Weather Forecasting* 29:601–613
- Clark AJ, Kain JS, Stensrud DJ, Xue M, Kong F, Coniglio MC, Thomas KW, Wang Y, Brewster K, Gao J, Wang X, Weiss SJ, Bright D, Du J (2011) Probabilistic precipitation forecast skill as a function of ensemble size and spatial scale in a convection-allowing ensemble. *Mon Weather Rev* 139:1410–1418
- Courtier P, Thépaut J-N, Hollingsworth A (1994) A strategy for operational implementation of 4D-Var, using an incremental approach. *Q J R Meteorol Soc* 120:1367–1388
- Courtier P (1997) Dual formulation of four dimensional variational assimilation. *Q J R Meteorol Soc* 123:2449–2461
- Daley R (1991) Atmospheric data analysis. Cambridge University Press, Cambridge, 457pp
- Doviak RJ, Zrnic DS (1993) Doppler radar and weather observations, 2nd edn. Academic Press, 562pp
- Doviak RJ, Ray PS, Strauch RG, Miller LJ (1976) Error estimation in wind fields derived from dual-Doppler radar measurement. *J Appl Meteorol* 15:868–878
- Dowell DC, Wicker LJ, Snyder C (2011) Ensemble Kalman filter assimilation of radar observations of the 8 May 2003 Oklahoma city supercell: influences of reflectivity observations on storm-scale analyses. *Mon Weather Rev* 132:1982–2005
- Elmore K (2011) The NSSL hydrometeor classification algorithm in winter surface precipitation: evaluation and future development. *Weather Forecasting* 26:756–765
- Gao J, Xue M, Shapiro A, Droegelemeier KK (1999) A variational method for the analysis of three-dimensional wind fields from two Doppler radars. *Mon Weather Rev* 127:2128–2142
- Gao J, Xue M, Brewster K, Carr F, Droegelemeier KK (2002) New development of a 3DVAR system for a nonhydrostatic NWP model. Preprint. In: 15th conference on numerical weather prediction and 19th conference on weather analysis and forecasting, San Antonio, TX, American Meteor Society, pp 339–341
- Gao J, Xue M, Brewster K, Droegelemeier KK (2004) A three-dimensional variational data assimilation method with recursive filter for single-Doppler radar. *J Atmos Oceanic Technol* 21:457–469
- Gao J, Stensrud DJ, Xue M (2009a) Three-dimensional analyses of several thunderstorms observed during VORTEX2 field operations. In: 34th conference on radar meteorology, Williamsburg, VA, Online publication
- Gao J, Ge G, Stensrud DJ, Xue M (2009b) The relative importance of assimilating radial velocity and reflectivity data to storm-scale analysis and forecast. In: 23rd conference on weather analysis and forecasting/19th conference on numerical weather prediction, Omaha, NB, American Meteor Society, Paper 16A.3
- Gao J, Brewster K, Xue M, Brotzge J, Thomas K, Wang Y (2010) Real-time, low-level wind analysis including CASA and WSR-88D radar data using the ARPS 3DVAR. In: 25th conference severe local storms, American Meteor Society, Paper 7B.4 (online publication)
- Gao J, Stensrud D (2012) Assimilation of reflectivity data in a convective-scale, cycled 3DVAR framework with hydrometeor classification. *J Atmos Sci* 69:1054–1065
- Gao J, Smith TM, Stensrud DJ, Fu C, Calhoun K, Manross KL, Brogden J, Lakshmanan V, Wang Y, Thomas KW, Brewster K, Xue M (2013) A realtime weather-adaptive 3DVAR analysis system for severe weather detections and warnings with automatic storm positioning capability. *Weather Forecasting* 28:727–745
- Gao J, Stensrud DJ (2014) Some observing system simulation experiments with a hybrid 3DEnVAR system for stormscale radar data assimilation, *Mon Weather Rev* 142:3326–3346

- Ge G, Gao J (2007) Latest development of 3DVAR system for ARPS and its application to a tornadic supercell storm. In: 22nd conference on weather analysis and forecasting/18th conference on numerical weather prediction, on-line publication, 2B.6
- Ge G, Gao J, Brewster KA, Xue M (2010) Effects of beam broadening and earth curvature in radar data assimilation. *J Atmos Oceanic Technol* 27:617–636
- Ge G, Gao J, Xue M (2012) Diagnostic pressure equation as a weak constraint in a storm-scale three dimensional variational radar data assimilation system. *J Atmos Ocean Tech* 29:1075–1092
- Ge G, Gao J, Xue M (2013) Impact of a diagnostic pressure equation constraint on tornadic supercell thunderstorms forecasts initialized using 3DVAR radar data assimilation. *Adv Meteor* 2013:1–12. doi:[10.1155/2013/947874](https://doi.org/10.1155/2013/947874)
- Gilmore MS, Straka JM, Rasmussen EN (2004) Precipitation and evolution sensitivity in simulated deep convective storms: comparisons between liquid-only and simple ice and liquid phase microphysics. *Mon Weather Rev* 132:1897–1916
- Hu M, Xue M, Brewster Keith (2006a) 3DVAR and cloud analysis with WSR-88D Level-II data for the prediction of the Fort Worth tornadic thunderstorms. Part I: cloud analysis and its impact. *Mon Weather Rev* 134:675–698
- Hu M, Xue M, Gao J, Brewster K (2006b) 3DVAR and cloud analysis with WSR-88D Level-II data for the prediction of the Fort worth tornadic thunderstorms. Part II: impact of radial velocity analysis via 3DVAR. *Mon Weather Rev* 134:699–721
- Kain JS, Xue M, Coniglio MC, Weiss SJ, Kong F, Jensen TL, Brown BG, Gao J, Brewster K, Thomas KW, Wang Y, Schwartz CS, Levit JJ (2010) Assessing advances in the assimilation of radar data within a collaborative forecasting-research environment. *Weather Forecasting* 25:1510–1521
- Kalnay E, Li H, Miyoshi T, Yang S-C, Ballabrera-Poy J (2007) 4-D-Var or ensemble Kalman filter? *Tellus* 59A:758–773
- Klemp JB, Wilhelmson RB (1978) Simulations of right- and left-moving storms produced through storm splitting. *J Atmos Sci* 35:1097–1110
- Klemp JB, Wilhelmson RB, Ray PS (1981) Observed and numerically simulated structure of a mature supercell thunderstorm. *J Atmos Sci* 38:1558–1580
- Klemp JB, Rotunno R (1983) A study of the tornadic region within a supercell thunderstorm. *J Atmos Sci* 40:359–377
- Kong, Xue FM, Thomas KW, Gao J, Wang Y, Brewster K, Drogemeier KK, Kain J, Weiss S, Bright D, Coniglio M, Du J (2009) A realtime storm-scale ensemble forecast system: 2009 spring experiment. In: 23rd conference on weather analysis and forecasting/19th conference on numerical weather prediction, Omaha, NB, American Meteor Society, Paper 16A.3
- Lin Y-L, Farley RD, Orville HD (1983) Bulk parameterization of the snow field in a cloud model. *J Clim Appl Meteor* 22:1065–1092
- LeDimet F, Talagrand O (1986) Variational algorithms for analysis and assimilation of meteorological observations: theoretical aspects. *Tellus* 38A:97–110
- Lewis J, Derber J (1985) The use of adjoint equations to solve a variational adjustment problem with advective constraints. *Tellus* 37A:307–322
- Lewis J, Lakshminarayanan S, Dhall S (2006) Dynamic data assimilation: a least squares approach. Cambridge University Press, 654pp
- Lorenc AC (1992) Iterative analysis using covariance functions and filters. *Q J R Meteor Soc* 118:569–591
- Lorenc A (2003) The potential of the ensemble Kalman filter for NWP—a comparison with 4DVar. *Q J R Meteor Soc* 129:3183–3204
- McCarthy D, Ruthi L, Hutton J (2007) The Greensburg, KS tornado. In: 22th conference on weather analysis and forecasting and 18th conference on numerical weather prediction, Park City, UT, American Meteor Society
- McLaughlin D et al (2009) Short-wavelength technology and the potential for distributed networks of small radar systems. *Bull Am Meteor Soc* 90:1797–1817

- Miller LJ, Sun J (2003) Initialization and forecasting of thunderstorm: specification of radar measurement errors. Preprints, In: 31st conference on radar meteorology, Seattle, WA, American Meteor Society, pp 146–149
- Purser RJ, Wu W-S, Parrish D, Roberts NM (2003) Numerical aspects of the application of recursive filters to variational statistical analysis. Part I: spatially homogeneous and isotropic Gaussian covariances. *Mon Weather Rev* 131:1524–1535
- Qiu CJ, Xu Q (1996) Least squares retrieval of microburst winds from single-Doppler radar data. *Mon Weather Rev* 124:1132–1144
- Rabier F, Järvinen H, Klinker E, Mahfouf J-F, Simmons A (2000) The ECMWF operational implementation of 4D variational assimilation. Part I: experimental results with simplified physics. *Q J R Meteor Soc* 126:1143–1170
- Ray PS, Johnson BC, Johnson KW, Bradberry JS, Stephens JJ, Wagner KK, Wilhelmsen RB, Klemp JB (1981) The morphology of several tornadic storms on 20 May 1977. *J Atmos Sci* 38:1643–1663
- Rihan FA, Collier CG, Ballard SP, Swarbrick SJ (2008) Assimilation of Doppler radial winds into a 3D-Var system: errors and impact of radial velocities on the variational analysis and model forecasts. *Q J R Meteor Soc* 134:1701–1716
- Sasaki Y (1955) A fundamental study of the numerical prediction based on the variational principle. *J Meteor Soc Jpn* 33:30–43
- Sasaki Y (1970a) Some basic formalisms in numerical variational analysis. *Mon Weather Rev* 98:875–883
- Sasaki Y (1970b) Numerical variational analysis formulated under the constraints as determined by longwave equations and a lowpass filter. *Mon Weather Rev* 98:884–898
- Sasaki Y (1970c) Numerical variational analysis with weak constraint and application to surface analysis of severe storm gust. *Mon Weather Rev* 98:899–910
- Sasaki Y, Mizuno K, Allen S, Whitehead V, Wilk KE (1989) Optimized variational analysis scheme of single Doppler radar wind data. Preprints. In: 3rd international conference on aviation weather systems, American Meteor Society, Boston, MA, pp 9–14
- Sasaki Y (2003) A theory of variational assimilation with Kalman filter-type constraints: bias and Lagrange multiplier. *Mon Weather Rev* 131:2545–2554
- Schenkman A, Xue M, Shapiro A, Brewster K, Gao J (2011) The analysis and prediction of the 8–9 May 2007 Oklahoma tornadic mesoscale convective system by assimilating WSR-88D and CASA radar data using 3DVAR. *Mon Weather Rev* 139:224–246
- Smith PL Jr, Myers CG, Orville HD (1975) Radar reflectivity factor calculations in numerical cloud models using bulk parameterization of precipitation processes. *J Appl Meteor* 14:1156–1165
- Smith TM, Gao J, Calhoun KM, Stensrud DJ, Manross KL, Ortega KL, Fu C, Kingfield DM, Elmore KL, Lakshmanan V, Riedel C (2014) Performance of a real-time 3DVAR analysis system in the Hazardous Weather Testbed. *Weather Forecasting* 29:63–77
- Snyder C, Zhang F (2003) Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon Weather Rev* 131:1663–1677
- Stensrud DJ, Xue M, Wicker LJ, Kelleher KE, Foster MP, Schaefer JT, Schneider RS, Benjamin SG, Weygandt SS, Ferree JT, Tuell JP (2009) Convective-scale warn on forecast: a vision for 2020. *Bull Am Meteor Soc* 90:1487–1499
- Stensrud DJ, Gao J (2010) Importance of horizontally inhomogeneous environmental initial conditions to ensemble storm-scale radar data assimilation and very short range forecasts. *Mon Weather Rev* 138:1250–1272
- Sun J, Flicker DW, Lilly DK (1991) Recovery of three-dimensional wind and temperature fields from simulated Doppler radar data. *J Atmos Sci* 48:876–890
- Sun J, Crook NA (1997) Dynamical and microphysical retrieval from Doppler radar observations using a cloud model and its adjoint. Part I: model development and simulated data experiments. *J Atmos Sci* 54:1642–1661
- Sun J, Crook NA (2001) Real-time low-level wind and temperature analysis using single WSR-88D data. *Weather Forecasting* 16:117–132

- Talagrand O, Courtier P (1987) Variational assimilation of meteorological observations with adjoint vorticity equation. I: theory. *Q J R Meteor Soc* 113:1311–1328
- Thacker C, Long R (1988) Fitting dynamics to data. *J Geophys Res* 93:1227–1240
- Tong M, Xue M (2005) Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSS experiments. *Mon Weather Rev* 133:1789–1807
- Wang X, Barker DM, Snyder C, Hamill TM (2008a) A hybrid ETKF–3DVAR data assimilation scheme for the WRF model. Part I: observing system simulation experiment. *Mon Weather Rev* 136:5116–5131
- Wang X, Barker DM, Snyder C, Hamill TM (2008b) A hybrid ETKF–3DVAR data assimilation scheme for the WRF model. Part II: real observation experiments. *Mon Weather Rev* 136:5116–5131
- Wang X, Parrish D, Kleist D, Whitaker J (2013) GSI 3DVar-based ensemble–variational hybrid data assimilation for NCEP global forecast system: single-resolution experiments. *Mon Weather Rev* 141:4098–4117
- Weygandt SS, Benjamin SG (2007) Radar reflectivity-based initialization of precipitation systems using a diabatic digital filter within the Rapid Update Cycle. In: 18th conference on numerical weather prediction, Park City, UT, American Meteor Society
- Wood VT, Brown RA (1997) Effects of radar sampling on single-Doppler velocity signatures of mesocyclones and tornadoes. *Weather Forecasting* 12:928–938
- Wu W-S, Purser RJ, Parrish D (2002) Three-dimensional variational analysis with spatially inhomogeneous covariances. *Mon Weather Rev* 130:2905–2916
- Xiao Q, Kuo Y, Sun J, Lee W, Lim E, Guo Y, Barker DM (2005) Assimilation of Doppler radar observations with a regional 3DVAR system: impact of Doppler velocities on forecasts of a heavy rainfall case. *J Appl Meteorol Climat* 44:768–788
- Xie Y, Koch SE, McGinley JA, Albers S, Wang N (2005) A sequential variational analysis approach for mesoscale data assimilation. Preprints. In: 21st conference on weather analysis and forecasting/17th conference on numerical weather prediction, Washington, DC, American Meteor Society, 15B.7. <http://ams.confex.com/ams/pdfpapers/93468.pdf>
- Xie Y, Koch SE, McGinley JA, Albers S, Bieringer P, Wolfson M, Chan M (2011) A space and time multiscale analysis system: a sequential variational analysis approach. *Mon Weather Rev* 139:1224–1240
- Xu Q, Qiu CJ (1994) Simple adjoint methods for single-Doppler wind analysis with a strong constraint of mass conservation. *J Atmos Oceanic Technol* 11:289–298
- Xu Q, Qiu CJ (1995) Adjoint-method retrievals of low-altitude wind fields from single-Doppler reflectivity and radial-wind. *J Atmos Oceanic Technol* 12:1111–1119
- Xu Q, Gu H, Yang S (2001) Simple adjoint method for three-dimensional wind retrievals from single-Doppler radar. *Q J R Meteor Soc* 127:1053–1067
- Xue M, Droegemeier KK, Wong V (2000) The Advanced Regional Prediction System (ARPS)—a multiscale nonhydrostatic atmospheric simulation and prediction tool. Part I: model dynamics and verification. *Meteor Atmos Phys* 75:161–193
- Xue M, Wang D, Gao J, Brewster K, Droegemeier KK (2003) The Advanced Regional Prediction System (ARPS), storm scale numerical weather prediction and data assimilation. *Meteor Atmos Phys* 82:139–170
- Xue M, Kong F, Thomas KW, Wang Y, Brewster K, Gao J, Wang X, Weiss S, Clark A, Kain J, Coniglio M, Du J, Jensen T, Kuo Y.-H. (2010) CAPS realtime storm scale ensemble and high resolution forecasts for the NOAA Hazardous weather Testbed 2010 spring experiment. In: 25th conference severe local storms, American Meteor Society, Paper 7B.3
- Yussouf N, Mansell ER, Wicker LJ, Wheatley DM, Stensrud DJ (2013) The ensemble Kalman filter analyses and forecasts of the 8 May 2003 Oklahoma City tornadic supercell storm using single- and double-moment microphysics schemes. *Mon Weather Rev* 141:3388–3412
- Zhang F, Snyder C, Sun J (2004) Impacts of initial estimate and observations on the convective-scale data assimilation with an ensemble Kalman filter. *Mon Weather Rev* 132:1238–1253

- Zhang F, Zhang M, Poterjoy J (2013) E3DVar: coupling an ensemble Kalman filter with three-dimensional variational data assimilation in a limited-area weather prediction model and comparison to E4DVar. *Mon Weather Rev* 141:900–917
- Zhang J, Carr F, Brewster K (1998) ADAS cloud analysis. Preprints. In: 12th conference on numerical weather prediction, Phoenix, AZ, American Meteor Society, pp 185–188
- Zrnic SD, Ryzhkov A, Straka J, Liu Y, Vivekanandan J (2001) Testing a procedure for automatic classification of hydrometeor types. *J Atmos Oceanic Tech* 18:892–912

Data Assimilation Experiments of Refractivity Observed by JMA Operational Radar

Hiromu Seko, Ei-ichi Sato, Hiroshi Yamauchi and Toshitaka Tsuda

Abstract Small scale distributions of water vapor that express convection cells are needed to improve numerical forecasts of heavy rainfall. In this study, temporal variations of refractivity (TVR), which include information of small scale water vapor variations, were obtained from the phase data of radio waves of the JMA's operational Doppler radar. The TVR distributions on August 4th 2008 showed that the increased and decreased regions of TVR moved smoothly, corresponding to the movements of sea breeze fronts. These smooth variations indicated that the observed TVR were caused by the atmosphere. The TVR obtained by the operational Doppler radar was assimilated by a nested Local Ensemble Transform Kalman Filter system. The reproduced distributions indicated that the data assimilation of TVR made the rainfall distributions closer to the observed ones by modifying water vapor distributions. This result shows that TVR has the potential to improve rainfall forecasts.

1 Introduction

In general, heavy rainfalls are generated by convergences of low-level humid air-flows. Because water vapor greatly affects initiations and developments of convection cells of heavy rainfalls, water vapor distribution is needed as assimilation data to increase the accuracy of rainfall forecasts. Conventional observations of water vapor near the surface are conducted at the Meteorological Observatories of the Japan Meteorological Agency (JMA). However, the horizontal resolution of the

H. Seko (✉) · E. Sato
MRI, Tsukuba, Japan
e-mail: hseko@mri-jma.go.jp

H. Yamauchi
JMA, Tokyo, Japan

T. Tsuda
RISH/Kyoto University, Uji, Japan

Meteorological Observatories (several 10 km) is too coarse to express convection cells.

As for the observations that express small scale variations of water vapor, we focused on phase of radio waves of Doppler radars. The radio waves are delayed by the atmosphere while they travel back and forth between a Doppler radar and stationary structures (e.g. tall buildings, towers for electric power supply etc.) (Fabry et al. 1997). In this method, refractivity that causes delays of radio waves is obtained by monitoring their phases received by Doppler radar. Rainfall forecasts are expected to be improved when the refractivity is used as assimilation data, because refractivity that is a function of temperature and water vapor affects the initiations and developments of convection cells.

In the U.S., it was shown that the refractivity may have the potential for forecasting convection initiations by using the observation data of the International H2O Project (Weckwerth et al. 2005). Several studies for assimilations of refractivity data have been performed so far. For instance, the effect of vertical variation of refractivity on the refractive index along the path was investigated (Park and Fabry 2010; Feng et al. 2016). Nicol et al. (2013) showed that hourly refractivity changes up to a height of 100 m observed by a meteorological tower were well correlated with the water vapor near the surface. The subtract method of noisy data from refractivity field was considered by Nicol et al. (2012).

However, few studies that had showed the variations of refractivity have been conducted in Japan (e.g. Seko et al. 2009). Because the low-level atmosphere in the summer in Japan is more humid than that in the U.S., investigations of relations between refractivity and convection cells, and developments of the assimilation methods of refractivity in the humid atmosphere are still desired.

In this study, Local Ensemble Transform Kalman Filter (LETKF, Hunt et al. 2007) based on the Japan Meteorological Agency Non-Hydrostatic Model (JMA-NHM, Saito et al. 2006), known as the NHM-LETKF (Miyoshi and Aranami 2006), was used as a data assimilation system. LETKF is one method of ensemble Kalman filters (EnKFs), which can provide initial conditions by assimilation of observation data. Some previous studies have used the EnKF for mesoscale application and have showed promising results so far (e.g., Snyder and Zhang 2003; Zhang et al. 2004, 2006; Dowell et al. 2004; Tong and Xue 2005; Xue et al. 2006; Meng and Zhang 2008; Seko et al. 2011; Miyoshi and Kunii 2012).

In this study, temporal variations of refractivity (TVR), not total values of refractivity were used as assimilation data, because the total values of radio waves' delays between radar and stationary structures cannot be determined due to ambiguities of reflection points of stationary structures. The horizontal distributions and temporal variations of TVR, and the preliminary results of data assimilation of TVR are presented in this chapter.

2 An Estimation Method of Temporal Variations of Refractivity

In this study, TVR data was obtained from the phase data observed by the Tokyo operational radar of JMA on August 4th 2008. If the estimation method of TVR from the phase data of operational radars is developed, the numerical forecasts for initiations and developments of convection cells will be improved in many areas in Japan, because the developed method is applicable to other operational radars.

Firstly, the procedures to obtain TVR from the phase data, namely In-phase/Quadrature-phase (IQ) data of Doppler radar observation are explained. If TVR is returned from stationary structures in wider areas in the radar range, TVR works as assimilation data more effectively in improving the initial conditions of numerical forecasts. To obtain TVR data from wider areas, the elevation angle of 0.0° was adopted in the observation of TVR data.

As mentioned in the Introduction, it is difficult to obtain the total values of delays from the phase data. Instead of the total values, ‘temporal variation of delays’ (abbreviated to TVD), that is, ‘temporal variation of the difference between transmitted and received phases’, was used. Just like the relation between the delay and refractivity, TVD obtained from the phase data is the integrated value of TVR along the path of radio waves. The TVR, which will be used as assimilation data, was produced by taking the difference of TVD in the radial direction of radar (Fig. 1).

Because the radio waves received by radars are not only those reflected from stationary structures, it is necessary to remove the delays reflected from moving structures, such as trees and wind turbines and so on. In Seko et al. (2009), the radio waves reflected from the moving structures were removed from all of reflected radio waves by using the temporal dispersion of TVD as the threshold. Namely, radio waves of which TVD were greatly fluctuated during a short time were removed as those reflected from moving structures.

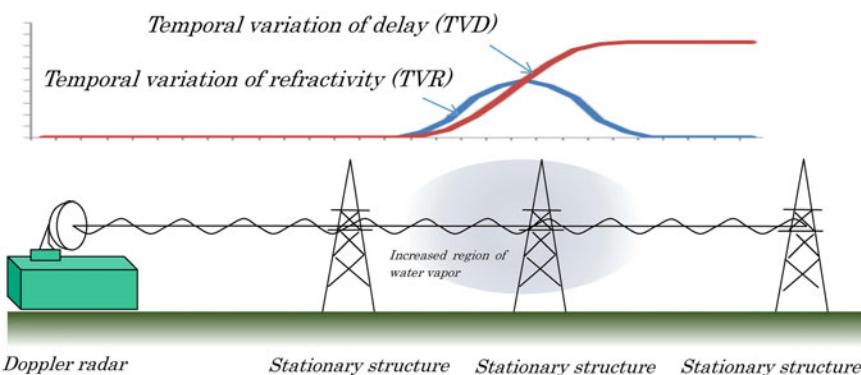


Fig. 1 Schematic illustration on the estimation method of temporal variation of refractivity (TVR)

As described earlier, the low elevation angle of 0.0° was used in the observation of TVD. However, the JMA operational radars conduct volume scan observations by changing the elevation angles to obtain 3-dimensional distributions of water substances and radial wind. That is, TVD data with the elevation angle of 0.0° are obtained once every 10 min. Because it is difficult to apply the method of Seko et al. (2009) to the intermittent data obtained by the operational radars, horizontal variations of TVD were used, instead of the temporal fluctuations of TVD. The radio waves that were reflected from moving structures were removed as follows; (1) averaged values of TVD over small areas (2.0° in tangential direction and 0.75 km in radial direction), of which areas were produced by partitioning whole radar range, were obtained. (2) The radio waves, of which TVD were far from their area averages, were removed. The radio waves reflected from stationary structures are expected to be remained through a few iterations of these procedures. The threshold value used in identifying the moving structures in this study was determined by the trial and error method.

3 Temporal Variations of Refractivity on August 4th 2008

Figure 2 shows the increment distributions of TVD during 1607-1617 JST and 1527-1617 JST of August 4th 2008. These distributions, which were produced from the observation data of the JMA's Tokyo radar with the elevation angle of 0.0° , show that signals were returned from many stationary structures around the radar. In addition to the echoes near the radar, there were line-shaped echoes (indicated by arrows in Fig. 2), which seems to be reflected from the towers for electric power supply. Increments during 10 min (abbreviated to 10-min increment, hereinafter) had

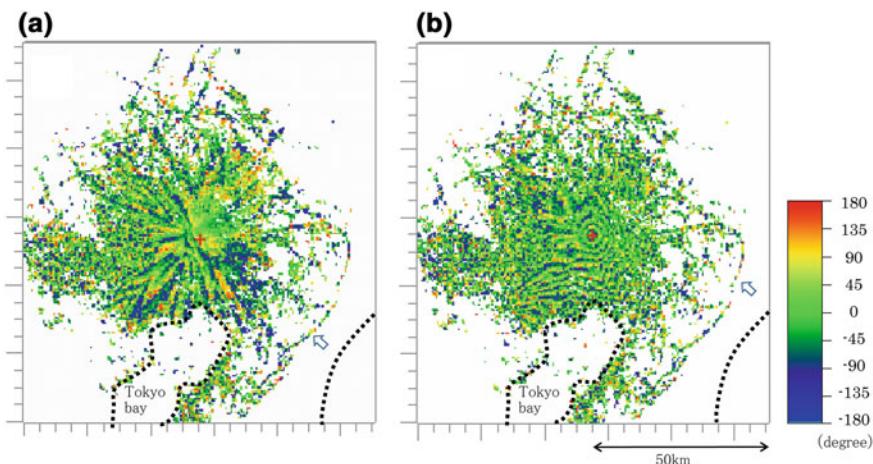


Fig. 2 **a** 10-min increment of TVR (1607-1617 JST) and 50-min increment of TVR (1527-1617 JST) on August 4th 2008. White arrows indicate the line-shaped regions. Red crosses are the positions of Tokyo radar of JMA. Black broken lines indicate coastlines

a radial pattern (Fig. 2a). This pattern suggests that water vapor near the radar varied greatly and affected TVD of which the radio waves returned from the stationary structures far from the radar. The pattern was changed to a concentric circle when the monitoring period became as long as 50 min (Fig. 2b). This pattern indicates that the phases in wide areas around the radar were greatly changed during 50 min.

Next, TVR distributions were produced by taking the difference of TVDs between adjacent areas in the radial direction. In this study, radio waves, of which spatial dispersion of TVDs was less than 60.0° from the area average, were used. Figure 3 is the 10-min increment distributions of TVR around the radar from 1547 JST to 1642 JST. The line-shaped increased regions (warm color regions indicated by arrows) moved from south and east, and merged around the radar. From 1637 JST, the line-shaped decreased region (indicated by cold colors), which extended in the northwest-southeast direction, was gradually moving northward. These smooth variations of increments indicate that they were caused by the atmosphere such as refractivity.

To confirm that these complicated distributions of increments can be produced by the atmosphere, a deterministic forecast with the initial time of 9 JST August 4th 2008, in which the mesoscale analyses of JMA were used as initial and boundary conditions, was performed. The distributions reproduced by the deterministic forecast show that there were a few line-shaped increased and decreased regions of TVR that extended in the north-south and east-west directions (indicated by green arrows in Fig. 4). These regions were located along the fronts of sea breezes, entering the Kanto Plain from east and south. These features of TVR distribution are common with the observed ones, though the timing and location were different from the observed ones. This similarity of line-shaped increased and decreased regions suggests that the increments of TVR were produced by the airflows of sea breezes.

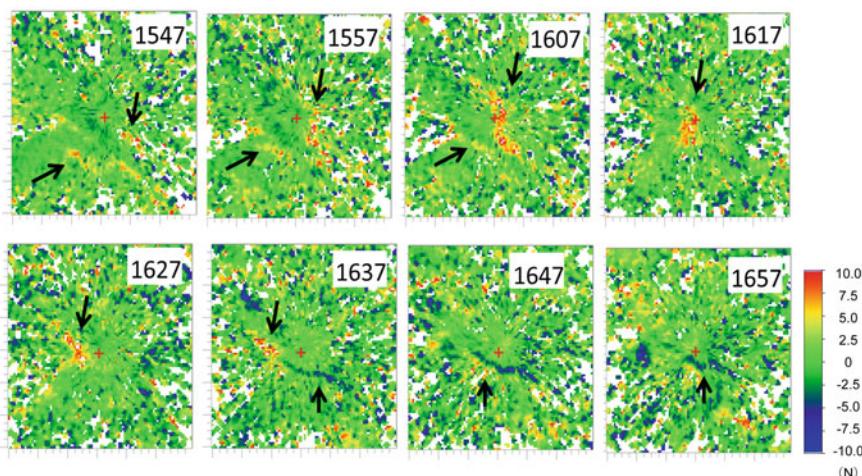
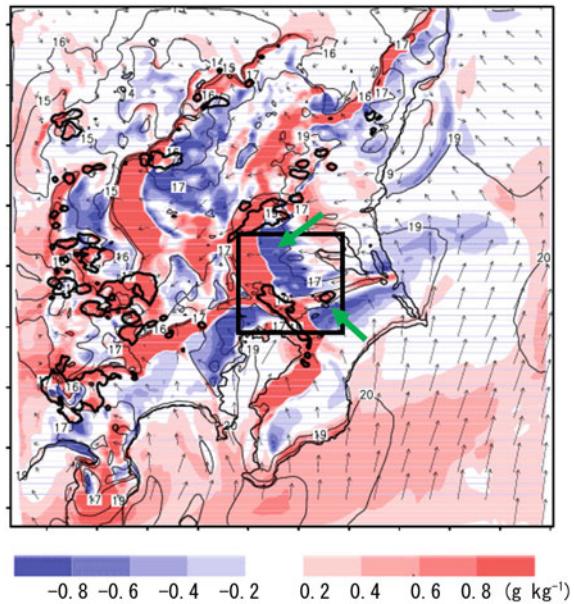


Fig. 3 Horizontal distributions of 10-min increment of TVR from 1547 JST to 1657 JST. Red crosses indicate the positions of the Tokyo radar. Arrows indicate increased or decreased regions of TVR that moved smoothly

Fig. 4 Increment of water vapor during 1 h at 1530 JST (colored region) reproduced by the deterministic forecast. Arrows and contours indicate the horizontal wind and water vapor near the surface (thin lines) and rainfall regions (thick lines), respectively. Thick green arrows indicate line-shaped increased and decreased regions of TVR. A black square indicates the region shown in Fig. 3



4 Data Assimilation Methods and Impact of Refractivity

The system used in this experiment is a two-way nested system of LETKF (Seko et al. 2013). The schematic illustration of this experiment is shown in Fig. 5. This system is composed of two LETKFs; Outer and Inner LETKFs. Outer and Inner LETKFs were performed to reproduce mesoscale convergences and convection cells, respectively. Outer LETKF with the grid interval of 15 km was performed

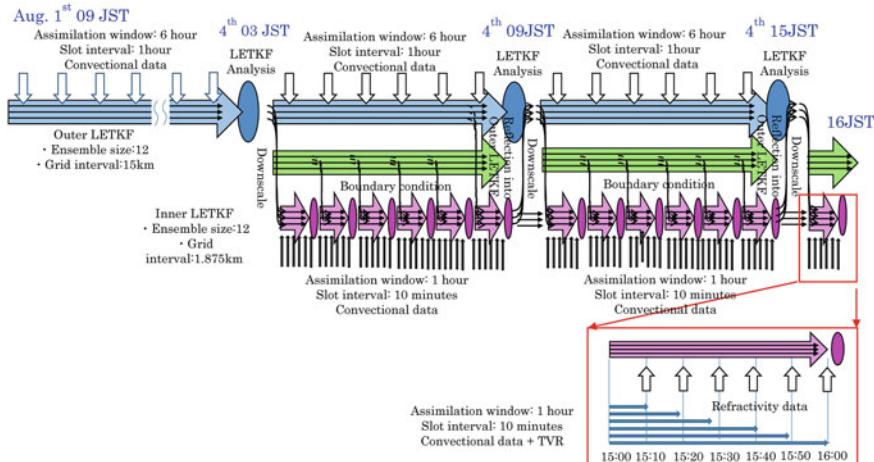


Fig. 5 Schematic illustration of the experiments using a two-way nested LETKF system

from 9 JST 1st August, and the assimilation of Inner LETKF with the grid interval of 1.875 km was started at 3 JST 4th. The grid number of Outer LETKF was $80 \times 80 \times 50$ so that the domain of Outer LETKF covered Japan except Kyushu and Hokkaido. The grid number of Inner LETKF was $160 \times 160 \times 50$ to cover the Kanto Plain and surrounding area. The ensemble size of both LETKFs was 12. The conventional data of JMA, such as upper sounding data, surface pressure and so on, were assimilated by Outer and Inner LETKFs, and TVR data was added to the conventional data of Inner LETKF. The slot intervals of Outer and Inner LETKFs were 1 h and 10 min, respectively.

As explained in Sect. 2, the temporal variations of TVR from a given time were obtained from the phase data of Doppler radar. We assumed that the analyzed distribution of 1500 JST is close to the reality, and the refractivity distributions from 1510 JST to 1600 JST were obtained by adding the increments of TVR to this analyzed distribution. The refractive index distributions from 1510 JST to 1600 JST were used as the assimilation data of Inner LETKF. Observation operator for refractive index is as follows;

$$N = (n - 1) \times 10^6 = 77.6 \frac{P}{T} + 373000 \frac{e}{T^2},$$

where n , N , e , P and T are refractivity, refractive index, partial pressure of water vapor, pressure and temperature, respectively. The observation error of refractive index was 10 N, which was determined by trial and error. The height of TVR data was assumed to be 60 m, which is the height of the antenna of the radar.

Figure 6 shows the rainfall regions observed by the conventional radars at 15 JST and 16 JST on August 4th 2008 and the increment distribution of TVR during 1600–1610 JST. At 15 JST, there were rainfall regions over the mountainous area

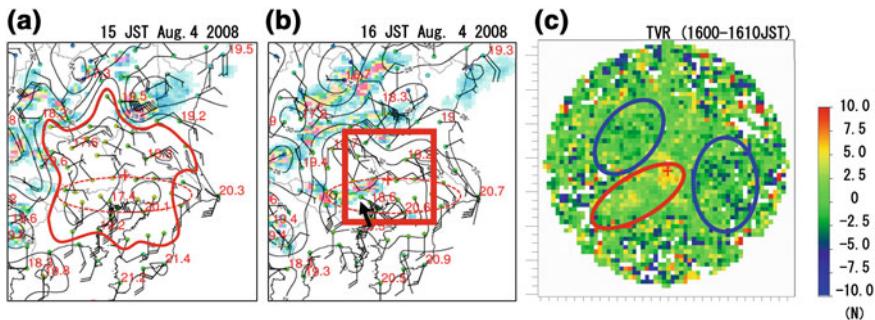


Fig. 6 **a, b** Rainfall distributions observed by operational radars. *Contours and barbs* indicate the surface temperature and horizontal wind observed by AMeDAS of JMA. *Red numbers* indicate water vapor mixing ratio observed by Meteorological Observatories of JMA. A *red contour* in (a) indicates the region of temperature at 32°C . A *red square* in (b) is the region of (c). *Thick black arrow* in (b) indicates a rainfall band generated at the southern part of the Kanto Plan. **c** Horizontal distribution of 10-min increment of TVR (1600–1610 JST). *Red crosses* indicate the positions of the Tokyo radar

surrounding the Kanto Plain. The temperature in the Kanto Plain exceeded 32 °C, and the airflows from east and south were converged at the southern part of the Kanto Plain (indicated by red circles with broken lines in Fig. 6a, b). This convergence of high temperature airflows is a favorable condition for the generation of convection cells. The rainfall band extending in the east-west direction was generated there at 16 JST (indicated by a black arrow). The increment distribution of TVR when the rainfall band was generated indicates that the increased region of refractivity extended west-southwestward from the radar (indicated by a red circle). This region where TVR was increased is consistent with the generation of rainfall band there. In addition to this increased region, refractivity was decreased at the northwestern and eastern sides of the radar (indicated by blue circles in Fig. 6c).

The rainfall distributions that were reproduced by Inner LETKF with or without the assimilation of TVR data are shown in Fig. 7a, b. The rainfall regions at the mountainous area surrounding the Kanto Plain were reproduced in both analyses. In addition to the mountainous area, a rainfall region extending in the east-west direction was reproduced at the southern part of the Kanto Plain (indicated by black arrows in Fig. 7a, b), though the time of generation was 1 h earlier than the observed one. On the eastern side of the radar, fake rainfall regions were generated (indicated by red circles with broken lines). These fake rainfall regions became weaker (Fig. 7b) when the TVR data was assimilated. The difference of the analyzed water vapor distributions (Fig. 7c) shows that water vapor was decreased there (indicated by red circles with broken line). It is deduced that the assimilation of TVR data suppressed the fake rainfall through the decrease of water vapor in the eastern part of the Kanto Plain. This result indicates that TVR has the potential to improve the rainfall forecasts.

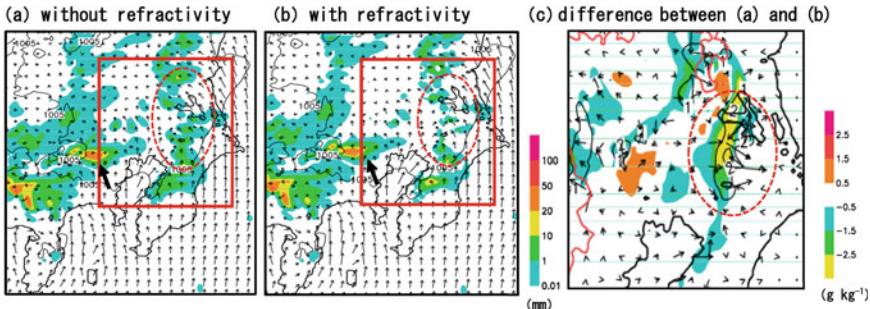


Fig. 7 Data assimilation results of TVR data. Colored regions in (a) and (b) are ensemble mean distributions of rainfall at 16 JST. Arrows show the horizontal wind near the surface. Thick black arrows in (a) and (b) indicate the rainfall band generated at the southern part of the Kanto Plain. Colored regions, black contours and arrows in (c) show the difference of analyzed water vapor, temperature at the height of 60 m and horizontal wind near the surface. Red squares in (a) and (b) indicate the region of (c). Red circles with broken lines indicate the regions where fake rainfall regions were weakened by the assimilation TVR data. Red contours in (c) depict the topography

5 Summary

In this study, TVR data was obtained from the phase data of JMA's operational Doppler radar. The potential of TVR as the assimilation data for the improvement of rainfall forecasts was shown in this chapter. In addition to the assimilation data, TVR is expected to be useful data for monitoring the initiations of convection cells because TVR shows the positions of sea breezes.

However, there are several points that should be investigated. Namely, the observation error was determined by trial and error, and the paths of radio waves were assumed to propagate at the altitude of the radar's antenna in this study. Although the effect of vertical variation of refractivity on the refractive index along the path in the U.S. was investigated by Park and Fabry (2010), the effect in Japan might be different from that in the U.S. because the low-level atmosphere in the summer in Japan is more humid. It should also be confirmed that TVR caused by sea breezes is larger than the effects of the bending of radio waves caused by the vertical variation of refractivity. To get more conclusive results about the impact of radar refractivity data on rainfall forecasts, the effects of radio waves' bending and the observation error of TVR should be investigated, and more data assimilation experiments using TVR data are needed.

Acknowledgements The authors would like to thank Mr. Osamu Suzuki, Mr. Yoshihisa Kimata and Mr. Takanori Sakanashi of JMA for many important suggestions on this study. The initial seeds and boundary conditions for Outer LETKF and the conventional observation data, the phase data of Tokyo operational radar were provided from the JMA. This work was supported in part by the research projects of "Tokyo Metropolitan Area Convection Study for Extreme Weather Resilient Cities (TOMACS)" and Grants-in-Aid for Scientific Research "Establishment of extraction methods of water vapor information from phase data of Doppler radar".

References

- Dowell DC, Zhang F, Wicker LJ, Snyder C, Crook NA (2004) Wind and temperature retrievals in the 17 May 1981 Arcadia, Oklahoma, Supercell: ensemble Kalman filter experiments. *Mon Weather Rev* 132:1982–2005
- Fabry F, Frush C, Zawadzki I, Kilambi A (1997) On the extraction of near-surface index of fraction using radar phase measurements from ground targets. *J Atmos Oceanic Technol* 14:978–987
- Feng Y, Fabry F, Weckwerth T (2016) Improving radar refractivity by considering the change of vertical refractivity profile and the varying altitudes of ground targets. *J Atmos Oceanic Technol* doi:10.1175/JRECH-D-15-0224.1, in press
- Hunt BR, Kostelich EJ, Szunyogh I (2007) Efficient data assimilation for spatiotemporal chaos: a local ensemble transform Kalman filter. *Physica D* 230:112–126
- Meng Z, Zhang F (2008) Test of an ensemble Kalman filter for mesoscale and regional-scale data assimilation. Part III: comparison with 3DVar in a real-data case study. *Mon Weather Rev* 136:522–540
- Miyoshi T, Aranami K (2006) Applying a four-dimensional local ensemble transform Kalman filter (4D-LETKF) to the JMA nonhydrostatic model (NHM). *SOLA* 2:128–131

- Miyoshi T, Kunii M (2012) The local ensemble transform Kalman filter with the weather research and forecasting model: experiments with real observations. *Pure appl Geophys* 169:321–333
- Nicol JC, Illingworth AJ (2012) The effect pf phase-corrected returns and spatial smoothing on the accuracy of radar refractivity retrievals. *J Atmos Ocean Technol* 30:22–39
- Nicol JC, Illingworth AJ, Darlington T, Kitchen M (2013) Quantifying errors due to frequency changes and target location uncertainty for radar refractivity retrievals. *J Atmos Ocean Technol* 30:2006–2024
- Park S, Fabry F (2010) Simulation and interpretation of the phase data used by the Radar refractivity retrieval algorithm. *J Atmos Ocean Technol* 27:1287–1301
- Saito K, Fujita T, Yamada Y, Ishida J, Kumagai Y, Aranami K, Ohmori S, Nagasawa R, Kumagai S, Muroi C, Kato T, Eito H, Yamazaki Y (2006) The operational JMA nonhydrostatic mesoscale model. *Mon Weather Rev* 134:1266–1298
- Seko H, Miyoshi T, Shoji Y, Saito K (2011) Data assimilation experiments of precipitable water vapor using the LETKF system: intense rainfall event over Japan 28 July 2008. *Tellus* 63A:402–412
- Seko H, Yamauchi H, Suzuki O, Saito K (2009) Estimation of temporal variation of refractive index using c-band doppler radar equipped with magnetron transmitter. *SOLA* 5:145–148
- Seko H, Tsuyuki T, Saito K, Miyoshi T (2013) Development of a two-way nested LETKF system for cloud-resolving model. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications, vol 2. Springer, pp 489–507
- Snyder C, Zhang F (2003) Assimilation of simulated Doppler radar observations with an ensemble Kalman filter. *Mon Weather Rev* 131:1663–1677
- Tong M, Xue M (2005) Ensemble Kalman filter assimilation of Doppler radar data with a compressible nonhydrostatic model: OSSE experiments. *Mon Weather Rev* 133:1789–1807
- Weckwerth TM, Pettet CR, Fabry F, Park SJ, LeMone MA, Wilson JW (2005) Radar refractivity retrieval: validation and application to short-term forecasting. *J Appl Meteorol* 44:285–300
- Xue M, Tong M, Droegeemeier KK (2006) An OSSE framework based on the ensemble square root Kalman filter for evaluating the impact of data from radar networks on thunderstorm analysis and forecasting. *J Atmos Ocean Technol* 23:46–66
- Zhang F, Snyder C, Sun J (2004) Impacts of initial estimate and observation availability with an ensemble Kalman filter. *Mon Weather Rev* 132:1238–1253
- Zhang F, Meng Z, Aksoy A (2006) Test of an ensemble Kalman filter for mesoscale and regional scale data assimilation. Part I: perfect model experiments. *Mon Weather Rev* 134:722–736

Assessment of Radiative Effect of Hydrometeors in Rapid Radiative Transfer Model in Support of Satellite Cloud and Precipitation Microwave Data Assimilation

Peiming Dong, Wei Han, Wei Li and Shuanglong Jin

Abstract To date, various satellite observations have been assimilated in numerical weather prediction (NWP), making a dramatic contribution to the improvement of forecast accuracy. However, the focus has mainly been on satellite measurements in a clear atmosphere, with relatively little attention paid to exploring the use of satellite observations under cloudy and rainy conditions. One of the key issues related to cloud and precipitation data assimilation is the radiative effect of hydrometeors in the rapid radiative transfer model (RRTM). This paper presents a detailed assessment of the radiative effect of hydrometeors in the RRTM, in support of satellite cloud and precipitation microwave data assimilation. Using the output of hydrometeors from the Weather Research and Forecasting (WRF) numerical forecast model as the input for the Community Radiative Transfer Model (CRTM), the radiative effect of hydrometeors on the simulation of Advanced Microwave Sounding Unit (A and B) (AMSU-A and AMSU-B, respectively) satellite observations are analyzed. Then, the sensitivity of the satellite simulation to hydrometeors' properties, including the water content, particle size and vertical distribution, are investigated. Finally, the result of CRTM is compared and discussed with that of Radiative Transfer for TOVS (RTTOV)—another popular RRTM used in the numerical weather prediction community. The results show that the inclusion of the radiative effect of hydrometeors in the RRTM makes the satellite brightness temperature simulation match up, reasonably successfully, with the observation. Overall, the radiative effects of hydrometeors have diverse influences in most of the channels of the satellite microwave observations, except the AMSU-A high-level temperature sounding channels 10–14. Investigation of the radiative effect of the individual hydrometeors verifies that the cloud and rain water mainly have a warming effect, owing to radiative emission. This effect dominates three of the

P. Dong

Beijing Piesat Information Technology Co., Ltd., Beijing 100195, China

P. Dong (✉) · W. Han · W. Li · S. Jin

Numerical Prediction Center, Chinese Meteorological Administration, Beijing 100081, China
e-mail: dongpm@cams.cma.cn

AMSU-A window channels, 1–3, but is weakened both in the other AMSU-A window channel, 15, and in the AMSU-B window channel 1. The effect of cloud and rain water in the other channels is one of scattering, which decreases the brightness temperature. Ice, snow and graupel all present a cooling effect, owing to scattering. The variation produced by ice is extremely small, but is obvious in AMSU-B. The effect of both snow and graupel is notable in all AMSU-A and AMSU-B channels. The sensitivity of satellite microwave remote sensing to the water content of hydrometeors corresponds well with their radiative effect. The brightness temperature is not sensitive to the effective radius size of cloud and ice, and the sensitivity of satellite observations to the particle size of rain, snow and graupel is strong. Also, the sensitivity is complicated by the frequency. The sensitivity of the satellite simulation to the vertical distribution of hydrometeors is presented by the transfer of the particular channel affected. Additionally, the results of RTTOV and CRTM are generally consistent. The main discrepancy is the magnitude of the response function of hydrometeors and the corresponding deviation of brightness temperature produced by the radiative effect of hydrometeors. The result in CRTM appears to be at least double the magnitude of that in RTTOV.

1 Introduction

Owing to the great social, ecological and economic impacts of weather, the improvement of forecast accuracy has emerged as a high-priority topic in the atmospheric sciences. It is recognized that the use of observations has a great impact on numerical weather prediction. Through data assimilation, diverse observations of the atmosphere, land and ocean, sampled at different times, intervals and locations are combined with a priori knowledge of the evolving atmospheric state to produce the optimal analysis for NWP (Rodgers 2000). Satellites provide a very useful source of observations. Along with advancements in remote sensing and data assimilation techniques, the amount of satellite data that is usable in NWP has been increasing rapidly. To date, the greatest volume of data assimilated into numerical forecasts derives from satellite observations, contributing a dramatic improvement in forecast accuracy over the last two decades at publication (Mahfouf et al. 1999).

However, currently, most assimilated satellite measurements are of the clear-sky type, with many satellite data affected by cloud and precipitation rejected. To utilize the information provided by satellite observations fully, in all weather conditions, and acquire continual and substantial improvements in NWP, research associated with the assimilation of satellite cloudy data has been pursued, but remains challenging problems (McNally 2002, 2009; Weng and Liu 2003; Bauer et al. 2006a, b, 2010, 2011; Pu 2009; Geer et al. 2010; Županski 2013; Okamoto et al. 2014). Specifically, an advanced radiative transfer model to handle the scattering and emissions that dominate cloud and precipitation radiative transfer is needed, thus enabling better use of satellite observations made under cloudy and rainy conditions

(Errico et al. 2007; Stengel et al. 2010). Recently, the Rapid Radiative Transfer Model, which is a crucial component of satellite data assimilation in NWP, has been extended from clear atmospheric conditions to also incorporate cloud absorption and scattering from hydrometeors. The two popular RRTMs used in the NWP community are CRTM (Weng 2007) and RTTOV (Saunders et al. 1999), developed by the Joint Center for Satellite Data Assimilation, USA, and the European Organization for the Exploitation of Meteorological Satellites, respectively.

The validation and assessment of the RRTM under cloudy conditions is a very important and necessary step not only for improvement to the RRTM, but its application as well. It allows for forward model biases to be quantified and error characteristics to be well understood, under various cloudy conditions (Geer et al. 2011). For example, the observed brightness temperature from NOAA-18 instruments has been compared to those simulated by the CRTM using the inputs of CloudSat-retrieved hydrometeor profiles (Chen et al. 2008). Only the cloud and ice water absorption and scattering are computed in this study with the limitation of the hydrometeor retrievals. Because of the lack of accurate and sufficiently detailed cloud microphysical profiles, various hydrometeor profiles from weather forecast models and cloud resolving models have generally been used for radiative transfer simulations (e.g. Kummerow 1993; Burns et al. 1997; Skofronick-Jackson et al. 2002; Bennartz and Bauer 2003; Chevallier et al. 2003; Hong et al. 2005; Sreerekha 2008; Hong et al. 2010, Han and Dong 2012). The hydrometeor profiles from the MM5 model and Met Office mesoscale model were used in a simulation of the satellite observation of Super Typhoon Paka and a mid-latitude front by Kummerow (1993) and Sreerekha et al. (2008), respectively. Hong et al. (2010) input the hydrometeors from a three-dimensional deep convective cloud model into a radiative transfer model to understand the effects of clouds on the Advanced Microwave Sensor Unit-B channels. It is cloud and ice water only that were used by Sreerekha et al. (2008). Most of the work mentioned above utilized five hydrometeors, including cloud liquid water, rain water, cloud ice, snow and graupel. The same five-class hydrometeors from a WRF model forecast were used by Han and Dong (2012) in their study of the RRTM in the application of satellite data assimilation. Up to seven hydrometeor classes were used by Bennartz and Bauer (2003) in their investigation of a convective system over the eastern Mediterranean.

This paper presents a detailed assessment of the radiative effect of hydrometeors in the RRTM by taking inputs from a numerical model forecast. Section 2 briefly describes the process that accounts for the radiative effect of hydrometeors in the RRTM. The data and methods are introduced in Sect. 3. In Sect. 4, the simulation of satellite observations made under cloudy and rainy conditions is analyzed, with the sensitivity of the satellite simulation to hydrometeors' properties examined in Sect. 5. Results from CRTM and RTTOV are compared and discussed in Sect. 6. And finally, a summary and further discussion comprise Sect. 7.

2 The Process of the Radiative Effect of Hydrometeors in the RRTM

Satellite data assimilation using the variational analysis approach involves combining the satellite observation with *a priori* initial guess. The best estimate of the atmospheric state, x , is obtained by minimizing the cost function (Rodgers 2000):

$$J = \frac{1}{2}(x - x^b)^T B^{-1}(x - x^b) + \frac{1}{2}[H(x) - y^o]^T (E + F)^{-1}[H(x) - y^o],$$

where, x^b is the background state and B is the associated error covariance matrix; y^o is the observed satellite measurement; $H(x)$ is the simulated satellite data at the state variable x through the observation operator H , which is the RRTM for satellite measurement; and E and F are the error matrices associated with the observation and forward model, respectively.

The gradient of the cost function at the iteration process of the minimum is:

$$\nabla_x J = B^{-1}(x - x^b) + H^T (E + F)^{-1}[H(x) - y^o],$$

where, H^T is the adjoint operator and the derivative of the radiance with respect to the state variables. So, an RRTM includes both fast forward and Jacobian radiative transfer models that are capable of producing the model simulation of radiance from the satellite instrument and the gradients of a simulated brightness temperature with respect to input variables.

In CRTM, for a plane-parallel atmosphere, the radiance emanating to the top of the atmosphere is obtained from the differential form of the radiative transfer equation (Weng 2007; Han et al. 2012):

$$\mu \frac{dI(\tau, \mu, \phi)}{d\tau} = -I(\tau, \mu, \phi) + S(\tau, \mu, \phi; \mu_0, \phi_0) + \frac{\varpi}{4\pi} \int_0^{2\pi} \int_{-1}^1 M(\tau; \mu, \varphi; \mu', \phi') I(\tau, \mu', \phi') d\mu' d\phi' \theta,$$

where I is the radiance; S and M are the source term and scattering phase matrix, respectively; μ_0 and ϕ_0 are the cosines of the zenith angle and the azimuthal angle of the sun, respectively; μ and ϕ are the cosines of the zenith angle and the azimuthal angle in the scattering direction, respectively; ϖ is the single scattering albedo; and τ is the optical thickness.

The first and second terms on the right-hand side of the equation are the transmitted radiance from the bottom of the layer and the emitted radiance by the layer, which can be derived from the atmospheric temperature and optical parameters. The third integral term contains the contributions due to multiple scattering. The total radiance from the first two terms approaches the exact emission solution as the scattering approaches zero. The multiple scattering is evaluated using a two-stream solution with a satisfactory accuracy under more cloudy conditions;

although, the more streams that are used for multiple scattering, the more exact the solution becomes. To overcome the computational expense of the term-by-term Mie calculation of the scattering coefficients and phase-matrix parameters, the microphysical properties (extinction coefficient, single-scattering albedo, asymmetry factor, and Legendre phase function coefficients) of cloud particles are stored in lookup tables. Up to six particles, including water, rain, ice, snow, graupel and hail, are treated presently.

A similar radiative theoretical formula, albeit with different treatment and code construction, is used in RTTOV. The top of the atmosphere upwelling radiance, $L(v, \theta)$, at a frequency v and viewing angle θ , is combined linearly (Bauer et al. 2006a, b; James et al. 2014):

$$L(v, \theta) = (1 - N) L^{Clr}(v, \theta) + N L^{Cld}(v, \theta).$$

where, $L^{Clr}(v, \theta)$ and $L^{Cld}(v, \theta)$ are the clear-sky and cloudy or rainy top of the atmosphere upwelling radiance, respectively. Note that N , the effective cloud fraction in the vertical profile, is a key argument for the simulation in that the two independent columns are linearly combined by it to produce the final result. Moreover, the scattering effect for the infrared and microwave radiance uses completely different approaches: it is parameterized for the former and treated explicitly for the latter. For the simulation of microwave radiance scattered by cloud and precipitation, the core RTTOV module is used for the clear-air part, while a separate interface, RTTOV-SCATT, is provided to add the scattering effect from hydrometeors in the atmosphere profile. The scattering effect of hydrometeors at microwave frequencies is computed using the delta-Eddington approximation, and the Mie scattering properties are also handled through the lookup tables. The hydrometeor variables involved in cloudy profile include cloud liquid water, cloud ice water and solid precipitation rain and snow.

3 Data and Methods

3.1 Satellite Observation

It is satellite microwave observations that form the focus of the present study. The satellite data from the Advanced Microwave Sounding Unit (AMSU) onboard NOAA-16 are used. AMSU is the second—the replacement version—of the four-channel Microwave Sounding Unit, which was first launched in 1978, onboard TIROS-N. Also, it is the predecessor of the Advanced Technology Microwave Sounder, onboard the Suomi National Polar Orbiter Partnership satellite, which combines and inherits most of the channels from AMSU. AMSU is a cross-track scanning microwave radiometer for sounding the atmospheric temperature and humidity. It is divided into Unit-A (AMSU-A) and Unit-B (AMSU-B), which provide a total of 15 channels and 5 observation channels, respectively. Twelve of

the AMSU-A channels are in the oxygen band frequency range from 50.3 to 57.3 GHz. The other three channels are located at 23.8, 31.4 and 89 GHz. AMSU-B has two window channels located at 89 and 150 GHz, along with three humidity sounding channels around the 183 GHz water vapor line. The field of view (FOV) size at nadir of AMSU-A is about 48 km. There are only 30 FOVs from each scan. AMSU-B, meanwhile, has an FOV of 16 km at nadir and a total of 90 FOVs at each beam position.

3.2 Case Description

A typhoon case is selected for the experiments. The cloud associated with the typhoon is systemic. More importantly, the accuracy of the simulated satellite radiance or brightness temperature in various channels is strongly affected by the surface emissivity. Compared to other surface types, the surface emissivity is less variable over the ocean and the microwave emissivity modeling over the sea surface is now considered as a well-handled issue. The selection obviously benefits the investigation.

Typhoon Luosha, numbered 0716, is the fifth tropical storm that has made landfall on the Chinese mainland in October since 1949, and was also the eighth landfalling typhoon in 2007. It initiated over the eastern Philippine Sea on 2 October 2007. Luosha moved northwestward and made its first landfall around 6 October near Yilan County, Taiwan. After landfall, it turned in a circular movement over the eastern Taiwan Strait before returning to make a second landfall in Yilan County. Luosha then moved northwestward across the Taiwan Strait and made a third landfall on the Chinese mainland around 7 October 2007.

3.3 Numerical Forecast Model

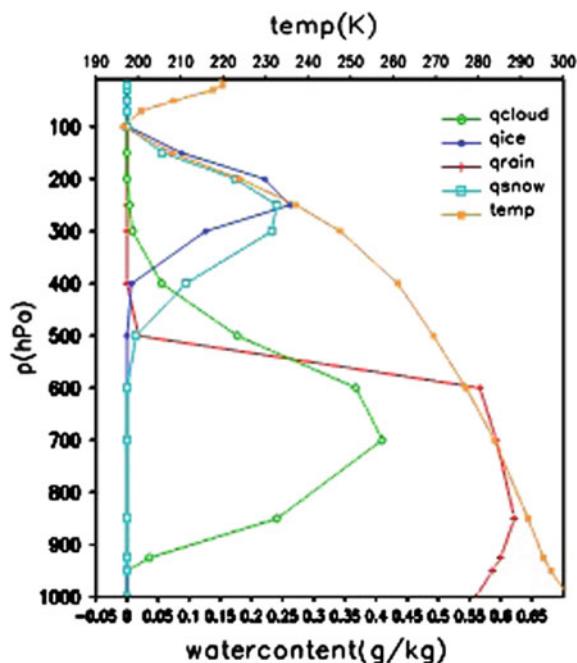
In this study, the Advanced Research WRF (ARW), version 3.3.1, developed at the National Center for Atmospheric Research (NCAR), USA, is employed. A double nested domain is configured. The parent domain is in the horizontal direction with a resolution of 54 km, while the two-way nesting domain has a resolution of 18 km. The Lin scheme is used for microphysical parameterization. This is a sophisticated scheme that has a 6-class hydrometeor setup, including water vapor, cloud liquid water, rain water, ice, snow and graupel, suitable for real-data high-resolution simulation. The other schemes used for physical processes are the Betts–Miller–Janjic cumulus parameterization scheme, the Yonsei University planetary boundary layer scheme, RRTM longwave radiation scheme, and the Dudhia shortwave radiation scheme (Wang et al. 2010). The WRF model is integrated up to 24 h, starting from 1800 UTC 2 October 2007, with National Centers for Environmental Prediction (NCEP) reanalysis data used as the initial and boundary conditions.

3.4 Inputs to the RRTM

The output of the numerical model's inner domain forecast with 18 km horizontal resolution is taken as the input to the RRTM. It includes not only atmospheric temperature, water vapor and pressure at user-defined layers, wind at 10 m above-ground, surface temperature and pressure, but also hydrometeor profiles of five cloud types, to evoke the inclusion of the multiple scattering of atmospheric particles in the RRTM. The five hydrometeor types are cloud water, rain water, ice, snow and graupel. Figure 1 shows the vertical profiles averaged for the hydrometeors in the typhoon area (18° – 24° N, 125° – 131° E). The cloud water and rain water are situated mainly below 500 hPa, with temperature at around 273 K. A small amount of cloud water overruns the zero degree line and is situated up to 300 hPa. Ice, snow and graupel water concentrate at the higher levels (Graupel is not shown). The altitude of ice is the highest. Snow is located below ice, and graupel water is the lowest with a much larger extent from 600 to 300 hPa.

For CRTM, the effective radius and variety of hydrometeor particle sizes are required, besides the inputs mentioned above. Taking account of the lack of adequate and authentic observations of cloud microphysical data, the effective radius of particle size is generally set to a constant. In this study, the effective radius of cloud liquid water, rain water, ice, snow and graupel are 15, 200, 200, 600 and 600 μ m, respectively.

Fig. 1 Vertical profiles of hydrometeors averaged in the typhoon area (18° – 24° N, 125° – 131° E)



For RTTOV, there is a need for cloud cover, and not the effective radius and variety of hydrometeor particle sizes. This is retrieved from the relative humidity in this study. The experiential formula between cloud cover N and relative humidity is:

$$N = \left(\frac{f - f_0}{1.0 - f_0} \right)^b,$$

where the range of N is 0–1.0; f is relative humidity; and f_0 is the critical value of relative humidity, which is dependent on the altitude. Here, this value is taken as 0.9 below 600 m, 0.5 for 600–1500 m, 0.6 for 1500–2500 m, and 0.5 when the altitude is above 2500 m. b is an experiential constant that is generally equal to 2.

4 Analysis of the Simulation of Cloudy and Rainy Satellite Microwave Observations

4.1 Influence of Hydrometeors on the Satellite Microwave Simulated Brightness Temperature

The simulated brightness temperature with and without hydrometeor profiles, using CRTM, as well as the satellite observations, are presented in Fig. 2. A five-class hydrometeor range—including cloud, rain, snow, ice and graupel—is used in the cloudy and rainy simulations. For brevity, only the results for AMSU-A channel 1 at 23.8 GHz and AMSU-B channel 1 at 89 GHz are given. As can be seen, large deviation exist between the simulated brightness temperature without hydrometeors and the observation. The inclusion of the radiative effect of hydrometeors in the RRTM leads to a notable matching up of the satellite brightness temperature simulation with the observation.

4.2 Deviation and Root-Mean-Squared Error

The area average of deviation and root-mean-squared error (RMSE) between the simulated brightness temperatures with and without hydrometeors are presented in Fig. 3. In terms of the channel numbers shown on the horizontal axis, numbers 1–15 are the 15 AMSU-A channels, while 16–20 correspond to AMSU-B channels 1–5. The legend shows the results for the different hydrometeors, in which ‘Ice*10’ implies that the result of ice is multiplied by 10, as the original magnitude is too small, and ‘4wat’ and ‘5wat’ represent the results of four-class and five-class hydrometers, respectively. The four-class result, which excludes graupel, is provided for convenience of comparison between CRTM and RTTOV in Sect. 6,

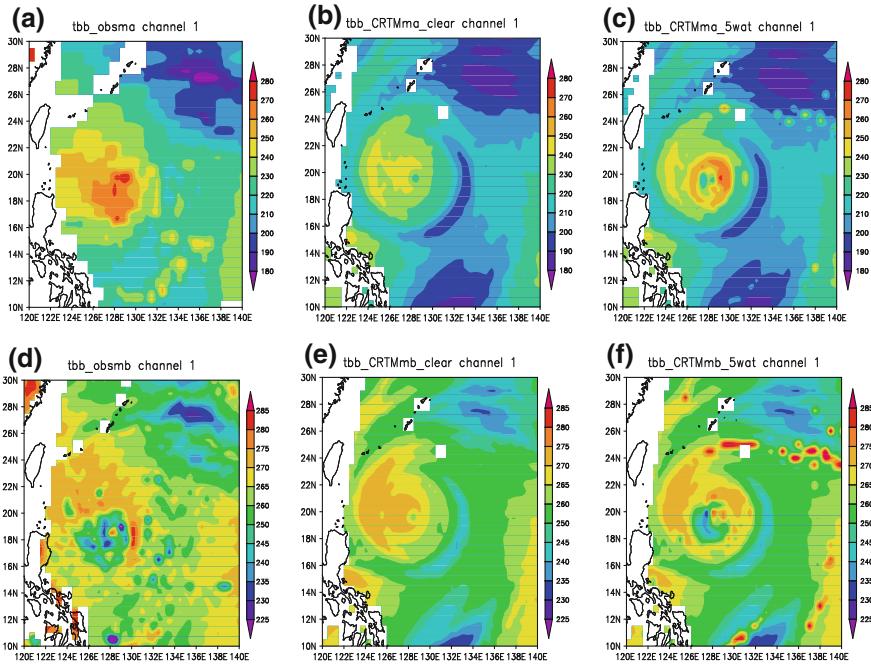


Fig. 2 Observed and CRTM-simulated brightness temperature for **a–c** AMSU-A channel 1 and **d–f** AMSU-B channel 1. **a, d** observations; **b, e** simulations without hydrometeors; **c, f** simulations with hydrometeors

because graupel is not currently included in RTTOV. The difference between the ‘5wat’ and ‘4wat’ results implies the effect of graupel.

In terms of the deviation, it can be seen that the radiative effect of hydrometeors has a diverse influence on the simulation of AMSU-A and AMSU-B brightness temperature, apart from in the AMSU-A high-level channels 10–14. The total radiative effect increases the brightness temperature for the AMSU-A channels 1–3 and decreases the temperature for the other AMSU-A channels and all the AMSU-B channels, both in the ‘5wat’ and ‘4wat’ results. The increment for channel 2 is the largest. Channel 3 yields the minimum warming from the cloudy and rainy particles. The cooling effect reduces gradually from AMSU-A channel 4 to 9. The decrement of brightness temperature becomes larger again in the AMSU-A window channel 15, and the two AMSU-B window channels, 16 and 17. The radiative effect of hydrometeors strengthens with the channel number in the AMSU-B sounding channels, from 18 to 20. All window channels—including AMSU-A 1–3, 15 and AMSU-B 16, 17—have large RMSE between the simulation with and without hydrometeors. The largest RMSE exceeds almost 20 K in channel 2. There are also several degrees of RMSE in simulated brightness temperature brought by the radiative effect of hydrometeors in both the AMSU-A low-level channels 4–8 and all AMSU-B humidity sounding channels. This corresponds well with the cloud

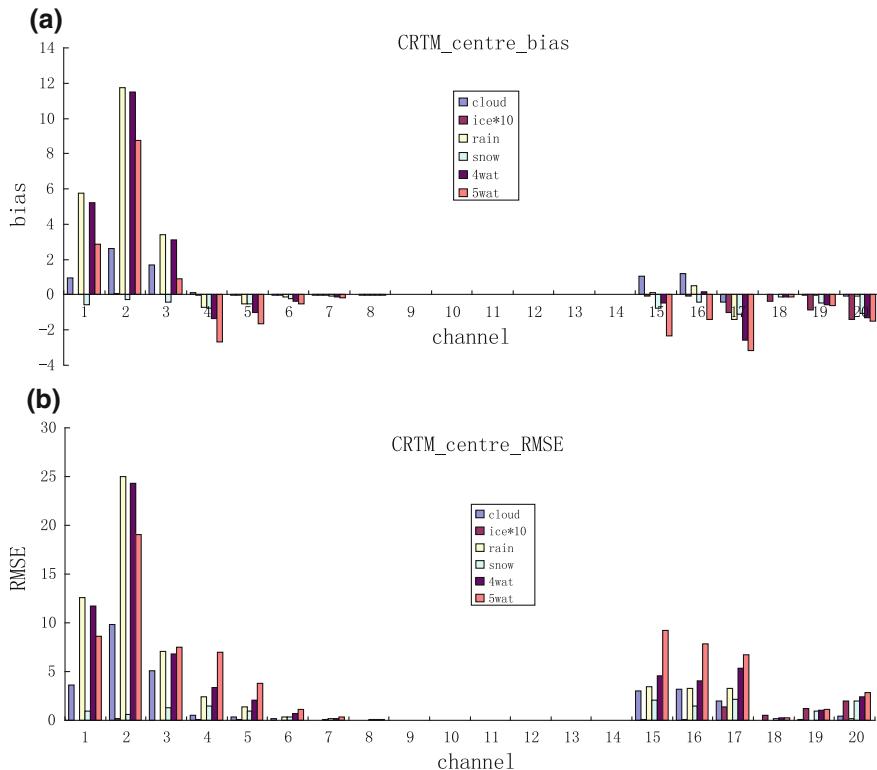


Fig. 3 Area average of simulated brightness temperature **a** deviation and **b** RMSE between the simulation with and without hydrometeors for AMSU-A and AMSU-B

detection performed in these channels in present data assimilation systems, in that the satellite observations of these channels are indeed affected by cloud and precipitation and must be removed in clear-sky satellite data assimilation.

Given that the focus is on the radiative effect of individual hydrometeors, it is found that the cloud and rain water play their roles mainly in the window and low-level channels, because cloud and rain water concentrate in the low-level atmosphere. The radiative emission of cloudy and rainy particles in the microwave low frequency increases the brightness temperature, and scattering becomes prevalent with the increase of frequency. Although there is still a warming of cloud and rain water in AMSU-A channel 15 and AMSU-B channel 16, the magnitude is small. The radiative effect of cloud and rain is one of cooling in AMSU-B window channel 17. In AMSU-A channels 4–8 and the AMSU-B humidity sounding channels 18–20, it is mainly rain that depresses the simulated brightness temperature. The higher the corresponding altitude of the channel, the smaller the magnitude, which is in accordance with the fewer cloud and rain water particles in high-level layers. This is also true for the AMSU-B sounding channels, as the

channel number is ordered opposite with the corresponding altitude. Ice and snow reduce the simulated brightness temperature. The variation produced by ice is extremely small and is obvious in AMSU-B channels, with a tenth of several degrees. The cooling effect of snow is notable not only in AMSU-B, but also AMSU-A, always bringing a reduction of about 1–2 K. In addition, the radiative effect of graupel can be determined from the difference between the ‘5wat’ and ‘4wat’ results. Specifically, it is the cooling in brightness temperature owing to scattering from graupel with large particle radius in all AMSU-A and AMSU-B channels. Graupel plays a more important role than ice and snow, and is even the key role among all hydrometeors in several channels.

4.3 Contribution Ratio

The percentage of brightness temperature variation contributed by each hydrometeor in deviation between the cloudy and clear simulation is shown in Fig. 4 to illustrate the comparative importance of the radiative effect of hydrometeors on the microwave satellite observation. The results of AMSU-A channels 10–14, in which hydrometeors play no role, are ignored.

It can be seen that the radiative effect is completely predominated by the cloud and rain water in AMSU-A channels 1–3. The total percentages of cloud and rain contribution exceed more than 98 % and 80 % in channels 1–2 and channel 3, respectively. The ratio of rain water is always larger than that of cloud water. From AMSU-A channels 4–9, the increase in the channel central frequency reduces the effect of cloud and rain water. The total contribution drops to below 40 % after channel 4 and fades completely in channel 9. The contributions of snow, ice and graupel all increase with channel number in AMSU-A channels 1–9, with the effect of snow increasing uniformly; it exceeds 50 % after channel 7 and reaches 84 % at channel 9. The effect of ice only emerges as being notable in the mid-level channels, 7–9. Graupel, meanwhile, takes up an important position from channel 3, making its maximum contribution (33 %) in channel 4. The contribution ratio of graupel decreases from 29 % to 8 % from channel 5 to 9.

In the AMSU-A window channel 15 and the five AMSU-B channels, it is graupel that makes the most important contribution. Its contribution is greater than 50 % in all three window channels and one humidity sounding channel: AMSU-A channel 15 and AMSU-B channels 16, 17 and 20. Snow is the second highest contributor in those channels and has the leading role in channels 18 and 19. The effects of cloud and rain water are concentrated in the window channels. The effect of ice is shown in three humidity sounding channels; the maximum is 13 % in channel 18.

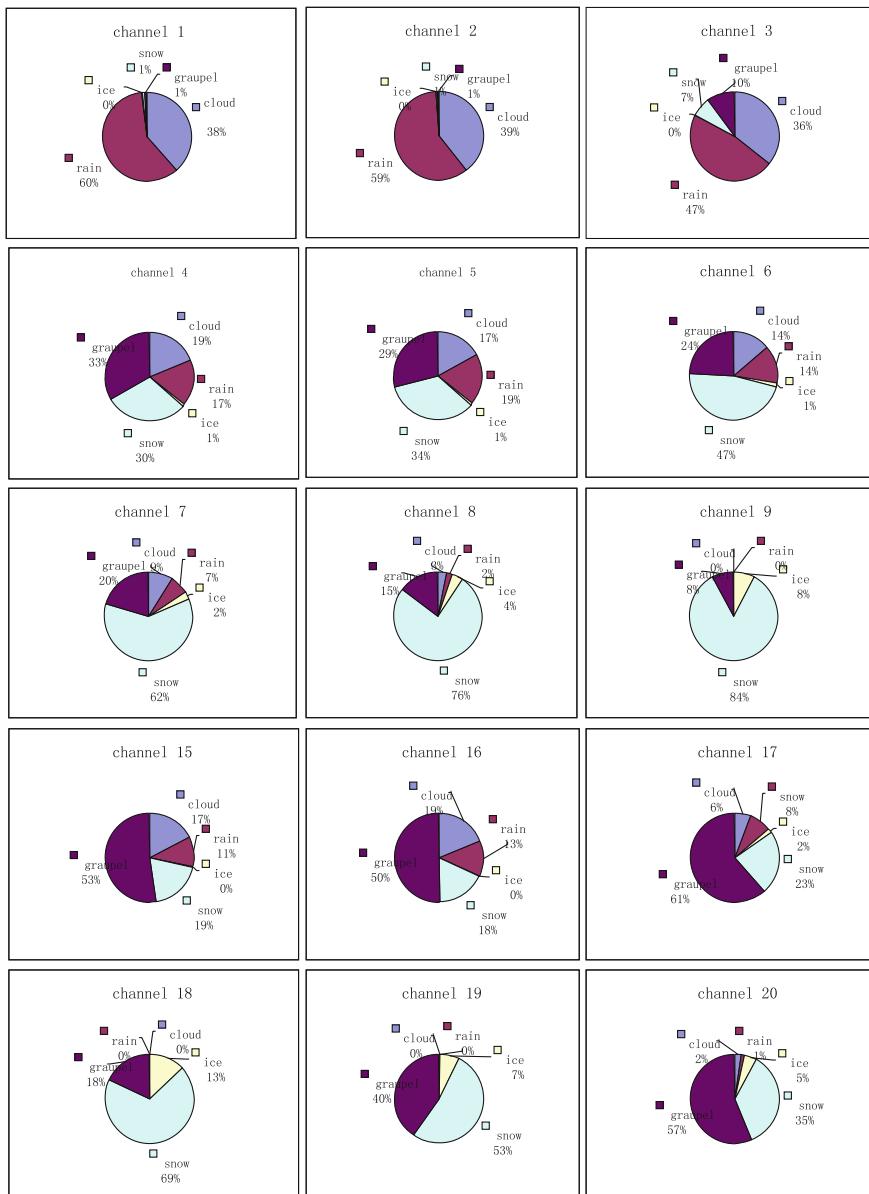


Fig. 4 Contribution ratio of each hydrometeor to the brightness temperature variation

5 Sensitivity of the Satellite Simulation to Hydrometeors' Properties

5.1 Sensitivity to Water Content

The sensitivity of satellite microwave brightness temperature to the water content of hydrometeors is verified by experiments in which the hydrometeors are each given a content increase or decrease of 10, 20 and 50 %.

The area-averaged deviation of brightness temperature produced by the changes in hydrometeor water content is illustrated in Fig. 5. As can be seen, the brightness temperatures of the AMSU-A window channels 1–3, 15 and the AMSU-B window channels 16–17, are the most sensitive to the content of cloud and rain water. The channel with the largest sensitivity is channel 2. The sensitivity of ice is apparent only in AMSU-B channels. It is worth pointing out that the temperature deviation is extraordinarily large when the content of ice is increased by 50 %. An explanation may be the insufficient ice water in the numerical forecast. Thus, the radiative effect of ice is likely underestimated in this study (Hong et al. 2010), and there is a need to revisit this later. Snow and graupel have strong influences on the simulated brightness temperature of the AMSU-A low-level channels, window channel 15, and all AMSU-B channels—especially those channels with high frequencies. The sensitivity of satellite microwave remote sensing to the water content corresponds well with the radiative effects of hydrometeors revealed above.

5.2 Sensitivity to Particle Size

In a similar way to water content in Sect. 5.1, the sensitivity to particle size is discussed using the results of experiments in which the effective radii of hydrometeors are each increased or decreased by 10, 20 and 50 %. The area-averaged deviation of brightness temperature is shown in Fig. 6.

There is no variation in brightness temperature produced by the changes in cloud and ice water particle size. The radiative effect of rain is mainly one of warming, owing to the emission process, in the AMSU-A window channels 1–3, 15 and AMSU-B window channel 16; while in other channels, the effect is one of cooling associated with scattering. However, enlargement of the effective radius only increases the temperature in AMSU-A channels 1–2. The brightness temperatures in AMSU-A channels 3, 15 and AMSU-B 16 are depressed when the particle size becomes large. At the same time, the brightness temperatures of AMSU-A channels 1–3 are depressed, and those in other channels are increased, as the effective radius is reduced. This implies that emitting particles dominate only in the channels with

Fig. 5 Area-averaged deviation of brightness temperature produced by the changes in hydrometeor water content (indicated by the different colored bars; see legend at the bottom of the figure) for AMSU-A and AMSU-B

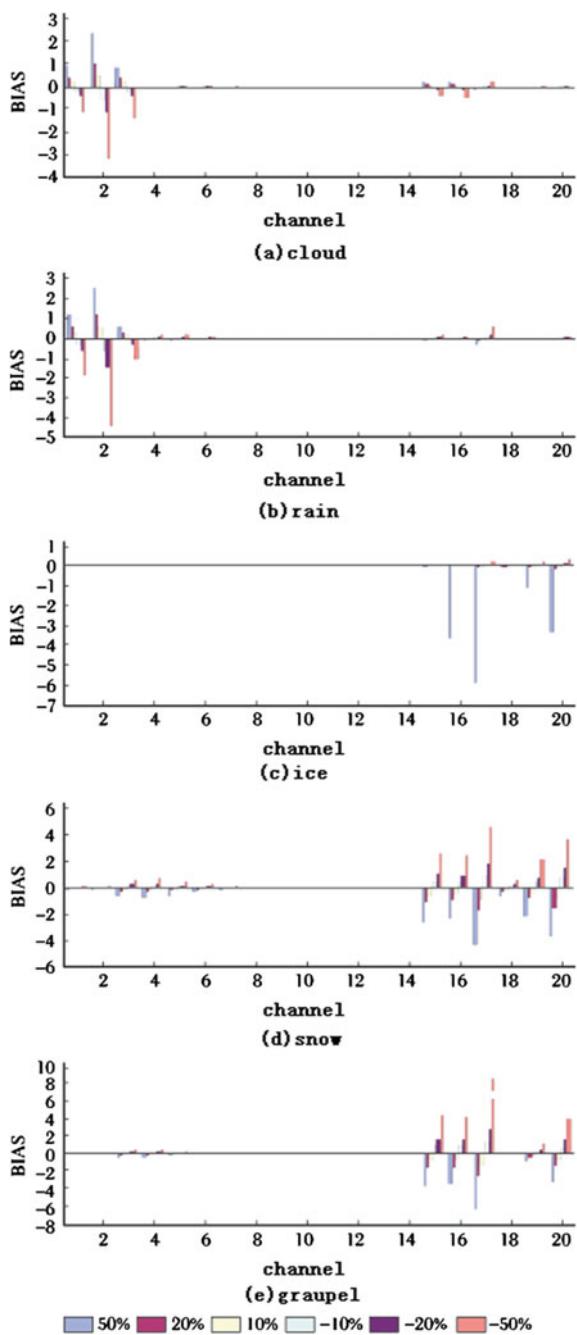
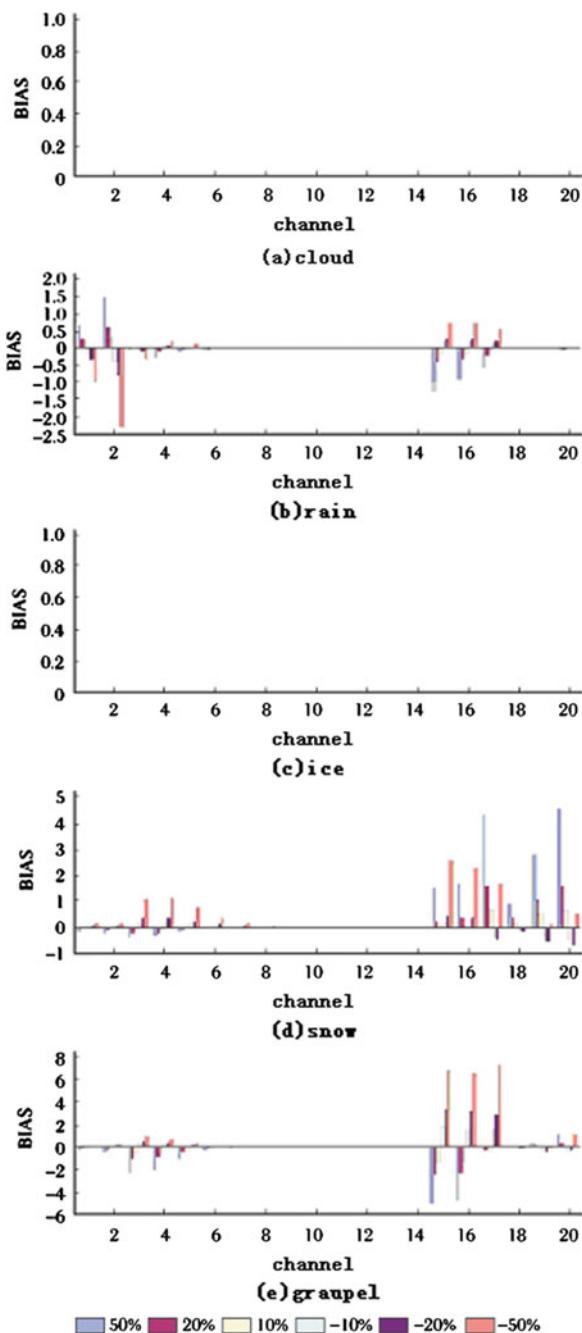


Fig. 6 Area-averaged deviation of brightness temperature produced by the changes in the effective radii of hydrometeors (indicated by the different colored bars; see legend at the bottom of the figure) for AMSU-A and AMSU-B



low frequency, and scattering becomes prevalent with an increase in particle size. The cooling of scattering overruns the warming due to emission in those high-frequency channels.

The brightness temperatures of AMSU-A channels 1–7, 15 and all AMSU-B channels are sensitive to the particle size of snow and graupel water. The sensitivities of AMSU-A channel 15 and AMSU-B channels are much stronger. For snow, the temperature generally increases, regardless of whether the effective radius is enlarged or reduced. The temperature associated with graupel decreases with increasing particle size. On the contrary, the temperature becomes warmer when the particle size is reduced. Meanwhile, the sensitivity seems to be complicated by the frequency. This is attributed to the scattering associated not only with particle size, but also the frequency.

5.3 *Sensitivity to the Vertical Distribution of Hydrometeors*

To study the sensitivity of the satellite simulation to the vertical distribution of hydrometeors, the hydrometeor profiles are adjusted by moving them each up or down by one and two layers. The one and two layers correspond to 50 and 100 hPa, respectively, because the hydrometeor profiles are on the even interval pressure layer with 50 hPa. Figure 7 shows the area-averaged deviation of brightness temperature brought about by the adjustment of the hydrometeor profiles.

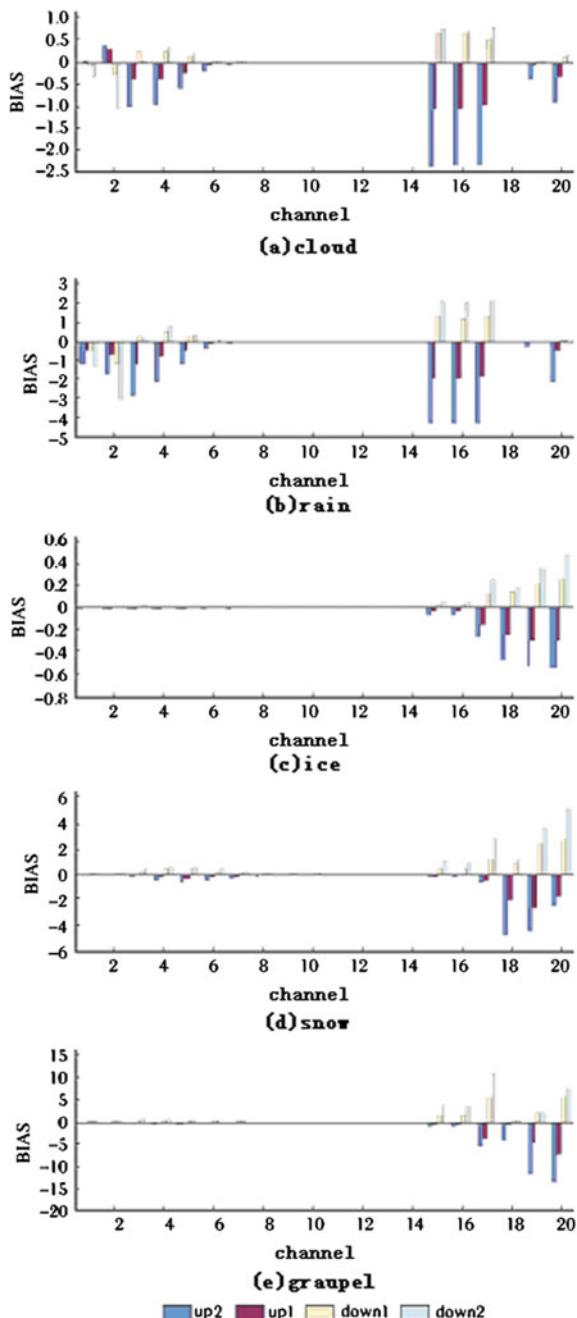
It can be seen that the sensitivity of the satellite simulation to the vertical distribution of hydrometeors is not only apparent in the change of brightness temperature in those channels affected by cloud and precipitation, but also by the transfer of the particular channel affected. It is understandable that the brightness temperature in the channel associated with its response function at a certain altitude is greatly affected by the change of hydrometeors in this layer.

6 Inter-Comparison Between RTTOV and CRTM

6.1 *Simulated Satellite Brightness Temperature*

Figure 8 presents the simulated brightness temperature with and without hydrometeor profiles by using RTTOV. The simulation by CRTM with its four-class hydrometeor setup, i.e., cloud, rain, snow and ice, the same as that of RTTOV, is presented. Compared to Fig. 2, the RTTOV clear-sky simulations agree well with those of CRTM without hydrometeors. It has large deviation with the observation and cannot capture the observed brightness temperature. Simultaneously, the simulation is improved greatly by the consideration of the radiative effect of hydrometeors. This makes both RTTOV and CRTM a practical tool under cloudy

Fig. 7 Area-averaged deviation of brightness temperature brought about by the adjustment of hydrometeor profiles (indicated by the different colored bars; see legend at the bottom of the figure) for AMSU-A and AMSU-B



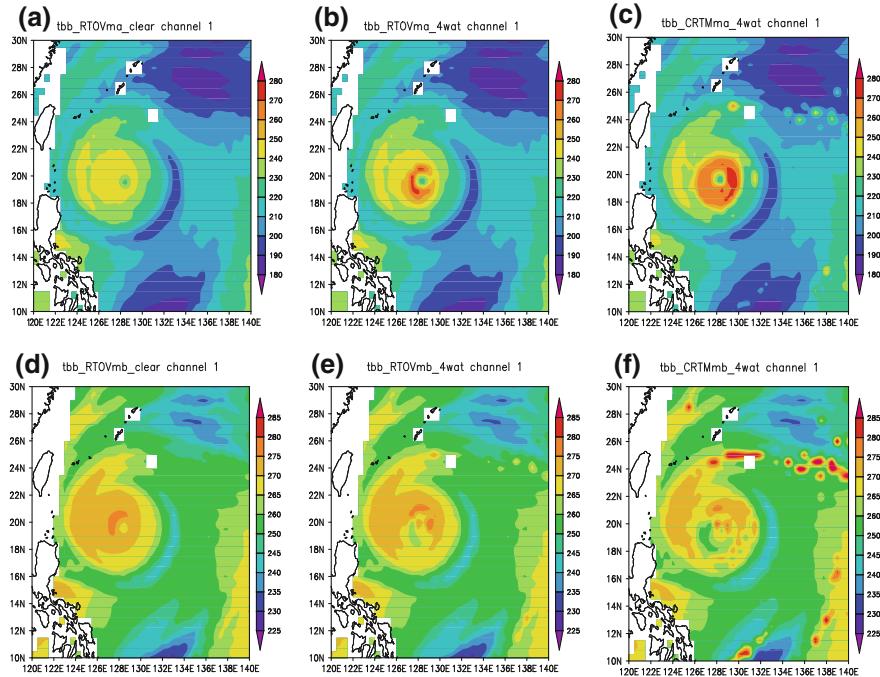


Fig. 8 RTTOV- and CRTM-simulated brightness temperature for **a–c** AMSU-A and **d–f** AMSU-B channel 1. **a, d** RTTOV simulations without hydrometeors; **b, e** RTTOV simulations with hydrometeors; **c, f** four-class hydrometeor CRTM simulations

and rainy conditions, through the capability of incorporating absorption and scattering from hydrometeors. It should be noted, however, that the depression of RTTOV-simulated brightness temperature in AMSU-B channel 1 is not as evident as in the CRTM simulation with its five-class hydrometeor range, while it is almost the same as the simulation of CRTM with only four classes of hydrometeors. This demonstrates that it is the cooling effect of graupel particles that is missing in the present RTTOV. In general, the five-class hydrometeor (with graupel) CRTM simulation is the closest match to the observation.

6.2 Deviation and RMSE

The RTTOV results for the area-average deviation and RMSE between the simulated brightness temperatures with hydrometeors and those without hydrometeors are shown in Fig. 9. In general, the radiative effect of hydrometeors represented in

RTTOV agrees well with that of CRTM. Only the AMSU-A high-level channels 10–14 are not affected by cloud and precipitation. The brightness temperature is increased in AMSU-A channels 1–3 and decreased in other channels. The warming effect, owing to the radiative emission of cloud and rain water dominates in three of the AMSU-A window channels, 1–3, but is weakened in both the other AMSU-A window channel, 15, and the AMSU-B window channel 16. The effect of rain water is already one of cooling in AMSU-A channel 15 and AMSU-B channel 16. A subtle difference is found between RTTOV and CRTM. The effect of cloud and rain water in other channels is one of scattering, which decreases the temperature. The scattering of both ice and snow leads to a reduction in temperature.

However, a significant discrepancy exists between the results of the two RRTMs. That is, the magnitude of the deviation produced by the radiative effect of hydrometeors: CRTM appears to be at least double the magnitude of RTTOV.

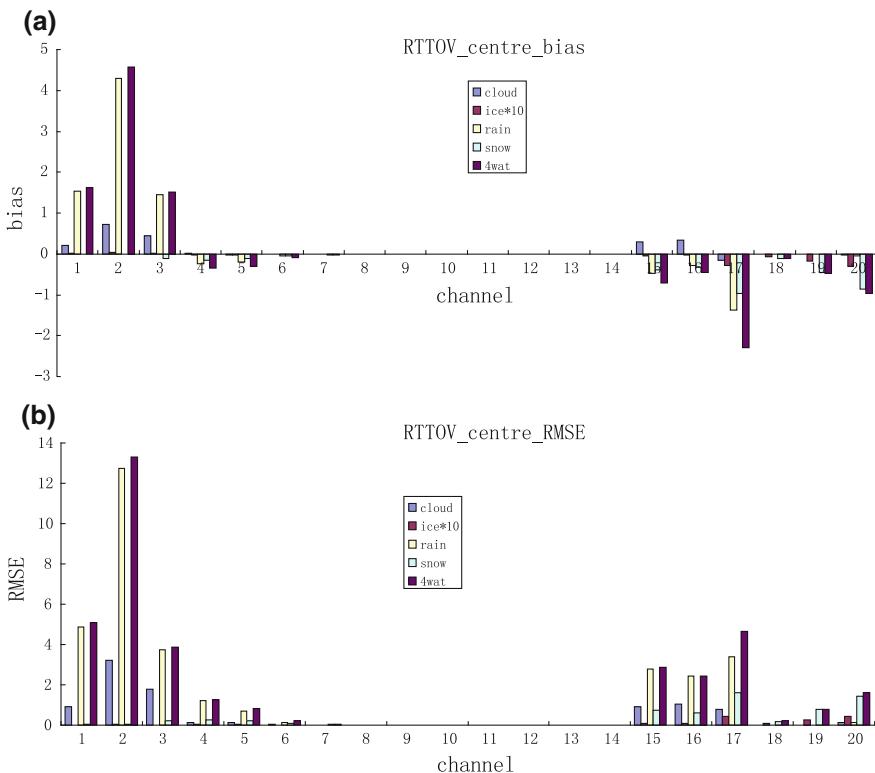


Fig. 9 RTTOV area-averaged simulated brightness temperature **a** deviation and **b** RMSE between the simulation with and without hydrometeors for AMSU-A and AMSU-B

6.3 Response Function of Hydrometeors

As mentioned above, the derivative of the satellite brightness temperature with respect to the state variables obtained by the adjoint model of the RRTM is used in data assimilation to generate the minimum cost function. It is also the response function that expresses the sensitivity of the brightness temperature to the input variables. Figures 10 and 11 are the response functions of four cloud-type hydrometeors—cloud, rain, ice and snow—produced in both RTTOV and CRTM.

The response functions of cloud water and rain water are situated in the low level and those of ice and snow are located in the high level, corresponding well with the vertical distribution of hydrometeors. Of note is that there is also a peak around 300 hPa for cloud water. This should contribute less to the variation of brightness temperature, taking into account the limited cloud water content in the high-level atmosphere. The response functions of cloud and rain water for the AMSU-A window channels and AMSU-B window channel 1 are positive, expressing the radiative emission of cloud and rain particles and increasing the brightness temperature. The response functions, especially that of rain water, become negative for the AMSU-B humidity sounding channels. This is consistent with the temperature depression in these channels, which suggests that the radiative effect has turned to one of cooling, owing to scattering. The negative response functions of ice and snow show that the radiative effect is one of scattering, and there is a depression in the brightness temperature. Moreover, it is illustrated by the order of magnitude that

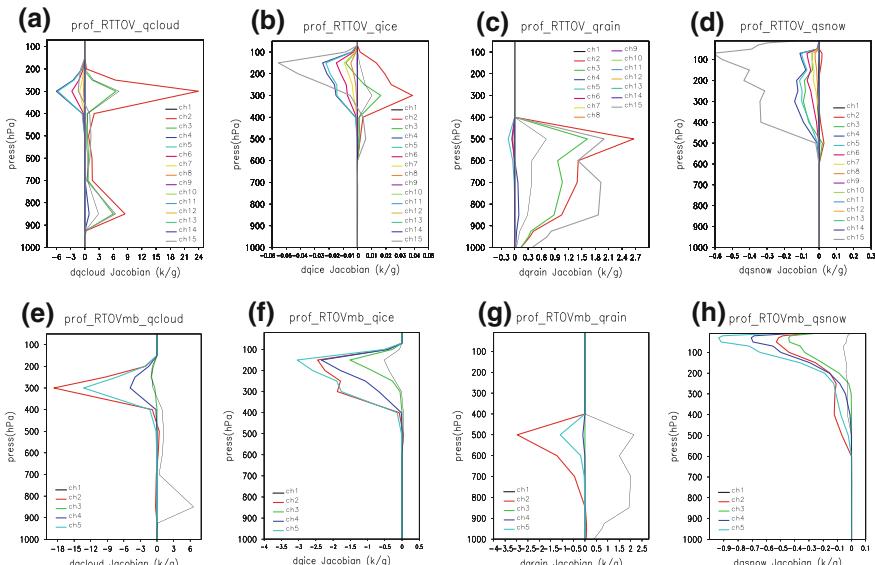


Fig. 10 Response functions of hydrometeors obtained from the RTTOV Jacobian model. **a-d** AMSU-A; **e-h** AMSU-B

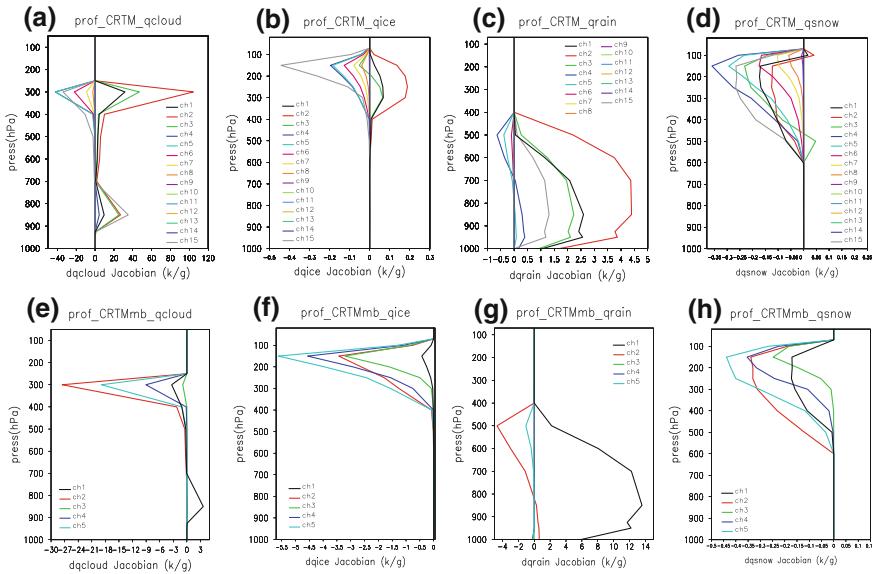


Fig. 11 Response functions of hydrometeors obtained from the CRTM Jacobian model. **a-d** AMSU-A; **e-h** AMSU-B

the radiative effect of rain is the strongest, followed by cloud and snow. Ice has the weakest radiative effect among these four classes of hydrometeors.

It is important to note at this point that it is still the magnitude of the response function that causes the level of discrepancy between the results of RTTOV and CRTM. The stronger the response function is, the larger the deviation of temperature is produced. This is in accordance with the previous statistical results, i.e., the deviation and RMSE between the simulated brightness temperature with and without hydrometeors.

7 Summary and Discussion

In this paper, the radiative effect of hydrometeors on the simulation of microwave satellite observations in the RRTM is evaluated by taking the hydrometeor output from a WRF numerical forecast as the input. The evaluation includes analysis of the simulation of cloudy and rainy satellite microwave observations and an investigation of the sensitivity of the satellite simulation to hydrometeor properties, including the water content, droplet size and vertical distribution. An inter-comparison between the results of two popular RRTMs within the NWP community, CRTM and RTTOV, is also carried out.

The inclusion of the radiative effect of hydrometeors in the RRTM leads to a notable matching up between the satellite brightness temperature simulation and observations. The development of RRTMs with the capability of incorporating emitting and scattering processes from hydrometeors is a necessary development so as to supply assimilations of satellite observations affected by cloud and precipitation.

The radiative effects of hydrometeors have diverse influences on most channels of satellite microwave observations, except the AMSU-A high-level temperature sounding channels 10–14. The validation of the effect of hydrometeors in the RRTM corresponds well with the cloud detection performed in these channels in present data assimilation systems, in that the satellite observations of these channels are indeed affected by cloud and precipitation and must be removed in clear-sky satellite data assimilation.

Cloud and rain water mainly have a warming effect owing to their radiative emission. This effect dominates in three of the AMSU-A window channels, 1–3, but is weakened both in the other AMSU-A window channel, 15, and the AMSU-B window channel 16. The effect of cloud and rain water in other channels is one of scattering, which decreases the brightness temperature. Ice, snow and graupel all present a cooling effect, owing to scattering. The variation produced by ice is extremely small, but is obvious in AMSU-B. The effect of both snow and graupel is notable in all AMSU-A and AMSU-B channels.

The sensitivity of satellite microwave remote sensing to the water content corresponds well with the radiative effect of hydrometeors. The brightness temperature is not sensitive to effective radius size of cloud and ice, and the sensitivity of satellite observations is strong to the particle size of rain, snow and graupel. Meanwhile, the sensitivity becomes complicated by the frequency. The sensitivity of the satellite simulation to the vertical distribution of hydrometeors is presented by the transfer of the particular channel affected.

The radiative effect of hydrometeors is generally consistent in RTTOV and CRTM. The greatest discrepancy is the magnitude of the response function of hydrometeors and the corresponding deviation of brightness temperature produced by the radiative effect of hydrometeors. The results in CRTM appear to be at least double the magnitude of those in RTTOV.

It could be argued that the current assessment is limited by the fact that the input is model-predicted cloud condensate, which is unlikely to be accurate. It is undeniable that there are errors, possibly even large ones, in the numerical forecast, especially for the cloud microphysical variables. However, the study nevertheless provides a feasible approach to understanding the simulation of RRTMs under cloudy conditions, given the lack of accurate and sufficiently detailed cloud microphysical observations. Moreover, quantitative evaluations of the errors arising from RRTMs with a numerical forecast background are very important because the errors are typically used to define the observational error covariance matrices in direct assimilations of satellite radiance. Of course, more observational datasets (e.g. satellite active sensors, CloudSat, measurements from aircraft) will be used in future work to better characterize the hydrometeors predicted by the numerical model.

The performances of RRTMs under scattering cloudy and rainy atmospheres may vary substantially. Scattering by cloud and precipitation is a complicated process involving many hydrometeor-related factors. On the one hand, satellite simulations are sensitive to hydrometeors' properties; while on the other hand, these variables are hard to obtain precisely, and are sometimes even missing completely. Having a fixed particle size in this study is a definite limitation. The determination of particle size from diagnostic or alternative prognostic schemes in numerical models can be tested later.

In addition, it should also be noted that the channels in which the hydrometeors have their greatest radiative effects are the window channels, as well as the lower sounding channels. The measurements from these channels respond sensitively to the radiation emanating from the Earth's surface, such that the simulation errors are strongly dependent on the accuracy of the surface emissivity. Thus, there is also a need to properly handle the variability of surface emissivity in the use of satellite data affected by cloud and precipitation, especially over complex surface types such as land, snow and ice, whose surface emissivity varies significantly.

Acknowledgements This study was supported by funding from the Natural Science Foundation of China Project 41675027 and 41675108, the International S&T Cooperation Program of China (ISTCP) under Grant no. 2011DFG23210 and the National High Technology Research and Development Program of China (863 Program) under Grant No. 2013AA050601.

References

- Bauer P, Moreau E, Chevallier F et al (2006a) Multiple-scattering microwave radiative transfer for data assimilation applications. *Q J R Meteorol Soc* 132:1259–1281
- Bauer P, Philippe L, Angela B et al (2006b) Implementation of 1D + 4D-Var assimilation of precipitation-affected microwave radiances at ECMWF. I: 1D-Var. *Q J R Meteorol Soc* 132:2277–2306
- Bauer P, Geer AJ, Lopez P, Salmond D (2010) Direct 4D-Var assimilation of all-sky radiances. Part I: implementation. *Q J R Meteorol Soc* 136:1868–1885
- Bauer P, Auligné T, Bell W, Geer A, Guidard V, Heilliette S, Kazumori M, Kim MJ, Liu EH-C, McNally AP, Macpherson B, Okamoto K, Renshaw R, Riishøjgaard LP (2011) Satellite cloud and precipitation assimilation at operational NWP centres. *Q J R Meteorol Soc* 137:1934–1951
- Bennartz R, Bauer P (2003) Sensitivity of microwave radiances at 85–183 GHz to precipitation ice particles. *Radio sci* 38. doi:[10.1029/2002RS002626](https://doi.org/10.1029/2002RS002626)
- Burns BA, Wu X, Diak GR (1997) Effects of precipitation and cloud ice on brightness temperatures in AMSU moisture channels. *IEEE Trans Geosci Remote Sens* 35:1429–1437
- Chen Y, Weng F, Han Y, Liu Q (2008) Validation of the community radiative transfer model (CRTM) by using cloudsat data. *J Geophys Res* 113, doi:[10.1029/2007JD009561](https://doi.org/10.1029/2007JD009561)
- Chevallier F, Bauer P (2003) Model rain and clouds over oceans: Comparison with SSIM/I observation. *Mon Weather Rev* 131(10):1242–1250
- Errico RM, Bauer P, Mahfouf JF (2007) Issues regarding the assimilation of cloud and precipitation data. *J Atmos Sci* 64:3785–3858
- Geer AJ, Bauer P, Lopez P (2010) Direct 4D-Var assimilation of all-sky radiances. Part II: assessment. *Q J R Meteorol Soc* 136:1886–1905
- Geer AJ, Bauer P (2011) Observation errors in all-sky data assimilation. *Q J R Meteorol Soc* 137:2024–2037

- Han W, Dong P (2012) Study and comparison of simulation of satellite microwave observations in cloudy and rainy areas using RTTOV and CRTM. In: Proceeding of the 18th international TOVS study conference. <http://cimss.ssec.wisc.edu/itwg/itsc/itsc13/>
- Han Y, van Delst P, Liu Q, et al (2012) User's guide to the JCSDA community radiative transfer model. <http://www.emc.ncep.noaa.gov/jcsda/CRTM/>
- Hong G, Heygester G, Miao G, Kunzi K (2005) Sensitivity of microwave brightness temperatures to hydrometeors in a tropical deep convective cloud system at 89–190 GHz. *Radio sci* 40. doi:10.1029/2004RS003129
- Hong G, Yang P, Weng F, Liu M (2010) Simulations of microwave brightness temperatures at AMSU-B frequencies over a 3D convective cloud system. *Int J Remote Sens* 31:1781–1800
- James H, Peter R, David R, et al (2014) RTTOV v11 Users Guide. NWPSAF-MO-UD-028
- Kummerow C (1993) On the accuracy of the Eddington approximation for radiative transfer in the microwave frequencies. *J Geophys Res* 98:2757–2765
- Mahfouf JF, Beljaars A, Chevallier F et al (1999) The importance of the earth radiation mission for numerical weather prediction. ECMWF Tech Memorandum 288
- McNally AP (2002) A note on the occurrence of cloud in meteorologically sensitive areas and the implications for advances infrared sounders. *Q J R Meteorol Soc* 128:2551–2556
- McNally AP (2009) The direct assimilation of cloud-affected satellite infrared radiances in the ECMWF 4D-Var. *Q J R Meteorol Soc* 135:1214–1229
- Okamoto K, McNally AP, Bell W (2014) Progress towards the assimilation of all-sky infrared radiances: an evaluation of cloud effects. *Q J R Meteorol Soc* 140:1603–1614
- Pu ZX (2009) Assimilation of satellite data in improving numerical simulation of tropical cyclones: progress, challenge and development. Data assimilation for atmospheric, oceanic and hydrologic applications (Vol I). Springer, Berlin, pp 163–176
- Rodgers CD (2000) Inverse methods for atmospheric sounding: theory and practice. World Scientific Publishing Company, pp 200
- Saunders RW, Matricardi M, Brunel P (1999) An improved fast radiative transfer model for assimilation of satellite radiance observation. *Q J R Meteorol Soc* 125:1407–1425
- Skofronick-Jackson GM, Gasiewski AJ, Wang JR (2002) Influence of microphysical cloud parameterizations on brightness brightness temperature. *IEEE Trans Geosci Remote Sens* 40:187–196
- Sreerekha TR, Buehler SA, O'keeffe U et al (2008) A strong ice cloud event as seen by a microwave satellite sensor: simulations and observations. *J Quant Spectrosc Radiat Transfer* 109:1705–1718
- Stengel M, Linskog M, Undén P, Gustafsson N, Bennartz R (2010) An extended observation operator in HIRLAM 4D-VAR for the assimilation of cloud-affected satellite radiances. *Q J R Meteorol Soc* 136:1064–1074
- Wang W, Bruyere C, Duha M et al (2010) WRF-ARW version 3 modeling system User's guide. <http://www.mmm.ucar.edu/wrf/users/docs/>
- Weng F, Liu Q (2003) Satellite data assimilation in numerical weather prediction models. Part I: forward radiative transfer and jacobian modeling in cloudy atmospheres. *J Atmos Sci* 60:2633–2646
- Weng F (2007) Advances in radiative transfer modeling in support of satellite data assimilation. *J Atmos Sci* 64:3799–3807
- Županski M (2013) All-Sky satellite radiance data assimilation: methodology and challenges, data assimilation for atmospheric, oceanic and hydrologic applications (Vol II). Springer, Berlin, pp 465–488

Toward New Applications of the Adjoint Sensitivity Tools in Data Assimilation

Dacian N. Daescu and Rolf H. Langland

Abstract Novel applications of the adjoint-based sensitivity tools are investigated to obtain *a priori* guidance on the forecast impact of modeling correlated observational errors in a four-dimensional variational data assimilation system (4D-Var DAS). A synergistic framework is considered that combines *a posteriori* estimates to the observation error covariance (\mathbf{R}) and derivative information extracted from the adjoint-DAS forecast error \mathbf{R} -sensitivity (FSR). It is explained that the FSR approach allows the analysis of structured error correlation models and estimation of their potential impact on reducing the forecast errors. Theoretical aspects are discussed and a proof-of-concept is provided with Lorenz's 40-variable model. The practical ability to exercise these new adjoint capabilities is shown in experiments performed with the Naval Research Laboratory Atmospheric Variational Data Assimilation System-Accelerated Representer (NAVDAS-AR) and the Navy's Global Environmental Model (NAVGEM). In particular, the FSR analysis of radiances assimilated from the Infrared Atmospheric Sounding Interferometer (IASI) indicates that modeling inter-channel observation error correlations may provide an increased benefit to the forecasts, as compared with tuning procedures that ignore the error correlations and only adjust the assigned observation error variance parameters.

1 Introduction

The information content of observations ingested into atmospheric data assimilation systems (DASs) is closely determined by the representation of the statistical properties of the errors in the prior state estimate (background) and observations. Improving the specification of the background error covariance (\mathbf{B} -matrix) is a continu-

D.N. Daescu (✉)

Portland State University, Portland, OR, USA

e-mail: daescu@pdx.edu

R.H. Langland

Marine Meteorology Division, Naval Research Laboratory, Monterey, CA, USA

e-mail: Rolf.Langland@nrlmry.navy.mil

ous research effort in the numerical weather prediction (NWP) community (Bouttier et al. 1997; Buehner 2005; Bannister 2008a, b; Brousseau et al. 2012; Clayton et al. 2013; Kuhl et al. 2013; Wang et al. 2013; Lorenc et al. 2015). The rapid growth in the data volume provided by remote sensing instruments (Thépaut and Andersson 2010; Lahoz 2010) has prompted research to estimate and improve the representation of the observation error covariance (\mathbf{R} -matrix) in the DAS.

Studies based on *a posteriori* statistical analysis of observed-minus-background (\mathbf{d}_b^o) and observed-minus-analysis (\mathbf{d}_a^o) departures (Desroziers et al. 2005; Garand et al. 2007; Bormann and Bauer 2010; Bormann et al. 2010; Stewart et al. 2014) indicate that both spatial and inter-channel error correlations are present in the radiances assimilated from hyperspectral sounding instruments such as the Atmospheric Infrared Sounder (AIRS) and the Infrared Atmospheric Sounding Interferometer (IASI). Error correlations of increased magnitude for microwave imager radiances have been estimated in the work of Bormann et al. (2011). Spatial and spectral thinning, superobbing, and error variance inflation are standard procedures implemented at NWP centers to address the information redundancy and to alleviate the impact of correlated observation errors in high-density satellite data (Goldberg et al. 2003; Pauley 2003; McNally et al. 2006; Collard 2007). An open research question is whether modeling correlated observation errors in the DAS will entail substantial gains in the forecast skill. Stewart et al. (2013) illustrate the potential benefits of incorporating error correlation structures (\mathbf{C}^o -matrix) in the \mathbf{R} -matrix specification. For operational implementation, promising results have been put forward by Weston et al. (2014) who used an estimate derived from *a posteriori* consistency diagnosis to specify inter-channel error correlations to IASI radiances assimilated in the Met Office four-dimensional variational data assimilation system (4D-Var DAS). Weston et al. (2014) also explained that various simplifications and further adjustments (such as reconditioning) are necessary to overcome computational issues that arise from the use of a non-diagonal \mathbf{R} matrix e.g., the increased condition number of the Hessian matrix in the 4D-Var minimization problem.

Trial-and-error experimentation to design new correlation models that increase the information content of observations requires significant computational resources and software development efforts. The capability to identify high-impact error correlation structures and assess the potential gain in the model forecast skill *prior to* the actual implementation in the DAS may provide valuable insight for developing error covariance models that are effective in reducing the forecast errors. Theoretical aspects of adjoint-based estimation to forecast error sensitivity in nonlinear variational data assimilation are discussed by Daescu (2008) and Daescu and Navon (2013). The practical ability to analyze various DAS-input components was shown by Daescu and Todling (2010) with the Grid-point Statistical Interpolation (GSI) analysis implemented in NASA Goddard Earth Observing System (GEOS). Daescu and Langland (2013a, b) provided a sensitivity analysis and a priori forecast impact estimates of observation error variance (σ_o^2) parameters in a 4D-Var DAS, the Naval Research Laboratory Atmospheric Variational Data Assimilation System-Accelerated Representer (NAVDAS-AR, Xu et al. 2005; Rosmond and Xu 2006). The study of Lupu et al. (2015) for tuning IASI error variances indicates that an

improved performance may be achieved from a tuning approach that uses *a posteriori* diagnosis to obtain new estimates $\tilde{\sigma}_o^2$ and relies on the adjoint-based σ_o^2 -sensitivity guidance for instrument channel selection.

Our work brings forward the practical ability to obtain guidance on the performance of an observation error covariance model through the adjoint-DAS forecast \mathbf{R} -sensitivity (FSR) and impact assessment. The chapter is organized as follows. Section 2 outlines the diagnosis of the error covariance based on observation residuals and provides insight on its limitations for covariance tuning. Section 3 includes a review of the FSR approach and details the methodology to estimate the forecast impact of a structured model to observation error correlations. A proof-of-concept is given with Lorenz's 40-variable model. Section 4 presents results with the adjoint versions of NAVDAS-AR and Navy's Global Environmental Model (NAVGEN, Hogan et al. 2014). Diagnosis of inter-channel error correlations, FSR-based sensitivity, and *a priori* forecast impact estimates are presented for IASI and AIRS instrument channels assimilated in NAVDAS-AR during May of 2013. Summary and prospects for further applications are in Sect. 5. A few fundamental properties of the *a posteriori*-derived error covariance estimates are given in the appendix.

2 A Posteriori Observation Error Covariance Diagnosis

The framework considered in this study is of an analysis state expressed as

$$\mathbf{x}^a = \mathbf{x}^b + \mathbf{K} [\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \quad (1)$$

$$\mathbf{K} = \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1} \quad (2)$$

where \mathbf{x}^b is a prior (background) estimate of the true state \mathbf{x}^t , \mathbf{y} is the observation vector, \mathbf{h} is the observational operator and \mathbf{H} denotes its linearized version (Jacobian matrix) evaluated at \mathbf{x}^b . Statistical information on the background error $\boldsymbol{\varepsilon}^b = \mathbf{x}^b - \mathbf{x}^t$ and the observational errors $\boldsymbol{\varepsilon}^o = \mathbf{y} - \mathbf{h}(\mathbf{x}^t)$ is used to specify the matrices \mathbf{B} and \mathbf{R} that represent, respectively, the true (unknown) background error covariance \mathbf{B}_t and observational error covariance \mathbf{R}_t . Hereafter the analysis (1)–(2) associated with the error covariance specification (\mathbf{B}, \mathbf{R}) is referred to as the *status quo* DAS.

In the context of linear estimation theory, Desroziers et al. (2005) have shown that estimates $\tilde{\mathbf{R}}$ to \mathbf{R}_t and $\tilde{\mathbf{B}}$ to \mathbf{B}_t may be obtained from the statistical analysis of the observation residuals, respectively, as

$$\tilde{\mathbf{R}} = \mathcal{E} [\mathbf{d}_a^o (\mathbf{d}_b^o)^T] \quad (3)$$

$$\tilde{\mathbf{H}} \mathbf{B} \mathbf{H}^T = \mathcal{E} [\mathbf{d}_b^o (\mathbf{d}_b^o)^T] \quad (4)$$

where \mathcal{E} denotes the statistical expectation operator, $\mathbf{d}_b^o = \mathbf{y} - \mathbf{h}(\mathbf{x}^b)$, $\mathbf{d}_a^o = \mathbf{y} - \mathbf{h}(\mathbf{x}^a)$, and $\mathbf{d}_b^a = \mathbf{h}(\mathbf{x}^a) - \mathbf{h}(\mathbf{x}^b)$. This approach has prompted the investigation of observation error covariance tuning procedures based on the estimates (3). However, as pointed

out by Talagrand (1999), *a posteriori* evaluation relies on *a priori* hypotheses on the error correlation structures. Assuming a linear observational operator, $\mathbf{h}(\mathbf{x}) = \mathbf{H}\mathbf{x}$, the right side term in (3) is expressed as

$$\mathcal{E} [\mathbf{d}_a^o(\mathbf{d}_b^o)^T] = (\mathbf{I} - \mathbf{H}\mathbf{K}) (\mathbf{I} - \mathbf{H}\mathbf{K}_t)^{-1} \mathbf{R}_t \quad (5)$$

where \mathbf{K}_t denotes the *optimal* Kalman gain matrix,

$$\mathbf{K}_t = \mathbf{B}_t \mathbf{H}^T (\mathbf{H}\mathbf{B}_t \mathbf{H}^T + \mathbf{R}_t)^{-1} \quad (6)$$

By combining (3) and (5), it is noticed that the following equivalence holds (Daescu and Langland 2013c)

$$\tilde{\mathbf{R}} = \mathbf{R}_t \Leftrightarrow \mathbf{H}\mathbf{K} = \mathbf{H}\mathbf{K}_t \quad (7)$$

Therefore, the quality of the estimate $\tilde{\mathbf{R}} \approx \mathbf{R}_t$ is closely determined by the error covariance specification (\mathbf{B}, \mathbf{R}) in the *status quo* DAS. Todling (2015) provides a cautionary note on misrepresented observation error correlations when little is known about the true covariances $\mathbf{B}_t, \mathbf{R}_t$. Covariance estimation in the context of ensemble data assimilation may alleviate some of these issues, as presented by Waller et al. (2014). In practice, tuning the observation error covariance based on the estimate $\tilde{\mathbf{R}}$ may improve or degrade the analysis and further insight is given in the Appendix. Additional information is necessary to assess the potential benefit of the model $\tilde{\mathbf{R}}$ and to design error covariance tuning procedures that are effective in *improving* the DAS performance.

3 Forecast Sensitivity to Observation Error Covariance (FSR)

A comprehensive set of equations to evaluate the forecast sensitivity with respect to various input parameters of a 4D-Var DAS is provided by Daescu and Langland (2013a, b). The evaluation of the forecast σ_o^2 -sensitivity allows assessing the potential benefits of the observation error variance specification $\tilde{\sigma}_o^2$ derived from *a posteriori* diagnosis. By extending the adjoint-DAS methodology to error correlation parameters, FSR provides a computationally feasible approach to obtain guidance on the forecast performance of a trial error covariance model $\tilde{\mathbf{R}}$ *prior to* its actual implementation in the analysis scheme (1)–(2).

Consider a scalar measure to the forecast error defined as

$$e(\mathbf{x}^a) = (\mathbf{x}_f^a - \mathbf{x}_f^v)^T \mathbf{E} (\mathbf{x}_f^a - \mathbf{x}_f^v) \quad (8)$$

where $\mathbf{x}_f^a = \mathcal{M}_{t_0, t_f}(\mathbf{x}^a)$ is the analysis forecast at verification time t_f initiated at t_0 from \mathbf{x}^a , \mathbf{x}_f^v is the verifying analysis at t_f (a proxy for the true state \mathbf{x}_f^t), and \mathbf{E} is a diagonal

matrix of weights that gives (8) units of energy per unit mass. The FSR is expressed as the rank-one matrix (Daescu and Langland 2013a, b),

$$\frac{\partial e}{\partial \mathbf{R}} = -\frac{\partial e}{\partial \mathbf{y}} \mathbf{z}^T, \quad \frac{\partial e}{\partial \mathbf{R}_{ij}} = -\frac{\partial e}{\partial y_i} \mathbf{z}_j \quad (9)$$

where the vector

$$\frac{\partial e}{\partial \mathbf{y}} = \mathbf{K}^T \frac{\partial e}{\partial \mathbf{x}^a} \quad (10)$$

denotes the forecast sensitivity to observations (Baker and Daley 2000) and \mathbf{z} denotes the weighted residual vector

$$\mathbf{z} = \mathbf{R}^{-1} [\mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b)] \approx \mathbf{R}^{-1} [\mathbf{y} - \mathbf{h}(\mathbf{x}^a)] \quad (11)$$

The observation error covariance model is expressed as

$$\mathbf{R} = \Sigma^o \mathbf{C}^o \Sigma^o \quad (12)$$

where $\Sigma^o = \text{diag}(\sigma_o)$ is the diagonal matrix of the specified observational error standard deviation $\Sigma_{ii}^o = \sigma_{o,i}$, and \mathbf{C}^o is the observational error correlation model, a symmetric and positive definite matrix with all diagonal entries of 1. The forecast \mathbf{C}^o -sensitivity is the rank-one matrix (Daescu and Langland 2013a, b)

$$\frac{\partial e}{\partial \mathbf{C}^o} = - \left(\Sigma^o \frac{\partial e}{\partial \mathbf{y}} \right) (\Sigma^o \mathbf{z})^T, \quad \frac{\partial e}{\partial \mathbf{C}_{ij}^o} = -\sigma_{o,i} \sigma_{o,j} \frac{\partial e}{\partial y_i} z_j \quad (13)$$

The evaluation of the forecast \mathbf{C}^o -sensitivity (13) allows the investigation of structured error correlation models. Consider a structure of the observational error correlation model defined by the observation relationship \mathcal{C} . Let I and J denote the index set of observations in relationship \mathcal{C} ,

$$(i, j) \in I \times J \Rightarrow y_i \mathcal{C} y_j \quad (14)$$

For example, a model to inter-channel observational error correlations may be structured according to $I = \{i : y_i \in \text{instrument channel } \#I\}$ and $J = \{j : y_j \in \text{instrument channel } \#J\}$; models to account for homogeneous and isotropic spatial error correlations (Gaspari and Cohn 1999; Gneiting 1999) may be structured based on the separation distance d between observation pairs, $I \times J = \{(i, j) : \text{dist}(y_i, y_j) = d\}$. Assuming that a scalar correlation coefficient (parameter) ρ is associated with the observation error correlation structure \mathcal{C} ,

$$(i, j) \in I \times J \Rightarrow C_{ij}^o = \rho \quad (15)$$

and accounting for the symmetry of the correlation model, $C_{ij}^o = C_{ji}^o$, from (13) and (15) the sensitivity embedded in the correlation structure \mathcal{C} is evaluated as

$$\frac{\partial e}{\partial \rho} = - \sum_{(i,j) \in I \times J} \boldsymbol{\sigma}_{o,i} \boldsymbol{\sigma}_{o,j} \left[\frac{\partial e}{\partial y_i} z_j + \frac{\partial e}{\partial y_j} z_i \right] \quad (16)$$

3.1 Estimation of the Forecast Impact of the Model $\tilde{\mathbf{R}}$

The forecast impact of the observation error covariance model $\tilde{\mathbf{R}}$ as compared with the *status quo* specification \mathbf{R} is defined as

$$\delta e = e[\mathbf{x}^a(\tilde{\mathbf{R}})] - e[\mathbf{x}^a(\mathbf{R})] \quad (17)$$

where $\mathbf{x}^a(\tilde{\mathbf{R}})$ is the analysis state associated with the model $\tilde{\mathbf{R}}$ in (1)–(2). A first order approximation to δe may be obtained from the FSR derivative information as

$$\delta e_1 = \text{Tr} \left[\frac{\partial e}{\partial \mathbf{R}} (\delta \mathbf{R})^T \right] = \sum_{i,j} \frac{\partial e}{\partial \mathbf{R}_{ij}} \delta \mathbf{R}_{ij} \quad (18)$$

where Tr denotes the matrix trace operator and

$$\delta \mathbf{R} = \tilde{\mathbf{R}} - \mathbf{R} \quad (19)$$

is a symmetric matrix, $\delta \mathbf{R} = (\delta \mathbf{R})^T$, representing the variation in the observation error covariance model. The rank-one matrix structure of the FSR (9) allows an efficient estimation of the first order impact (18) as (Daescu and Langland 2013b)

$$\delta e_1 = -\mathbf{z}^T [\delta \mathbf{R}] \frac{\partial e}{\partial \mathbf{y}} \quad (20)$$

Since the FSR (9) is evaluated in the *status quo* DAS, the impact estimate (18)/(20) is obtained *prior to* the actual implementation of the trial model $\tilde{\mathbf{R}}$ in the analysis (1)–(2) and provides guidance on its forecast performance.

Typically, observational error correlations are not modeled in operational atmospheric data assimilation systems. Henceforth in this section it is assumed that \mathbf{C}^o in (12) is the identity matrix, $\mathbf{C}^o = \mathbf{I}$, and thus \mathbf{R} is specified as a diagonal matrix,

$$\mathbf{R} = (\Sigma^o)^2 = \text{diag}(\boldsymbol{\sigma}_o^2) \quad (21)$$

A common approach for tuning the model \mathbf{R} is to adjust the specified observational error variance parameters only, without accounting for error correlations. In this con-

text, $\tilde{\mathbf{R}} = \text{diag}(\tilde{\boldsymbol{\sigma}}_o^2)$ and a first order estimate to the forecast impact induced by the variation

$$\delta \mathbf{R}_\sigma = \left(\tilde{\boldsymbol{\Sigma}}^o \right)^2 - (\boldsymbol{\Sigma}^o)^2 \quad (22)$$

is obtained as

$$e \left[\mathbf{x}^a (\tilde{\boldsymbol{\sigma}}_o^2) \right] - e \left[\mathbf{x}^a (\boldsymbol{\sigma}_o^2) \right] \approx - \sum_{i=1}^p \left(\delta \sigma_{o,i}^2 \right) \left(\frac{\partial e}{\partial y_i} z_i \right) \quad (23)$$

where $\delta \sigma_{o,i}^2 = \tilde{\sigma}_{o,i}^2 - \sigma_{o,i}^2$ denotes the change in the observational error variance specification associated with the observation y_i and p denotes the dimension of the observation vector \mathbf{y} .

3.1.1 The Impact of Modeling Observational Error Correlations

Of particular significance to practical applications is the ability to assess the *added benefit* of an observational error correlation model $\tilde{\mathbf{C}}^o$, as compared with a tuning procedure that only adjusts the error variance parameters. Consider the covariance model $\tilde{\mathbf{R}}$ represented as

$$\tilde{\mathbf{R}} = \tilde{\boldsymbol{\Sigma}}^o \tilde{\mathbf{C}}^o \tilde{\boldsymbol{\Sigma}}^o \quad (24)$$

From (21) and (24), the variation (19) is expressed as

$$\delta \mathbf{R} = \tilde{\boldsymbol{\Sigma}}^o \tilde{\mathbf{C}}^o \tilde{\boldsymbol{\Sigma}}^o - (\boldsymbol{\Sigma}^o)^2 = \underbrace{\left(\tilde{\boldsymbol{\Sigma}}^o \right)^2 - (\boldsymbol{\Sigma}^o)^2}_{\delta \mathbf{R}_\sigma} + \underbrace{\tilde{\boldsymbol{\Sigma}}^o \left(\tilde{\mathbf{C}}^o - \mathbf{I} \right) \tilde{\boldsymbol{\Sigma}}^o}_{\delta \mathbf{R}_c} \quad (25)$$

First term in the right side of (25) is the diagonal matrix $\delta \mathbf{R}_\sigma$ defined in (22), whereas second term

$$\delta \mathbf{R}_c = \tilde{\boldsymbol{\Sigma}}^o \left(\tilde{\mathbf{C}}^o - \mathbf{I} \right) \tilde{\boldsymbol{\Sigma}}^o \quad (26)$$

is a matrix with all main diagonal entries set to zero and represents the additional contribution to $\delta \mathbf{R}$ induced by the variation $\delta \mathbf{C}^o = \tilde{\mathbf{C}}^o - \mathbf{I}$ in the error correlation model. The first order estimate (20) to the forecast impact is expressed as

$$\delta e_1 = \delta e_\sigma + \delta e_c \quad (27)$$

where

$$\delta e_\sigma = -\mathbf{z}^T [\delta \mathbf{R}_\sigma] \frac{\partial e}{\partial \mathbf{y}} \quad (28)$$

is the estimate (23) to the forecast impact of adjusting only the observational error variance and

$$\delta e_c = -\mathbf{z}^T [\delta \mathbf{R}_c] \frac{\partial e}{\partial \mathbf{y}} \quad (29)$$

is an estimate to the added impact associated with the error correlation model $\tilde{\mathbf{C}}^o$. If the correlation model $\tilde{\mathbf{C}}^o$ is structured as in (14) and (15), the estimate (29) to the forecast impact of the correlation structure \mathcal{C} embedded in the covariance model \mathbf{R} is evaluated as

$$\delta e_{\mathcal{C}} = -\rho \sum_{(i,j) \in I \times J} \tilde{\sigma}_{o,i} \tilde{\sigma}_{o,j} \left[\frac{\partial e}{\partial y_i} z_j + \frac{\partial e}{\partial y_j} z_i \right] \quad (30)$$

This approach is used in the numerical experiments for IASI and AIRS instruments to obtain *a priori* estimates (30) to the added benefit of inter-channel error correlations derived from observation-space residuals, as compared with estimates (23) to the forecast impact of tuning only the observational error variance parameters.

3.2 Proof-of-Concept with Lorenz Model

A proof-of-concept to *a priori* adjoint-based estimation of the forecast impact of correlated observational errors is presented with Lorenz's 40-variable model (Lorenz and Emanuel 1998)

$$\frac{d x_j}{d t} = (x_{j+1} - x_{j-2})x_{j-1} - x_j + F, \quad j = 1 : n \quad (31)$$

where $n = 40$, $x_{n+j} = x_j$, and the forcing constant is specified as $F = 8$. The system (31) is integrated with the standard fourth-order explicit Runge-Kutta method and a constant time step $\Delta t = 0.005$, hereafter identified with a 36 min time interval, to produce a reference trajectory ("the truth") \mathbf{x}^t . An initial state \mathbf{x}_0^t is obtained by performing a 90-days (3600 time steps) integration initialized from $x_j = 8$ for $j \neq n/2$ and $x_{n/2} = 8.008$. A 4D-Var data assimilation framework is considered as follows. The length of each assimilation interval $[t_0, t_N]$ is specified as $t_N - t_0 = 0.05$ i.e., a 6 h time interval. Observations are assumed to be taken by a rotating instrument that completes a full rotation in the 6-h assimilation window, as illustrated in Fig. 1a. At each time step $t_i = t_0 + i\Delta t$, $i = 1 : 10$, four observations are assimilated $y_{4i-3}, y_{4i-2}, y_{4i-1}, y_{4i}$ associated with the model variables $x_{4i-3}, x_{4i-2}, x_{4i-1}, x_{4i}$, respectively. The observational errors $\varepsilon^o = \mathbf{y} - \mathbf{x}^t$ are assumed to be normally distributed, unbiased, *uncorrelated in time*, and with the standard deviation of $\sigma_{o,t} = 0.25$. The errors in observations taken *at the same time* t_i are assumed to be correlated according to an autoregressive model of order one (AR-1) with parameter $0 < \rho < 1$, $\mathbf{C}_t^o(i_1, i_2) = \rho^{|i_1 - i_2|}$.

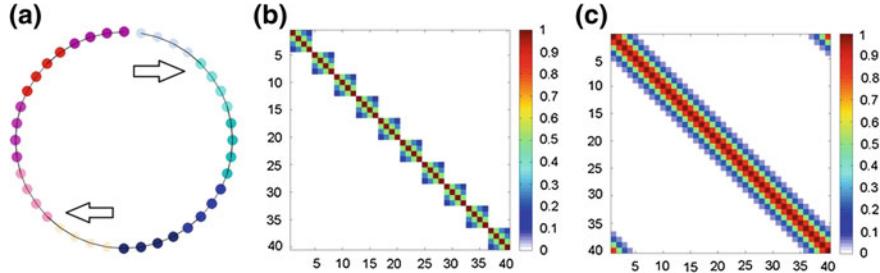


Fig. 1 **a** A rotating instrument provides observations of four state variables at each time step of the 6-h assimilation interval. **b** The structure of the true observational error correlation matrix \mathbf{C}_t^o . **c** The structure of the assigned background error correlation matrix \mathbf{C}^b

In this setup, the true error correlation matrix \mathbf{C}_t^o has a block diagonal structure, as illustrated in Fig. 1b, and $\mathbf{R}_t = \sigma_{o,t}^2 \mathbf{C}_t^o$. It is further assumed that observational error correlations are not modeled in the *status quo* DAS, $\mathbf{C}^o = \mathbf{I}$, and \mathbf{R} is specified as $\mathbf{R} = \sigma_{o,t}^2 \mathbf{I}$. Specifically, at each time step the diagonal blocks of the error covariance matrix \mathbf{R}_t and its representation \mathbf{R} in the DAS are respectively,

$$\mathbf{R}_t[1 : 4, 1 : 4] = \sigma_{o,t}^2 \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 \\ \rho & 1 & \rho & \rho^2 \\ \rho^2 & \rho & 1 & \rho \\ \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}, \quad \mathbf{R}[1 : 4, 1 : 4] = \sigma_{o,t}^2 \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad (32)$$

The background estimate \mathbf{x}^b to the initial state \mathbf{x}_0^t is prescribed by introducing random perturbations in \mathbf{x}_0^t taken from the standard normal distribution, $\epsilon^b \sim N(0, 1)$. The background error covariance matrix \mathbf{B}_t is modeled in the DAS by a static matrix $\mathbf{B} = \Sigma^b \mathbf{C}^b \Sigma^b$. The background error standard deviation is assigned as $\sigma_b = 1$ and the background error correlation matrix \mathbf{C}^b is constructed based on the 5th-order piecewise rational function defined in Eq. (4.10) of Gaspari and Cohn (1999) with a correlation length parameter $c = 3$, $\mathbf{C}^b(i, j) = 0$ if $|i - j| \geq 2c$. The specified correlation matrix \mathbf{C}^b is shown in Fig. 1c.

A twin experiments framework is used to investigate the ability of the adjoint-DAS FSR approach to provide an *a priori* estimate to the impact of observational error correlations. It is emphasized that the FSR estimates to the forecast impact are each valid for a single analysis episode and without accounting for the impact propagation (dependency) into subsequent assimilation cycles. Therefore, the long term forecast impact due to a permanent change in the specified parameter value is not quantified. For consistency with the theoretical formulation to the FSR-based \mathbf{R} -impact approximation, the setup of the observing system experiments (OSEs) is as illustrated in Fig. 2. At each assimilation cycle the only difference between the *status quo* (control) and experiment consists on the \mathbf{R} -specification: the experiment is initialized with the same input $(\mathbf{y}, \mathbf{x}^b, \mathbf{B})$ used to produce the control analysis \mathbf{x}^a and by replacing \mathbf{R} with \mathbf{R}_t to produce the analysis $\bar{\mathbf{x}}^a$.

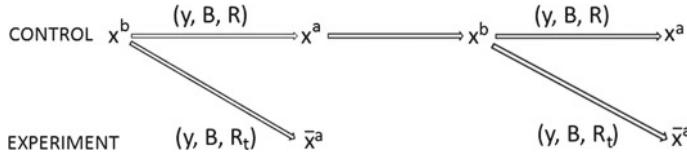


Fig. 2 The twin experiments setup to validate the FSR approach to forecast error **R**-impact estimation

The forecast aspect (8) is taken as the squared Euclidean norm of the error in the 24-h forecast verified against the reference state \mathbf{x}_f^t . The simplicity of the Lorenz model allows the investigation of the FSR approach to forecast **R**-impact estimation (18)–(20) for various values of the error correlation coefficient ρ in (32). It is noticed that the magnitude of the perturbation $\delta\mathbf{R}$, as measured by its Frobenius norm, increases with the correlation strength parameter ρ . Numerical results are presented for parameter values $\rho = 0.25$ (EXP-1, weak correlations), $\rho = 0.5$ (EXP-2, moderate correlations), and $\rho = 0.75$ (EXP-3, strong correlations). A new set of observations $\mathbf{y} = \mathbf{x}^t + \boldsymbol{\varepsilon}^o$, $\boldsymbol{\varepsilon}^o \sim N(\mathbf{0}, \mathbf{R}_t)$ was generated for each parameter specification and accordingly, both control and experiment data assimilation/forecast cycles were run for a 37-month period (4440 cycles). First month was considered a spin-up period and a statistical analysis of the forecast error **R**-impact determined through OSEs as

$$\delta e = e[\bar{\mathbf{x}}^a(\mathbf{R}_t)] - e[\mathbf{x}^a(\mathbf{R})] \quad (33)$$

and the first order FSR-based estimate δe_1 defined in (20) was performed over the remaining 36-month period. For each experiment, scatter plots of the **R**-impact δe and its *a priori* approximation δe_1 collected for each analysis/forecast cycle are shown in Fig. 3. From the time series of analyses/forecasts, the correlation coefficient of the two forecast error impact measures, $\text{corrcoef}(\delta e, \delta e_1)$, was found to be of ~ 0.98 in EXP-1, ~ 0.95 in EXP-2, and ~ 0.9 in EXP-3 which indicates the ability of the FSR approach to provide proper guidance on the impact of observational error correlations. In average, throughout the 36-month period, insertion of the \mathbf{R}_t model in the DAS was found to have a benefic impact on forecasts, $\delta e < 0$. The percentage of analysis/forecast episodes associated with $\delta e < 0$ was closely determined by the correlation strength parameter and increased from $\sim 61\%$ in EXP-1 to $\sim 78\%$ in EXP-3. For comparison, a plot of the 30-day (120 analysis/forecast cycles) moving average of the **R**-impact on the forecast error is shown in Fig. 4 for weak correlations (EXP-1, $\rho = 0.25$) and strong correlations (EXP-3, $\rho = 0.75$). It is noticed that in average, the first-order FSR-approximation δe_1 overestimates the actual forecast error impact δe of the error covariance \mathbf{R}_t . Specifically, the average relative error in the approximation δe_1 was found to be of $\sim 29\%$ in EXP-1 and of $\sim 42\%$ in EXP-3.

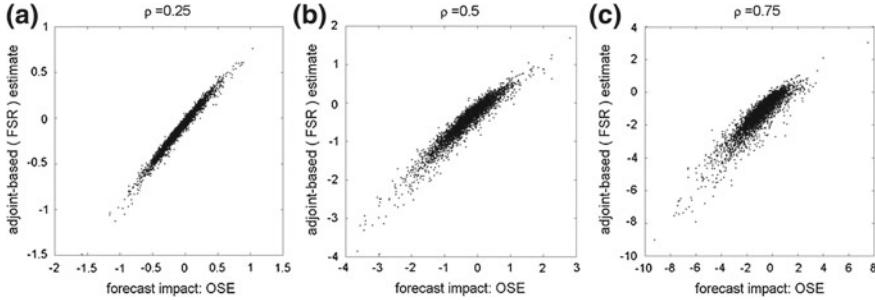


Fig. 3 Scatter plots of the **R**-impact on the forecast error determined through OSEs and the associated adjoint-based FSR approximation: **a** EXP-1, **b** EXP-2, and **c** EXP-3

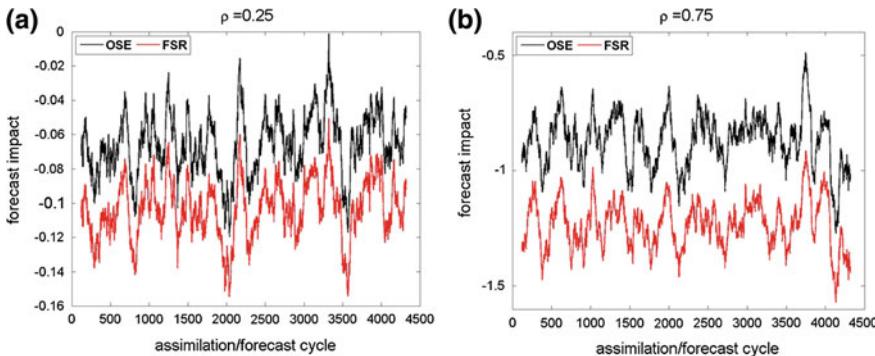


Fig. 4 The 30-day moving average of the forecast error **R**-impact determined through OSEs and the associated first order adjoint-based FSR approximation: **a** EXP-1 and **b** EXP-3

4 Results with NAVDAS-AR/NAVGEN

The practical ability to combine information derived from *a posteriori* error covariance diagnosis and adjoint-based FSR is shown in numerical experiments performed with NAVDAS-AR/NAVGEN and their adjoint versions. The forecast aspect (8) is specified as the error in the 24-h forecasts produced by NAVGEN at a resolution of T359L50, verified against the NAVDAS-AR analyses valid at $t_f = t_0 + 24$ h and measured in a moist total energy norm over the global domain. The results presented are valid for the analyses/forecasts produced for the time period May 1–31 of 2013. The contribution of various observation types assimilated in NAVDAS-AR to forecast error reduction has been evaluated using the second-order approximation measure of Langland and Baker (2004). Average values of the observation impact estimates associated with the analyses/forecasts produced at each 6-h cycle of the 4D-Var assimilation interval are shown in Fig. 5a. The complementary information provided by the sensitivity to observation error covariance is shown in Fig. 5b. Observational error correlations are not modeled in the NAVDAS-AR version considered

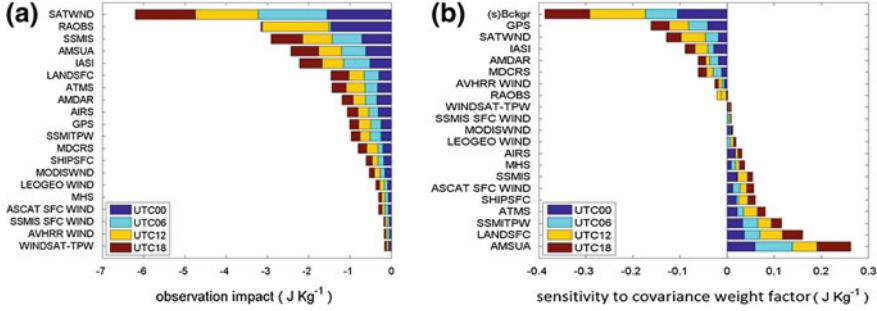


Fig. 5 **a** The observation impact (J kg^{-1}) on the 24 h forecast error reduction for various observation types assimilated in NAVDAS-AR. **b** The 24 h forecast error sensitivity (J kg^{-1}) with respect to the observation error covariance weight coefficient s_i^o ; also shown is the forecast error s^b -sensitivity, (s)Bckgr

in our experiments, $\mathbf{C}^o = \mathbf{I}$. To facilitate the comparison between various observation types \mathbf{y}_i , $i = 1 : I$, the forecast sensitivity is evaluated with respect to the error covariance weight coefficient s_i^o (non-dimensional scalar) in the parametric representation $\mathbf{R}_i(s_i^o) = s_i^o \mathbf{R}_i$ (Daescu and Langland 2013a, b)

$$\frac{\partial e}{\partial s_i^o} = -(\mathbf{R}_i \mathbf{z}_i)^T \frac{\partial e}{\partial \mathbf{y}_i}, \quad i = 1 : I \quad (34)$$

Also displayed in Fig. 5b is the forecast sensitivity with respect to the background error covariance weight factor, $\mathbf{B}(s^b) = s^b \mathbf{B}$, evaluated as

$$\frac{\partial e}{\partial s^b} = (\mathbf{R} \mathbf{z})^T \frac{\partial e}{\partial \mathbf{y}} \quad (35)$$

From (11) it is noticed that

$$\mathbf{R} \mathbf{z} = \mathbf{y} - \mathbf{h}(\mathbf{x}^b) - \mathbf{H}(\mathbf{x}^a - \mathbf{x}^b) \approx \mathbf{y} - \mathbf{h}(\mathbf{x}^a) \quad (36)$$

and the evaluation of (34) and (35) is performed by taking the corresponding inner product between the analysis residual and the observation sensitivity. The *status quo* corresponds to the parameter values $s_i^o = 1$, $i = 1 : I$, $s^b = 1$ and the sensitivities provide guidance to improve the relative weighting of various input components in the DAS. For the case study presented here, the s^b -sensitivity indicates that inflation of the background error covariance, $s^b > 1$, is of potential benefit to the forecasts. Satellite-derived cloud-drift and water vapor-motion winds (SATWND) have a particularly large impact on the forecast error reduction however, the FSR identifies this data type as over-weighted. Overall, assimilation of radiances from the Advanced Microwave Sounding Unit (AMSU)-A may benefit from reducing the assigned σ_o , whereas the guidance for the IASI instrument is to inflate the assigned σ_o values. It is

emphasized that the FSR analysis only identifies the direction to adjust the assigned error covariance parameters and without providing the actual parameter values for covariance tuning. The combined information derived from *a posteriori* covariance estimation and FSR is detailed next for IASI and AIRS instruments based on the analyses and 24 h forecasts valid at 0000 UTC.

4.1 Estimates of the Observational Error Standard Deviation

A statistical analysis of bias-corrected innovations \mathbf{d}_b^o and residuals \mathbf{d}_a^o produced by NAVDAS-AR was performed to assess the consistency of the specified observational error standard deviation for radiance data provided by the atmospheric sounders AIRS and IASI. During the time period considered in this study, NAVDAS-AR assimilated 64 channels from AIRS and 73 channels from IASI, as listed in Tables 1 and 2, respectively.

Figure 6 displays the values of σ_o (K) as specified in NAVDAS-AR and the estimates $\tilde{\sigma}_o$ derived from observation residuals (3) for each instrument channel. In general, for both instruments the estimates $\tilde{\sigma}_o$ are of significantly lower magnitude as compared with the values assigned in NAVDAS-AR. These findings are consistent with the results obtained at other NWP centers (e.g., Bormann et al. 2010; Stewart et al. 2014; Weston et al. 2014).

Table 1 AIRS channels assimilated in NAVDAS-AR in May 2013

Channel group	AIRS channel numbers
Long-wave CO ₂ , upper temperature-sounding	92 93 99 104 105 110 111 116 117 123 128 129 138 139 144 145 151 156 157 162 168 169 170 172 173 174 175 177 179 180 182 185 186 190 192 193 198 201 204 207 210 213 215 216 218 239 250 251
Long-wave CO ₂ , lower temperature-sounding	253 308 318
Short-wave channels	1881 1882 1883 1884 1897 1901 1911 1917 1918 1921 1923 1924 1928

Table 2 IASI channels assimilated in NAVDAS-AR in May 2013

Channel group	IASI channel numbers
Long-wave CO ₂ , upper temperature-sounding	122 128 135 141 148 154 161 173 185 187 193 199 205 207 210 212 214 217 219 222 224 226 230 232 236 239 246 249
Long-wave CO ₂ , lower temperature-sounding	252 254 260 262 265 267 269 282 306 323 329 335 347 350 354 356 360 366 371 373 375 377 379
Water vapor sensitive	2889 2944 2948 2951 2958 3098 3168 3248 3281 3309 3442 3444 3448 3450 3491 3506 3575 3577 3582 3589 3653 3661

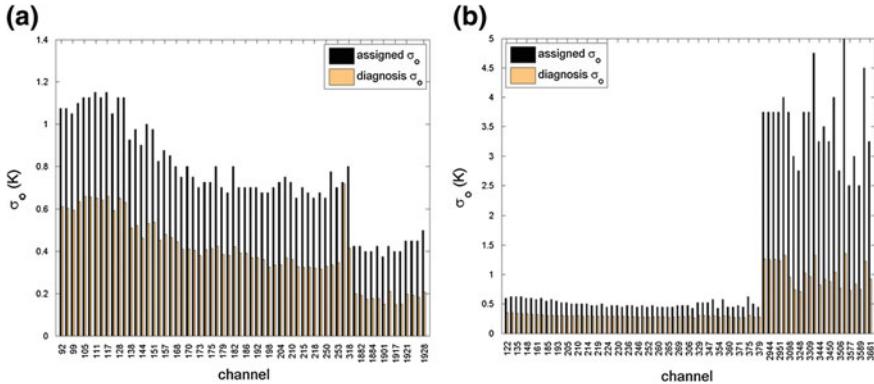


Fig. 6 The values of σ_o (K) specified in NAVDAS-AR and the *a posteriori* (Desroziers) estimates $\tilde{\sigma}_o$ (K) for **a** AIRS and **b** IASI

4.2 Estimates of the Inter-Channel Error Correlations

The inter-channel observational error correlations $\tilde{\mathbf{C}}_{ij}^o$ for AIRS and IASI instruments have been estimated based on the Desroziers diagnostic (3) as

$$\tilde{\mathbf{R}}_{ij} = \mathcal{E} \{ [\mathbf{d}_a^o]_{ch\#i} [\mathbf{d}_b^o]_{ch\#j} \}, \quad \tilde{\mathbf{C}}_{ij}^o = \frac{1}{\tilde{\sigma}_{o,i} \tilde{\sigma}_{o,j}} \tilde{\mathbf{R}}_{ij} \quad (37)$$

The evaluation of (37) incorporates each pair of observations from the instrument channels i and j taken at the same location in the time-space domain. For any practical applications, the symmetric form of (37) is considered by taking

$$\tilde{\mathbf{R}}_{sym} = \frac{1}{2} \left(\tilde{\mathbf{R}} + \tilde{\mathbf{R}}^T \right) \quad (38)$$

The estimated inter-channel error correlations for AIRS and IASI are shown in Fig. 7. For each instrument, the *a posteriori* diagnosis indicates that in general, there are weak (<0.2) or no correlations for the long-wave upper temperature sounding channels. A few AIRS channels ranging between 201–253, display mild correlations (0.2–0.5). Error correlation coefficients in the range of 0.2–0.6 have been obtained for the long-wave lower temperature-sounding channels of IASI (channels 252–379). Error correlations of increased magnitude have been estimated for the water-vapour (humidity-sensitive) channels assimilated from IASI and the short-wave channels assimilated from AIRS. These estimates are consistent with the results of Bormann et al. (2010) in the ECMWF assimilation system and of Weston et al. (2014) with the Met Office 4D-Var.

For practical applications, the performance of a covariance specification determined by the diagnostic estimates may be impaired by several factors such as the

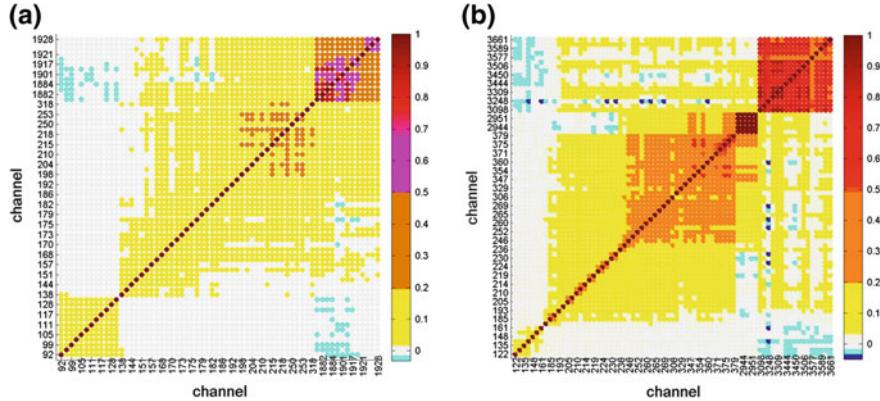


Fig. 7 Estimates of the inter-channel error correlations derived from Desroziers diagnostic for **a** AIRS and **b** IASI

validity of the *a priori* assumptions on the error correlation structures, tuning the **R**-model only in the presence of deficiencies in both **B**- and **R**-models, uncertainties in the estimation entailed by a nonlinear observational operator, observation and model biases, and the amount of information available for statistical analysis e.g., the size of the ensemble of innovation and residual vectors. An unresolved issue is to assess whether adjusting the error covariance parameters based on the diagnostic estimates will improve the forecasts. For each instrument, the FSR-based forecast error sensitivity was evaluated to obtain *a priori* guidance on the forecast error performance of the observation error variance specification $\tilde{\sigma}_o^2$ and the added benefit of incorporating the inter-channel error correlation model $\tilde{\mathbf{C}}^o$ in NAVDAS-AR.

4.3 The FSR Guidance and Forecast Error Impact Estimates

The forecast error sensitivity to the observation error variance weight parameters $\sigma_{o,i}^2(s_i^o) = s_i^o \sigma_{o,i}^2$ is evaluated from (34) for each instrument channel. The average s_i^o -sensitivity values associated with AIRS and IASI channels is shown in Fig. 8. Positive sensitivity values identify those instrument channels that may benefit from reducing the assigned σ_o , whereas negative sensitivity values point toward error variance inflation. The FSR guidance is that in general, forecasts will benefit from *reducing* the σ_o assigned to the majority of the AIRS channels and from further *inflating* the σ_o assigned to the long-wave lower temperature-sounding channels of IASI. In particular, it is estimated that tuning the long-wave lower temperature-sounding channels of IASI to the diagnosis estimates $\tilde{\sigma}_o$ will *degrade the forecasts*. The s_i^o -sensitivity displays mixed results and without a clear tendency for the long-wave upper temperature-sounding channels of IASI. The estimates (22) and (23) to the

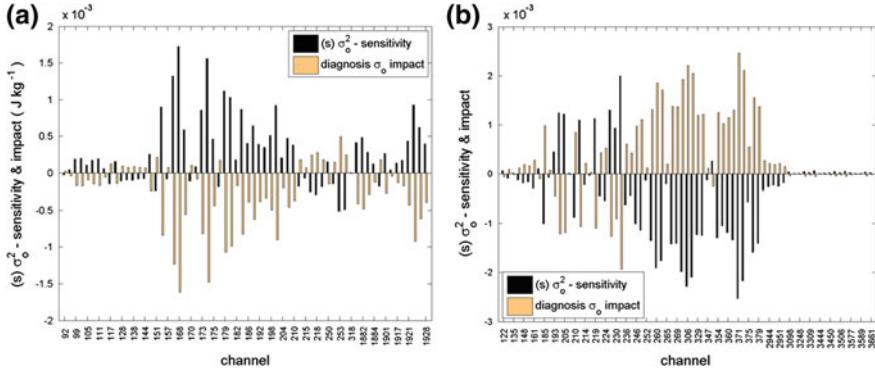


Fig. 8 Sensitivity of the 24 h forecast error (J kg^{-1}) to the observation error variance weight parameters $(s)\sigma_o^2$ and *a priori* estimates to the forecast error impact (J kg^{-1}) of tuning the observation error variances to the diagnosis estimates $\tilde{\sigma}_o^2$. Time-averaged values per assimilation cycle are shown for **a** AIRS and **b** IASI.

forecast error impact of adjusting the observational error variance from σ_o^2 to the diagnosis value $\tilde{\sigma}_o^2$ are also shown in Fig. 8.

4.3.1 The Impact of Inter-Channel Error Correlations

The FSR approach has been implemented to obtain information on the forecast sensitivity and impact estimates of inter-channel error correlations. The sensitivity evaluation is performed based on the structured formulation (15) and (16) and by accounting for the symmetry of the matrix \mathbf{C}^o . Having available the observation sensitivity vector, the evaluation of the forecast error \mathbf{C}^o -sensitivity (16) is performed together with the estimation of the inter-channel error correlations $\tilde{\mathbf{C}}^o$ (37) by operating on the same structures of observation pairs. The sensitivity to inter-channel error correlations for AIRS and IASI is displayed in Fig. 9. Since the *status quo* corresponds to $\mathbf{C}^o = \mathbf{I}$, the guidance provided by the \mathbf{C}^o -sensitivity matrix is as follows: entries with negative values indicate that insertion of a positive correlation coefficient is of potential benefit to the forecasts, whereas entries with positive values indicate that insertion of a negative correlation coefficient is of potential benefit to the forecasts. In particular, it is noticed that the \mathbf{C}^o -sensitivity matrix associated with IASI exhibits a well-defined structure with negative values of increased magnitude for the long-wave lower temperature-sounding channels 252–379. In conjunction with the diagnosis estimates in Fig. 7, the FSR guidance is that insertion of the IASI inter-channel error correlation model $\tilde{\mathbf{C}}^o$ in NAVDAS-AR is of potential benefit to the forecasts. For comparison, the \mathbf{C}^o -sensitivity matrix associated with AIRS shows entrywise values of relatively lower magnitude and of increased variation in sign.

A priori estimates to the forecast impact of the model $\tilde{\mathbf{R}} = \tilde{\Sigma}^o \tilde{\mathbf{C}}^o \tilde{\Sigma}^o$ are obtained based on the first-order approximation (18). As explained in Sect. 3, the FSR allows

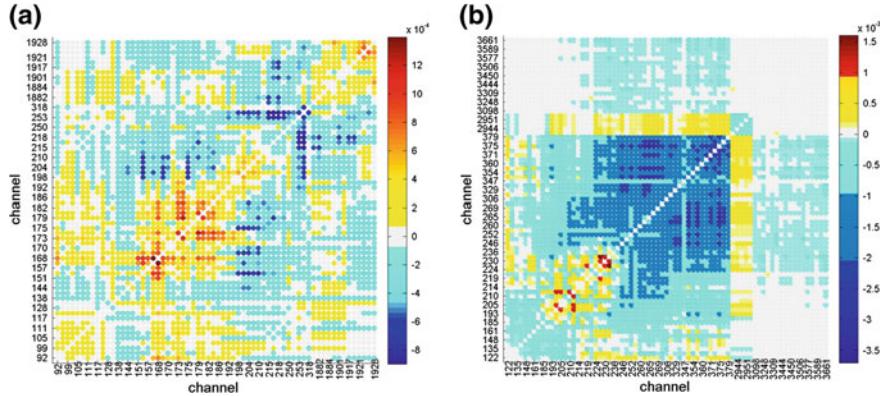


Fig. 9 Forecast error sensitivity (J kg^{-1}) to inter-channel error correlations. Displayed are time-averaged values per assimilation/forecast cycle for **a** AIRS and **b** IASI

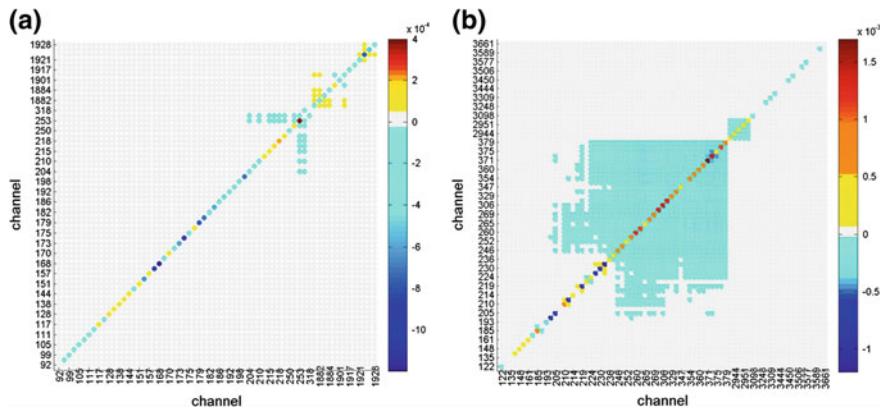


Fig. 10 First-order FSR-based estimates to the forecast error impact (J kg^{-1}) of the observational error covariance model $\tilde{\mathbf{R}}$. Displayed are time-averaged values per assimilation/forecast cycle for **a** AIRS and **b** IASI

estimation of the impact associated with a specified correlation structure \mathcal{C} in the model $\tilde{\mathbf{R}}$. The right side term in (18) provides an entrywise measure (impact matrix) of the forecast impact of the inter-channel error covariance model $\tilde{\mathbf{R}}$, as shown in Fig. 10. The main diagonal entries of the impact matrix provide estimates to the forecast error impact of tuning the observation error variances only, as expressed in (23)/(28). The off-diagonal elements of the impact matrix provide estimates to the *added impact* of modeling observation error correlations, as expressed in (29)/(30). The time-series of the estimated $\tilde{\sigma}_o^2$ -impact δe_σ and of the estimated $\tilde{\mathbf{R}}$ -impact δe_1 (27) are shown in Fig. 11. For the case study presented here, the FSR indicates little forecast impact from modeling inter-channel error correlations for AIRS radiances. The FSR guidance for IASI is that modeling inter-channel observation error corre-

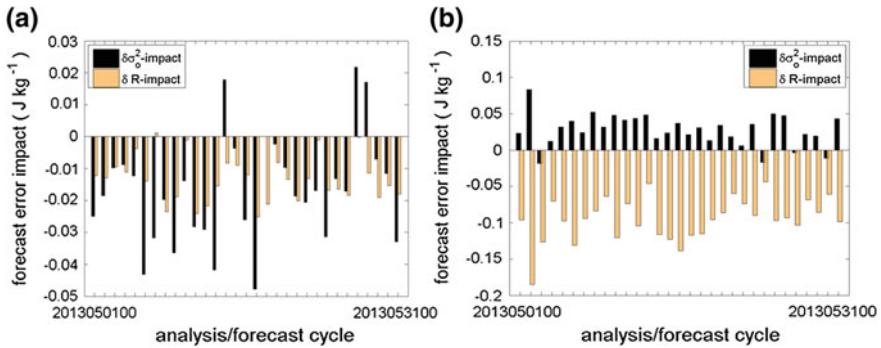


Fig. 11 The FSR-based estimates to the forecast error impact (J kg^{-1}) of tuning the observation error variance only ($\delta\sigma_o^2$ -impact) and of the error covariance specification $\tilde{\mathbf{R}}$ ($\delta\mathbf{R}$ -impact). Displayed is the time series of the estimates valid for May 2013 at 0000 UTC for **a** AIRS and **b** IASI

tions may provide an increased benefit to the forecasts and in particular, reduction of the assigned σ_o values is justified *if* inter-channel error correlations are incorporated in NAVDAS-AR.

5 Summary and Research Perspectives

Novel applications of the adjoint sensitivity tools have been investigated to analyze observation error correlation structures and obtain guidance on the forecast impact of a trial error covariance model *prior to* its actual implementation in the DAS. Until now, efforts to implement error correlation models for radiance data provided by hyperspectral instruments have relied mainly on *a posteriori* consistency diagnosis. The FSR approach allows identification of high-impact error correlation structures and provides valuable insight on the development of covariance models that are effective in reducing the forecast errors. A synergistic framework was proposed for linking these methodologies that combines complementary information on the error correlation structures and alleviates their individual shortcomings. The practical ability to exercise this novel adjoint capability was shown in a set of preliminary experiments with NAVDAS-AR/NAVGEN. Estimates of inter-channel error correlations for IASI and AIRS instruments are derived from the covariance diagnosis and first order estimates to the forecast error impact are obtained from the FSR derivative information. For the particular case study presented here, the FSR guidance is that modeling inter-channel error correlations would benefit the assimilation of IASI radiances and have little impact on the assimilation of AIRS radiances. The observation-space formulation of the NAVDAS-AR 4D-Var algorithm facilitates the specification of a non-diagonal observation error covariance model and research is ongoing at NRL-Monterey to account for inter-channel error correlations in the

assimilation of sounder radiance data (Campbell and Satterfield 2015). Further work is needed to perform the FSR analysis for various microwave and infrared sounders instruments and to validate the *a priori* FSR guidance through OSEs.

The observation sensitivity vector is the key ingredient to the FSR-based sensitivity analysis and may be also produced in an ensemble-based DAS (Liu and Kalnay 2008; Liu et al. 2009). Therefore, the FSR applications presented here in the adjoint-DAS 4D-Var framework may be as well implemented in ensemble-based DASs. Theoretical aspects of the adjoint-DAS estimation of the forecast error sensitivity to the **B**-matrix specification are discussed in our work Daescu and Langland (2013a,b) however, this capability has not been properly investigated for practical applications. It is expected that further research to obtain the forecast \mathbf{x}^b - and **B**-sensitivity information will find an extended range of applications in NWP.

Acknowledgements The work of D.N. Daescu was supported by the Naval Research Laboratory Atmospheric Effects, Analysis, and Prediction BAA #75-11-01 under award N00173-13-1-G903. Support for the second author from the sponsor ONR-PR-0602435N is gratefully acknowledged.

Appendix

As shown by Desroziers et al. (2005), the error covariance estimates $\tilde{\mathbf{R}}$ in (3) and $\tilde{\mathbf{B}}$ in (4) are expressed, respectively, as

$$\tilde{\mathbf{R}} = \mathbf{R} (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{H} \mathbf{B}_t \mathbf{H}^T + \mathbf{R}_t) \quad (39)$$

$$\mathbf{H} \tilde{\mathbf{B}} \mathbf{H}^T = \mathbf{H} \mathbf{B} \mathbf{H}^T (\mathbf{H} \mathbf{B} \mathbf{H}^T + \mathbf{R})^{-1} (\mathbf{H} \mathbf{B}_t \mathbf{H}^T + \mathbf{R}_t) \quad (40)$$

By analogy with (1) and (2), the analysis state associated with the covariance specification (\mathbf{B}, \mathbf{R}) is expressed as

$$\tilde{\mathbf{x}}^a = \mathbf{x}^b + \tilde{\mathbf{K}} [\mathbf{y} - \mathbf{h}(\mathbf{x}^b)] \quad (41)$$

$$\tilde{\mathbf{K}} = \tilde{\mathbf{B}} \mathbf{H}^T (\mathbf{H} \tilde{\mathbf{B}} \mathbf{H}^T + \tilde{\mathbf{R}})^{-1} \quad (42)$$

By adding (39) and (40), it is noticed that the estimates $(\tilde{\mathbf{B}}, \tilde{\mathbf{R}})$ are consistent with the innovation error covariance,

$$\mathbf{H} \tilde{\mathbf{B}} \mathbf{H}^T + \tilde{\mathbf{R}} = \mathbf{H} \mathbf{B}_t \mathbf{H}^T + \mathbf{R}_t \quad (43)$$

and this property has prompted research on covariance tuning procedures based on the diagnosis estimates (39), (40). However, as explained by Daescu and Langland (2013c), the operator **HK** remains invariant when the *status quo* specification (\mathbf{B}, \mathbf{R}) is replaced by the estimates $(\tilde{\mathbf{B}}, \tilde{\mathbf{R}})$,

$$\mathbf{HK} = \mathbf{H} \tilde{\mathbf{K}} \quad (44)$$

The relationship (44) was established in Daescu and Langland (2013c) as follows.

$$\begin{aligned}
 \mathbf{H}\tilde{\mathbf{K}} &\stackrel{(42)}{=} \mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T \left(\mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}} \right)^{-1} \\
 &\stackrel{(40)}{=} \mathbf{H}\mathbf{B}\mathbf{H}^T \left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R} \right)^{-1} \left(\mathbf{H}\mathbf{B}_t\mathbf{H}^T + \mathbf{R}_t \right) \left(\mathbf{H}\tilde{\mathbf{B}}\mathbf{H}^T + \tilde{\mathbf{R}} \right)^{-1} \\
 &\stackrel{(43)}{=} \mathbf{H}\mathbf{B}\mathbf{H}^T \left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R} \right)^{-1} \\
 &\stackrel{(2)}{=} \mathbf{H}\mathbf{K}
 \end{aligned}$$

From (1), (41), and (44) it follows that

$$\mathbf{H}\mathbf{x}^a = \mathbf{H}\tilde{\mathbf{x}}^a \quad (45)$$

and therefore, replacing the *status quo* model (\mathbf{B}, \mathbf{R}) by the *a posteriori* diagnosis model $(\tilde{\mathbf{B}}, \tilde{\mathbf{R}})$ has no impact on the observation-space representation of the analysis. In particular, if \mathbf{H} is the identity operator, $\mathbf{H} = \mathbf{I}$, then the covariance specification $(\tilde{\mathbf{B}}, \tilde{\mathbf{R}})$ has no impact on the analysis,

$$\mathbf{x}^a = \tilde{\mathbf{x}}^a \quad (46)$$

It is also noticed that the equivalence between (5) and (39) is established from the identities

$$\mathbf{I} - \mathbf{H}\mathbf{K} = \mathbf{R} \left(\mathbf{H}\mathbf{B}\mathbf{H}^T + \mathbf{R} \right)^{-1}; \quad (\mathbf{I} - \mathbf{H}\mathbf{K}_t)^{-1} \mathbf{R}_t = \mathbf{H}\mathbf{B}_t\mathbf{H}^T + \mathbf{R}_t \quad (47)$$

References

- Baker NL, Daley R (2000) Observation and background adjoint sensitivity in the adaptive observation-targeting problem. *Q J R Meteorol Soc* 126:1431–1454
- Bannister RN (2008a) A review of forecast error covariance statistics in atmospheric variational data assimilation. I: characteristics and measurements of forecast error covariances. *Q J R Meteorol Soc* 134:1951–1970
- Bannister RN (2008b) A review of forecast error covariance statistics in atmospheric variational data assimilation. II: modelling the forecast error covariance statistics. *Q J R Meteorol Soc* 134:1971–1996
- Bormann N, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. I: methods and applications to ATOVS data. *Q J R Meteorol Soc* 136:1036–1050
- Bormann N, Collard A, Bauer P (2010) Estimates of spatial and interchannel observation-error characteristics for current sounder radiances for numerical weather prediction. II: applications to AIRS and IASI data. *Q J R Meteorol Soc* 136:1051–1063
- Bormann N, Geer AJ, Bauer P (2011) Estimates of observation-error characteristics in clear and cloudy regions for microwave imager radiances from numerical weather prediction. *Q J R Meteorol Soc* 137:2014–2023

- Bouttier F, Derber J, Fisher M (1997) The 1997 revision of the Jb term in 3D/4D-Var. ECMWF Res Dept Tech Memo 238 (available from ECMWF, Reading UK)
- Brousseau P, Berre L, Bouttier F, Desroziers G (2012) Flow-dependent background-error covariances for a convective-scale data assimilation system. *Q J R Meteorol Soc* 138:310–322
- Buehner M (2005) Ensemble-derived stationary and flow-dependent background-error covariances: evaluation in a quasi-operational NWP setting. *Q J R Meteorol Soc* 131:1013–1043
- Campbell WF, Satterfield EA (2015) Accounting for correlated satellite observation error in NAVGEM. Presentation J9.1 to 11th annual symposium on new generation operational environmental satellite systems, 95th american meteorological society annual meeting, Phoenix, AZ. Recorded presentation. <https://ams.confex.com/ams/95Annual/webprogram/Paper257046.html>
- Clayton AM, Lorenc AC, Barker DM (2013) Operational implementation of a hybrid ensemble/4D-Var global data assimilation system at the Met Office. *Q J R Meteorol Soc* 139:1445–1561
- Collard AD (2007) Selection of IASI channels for use in numerical weather prediction. *Q J R Meteorol Soc* 133:1977–1991
- Daescu DN (2008) On the sensitivity equations of four-dimensional variational (4D-Var) data assimilation. *Mon Weather Rev* 136:3050–3065
- Daescu DN, Todling R (2010) Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Q J R Meteorol Soc* 136:2000–2012
- Daescu DN, Langland RH (2013a) Error covariance sensitivity and impact estimation with adjoint 4D-Var: theoretical aspects and first applications to NAVDAS-AR. *Q J R Meteorol Soc* 139:226–241
- Daescu DN, Langland RH (2013b) The adjoint sensitivity guidance to diagnosis and tuning of error covariance parameters. In: Park SK, Xu L (eds) Data assimilation for atmospheric, oceanic and hydrologic applications, vol II. Springer, pp 205–232
- Daescu DN, Langland RH (2013c) Adjoint estimation of the forecast impact of observation error correlations derived from a posteriori covariance diagnosis. Presentation 17.3 to 6th symposium on data assimilation, University of Maryland, College Park, MD. Webinar recording. http://das6.cscamm.umd.edu/program/das6_program.html
- Daescu DN, Navon IM (2013) Sensitivity analysis in nonlinear variational data assimilation: theoretical aspects and applications. In: Faragó I, Havasi Á, Zlatev Z (eds) Advanced numerical methods for complex environmental models: needs and availability, pp 276–300. eISBN:978-1-60805-778-8 ISBN:978-1-60805-777-1. Bentham Science
- Desroziers G, Berre L, Chapnik B, Poli P (2005) Diagnosis of observation, background, and analysis-error statistics in observation space. *Q J R Meteorol Soc* 131:3385–3396
- Garand L, Heilliette S, Buehner M (2007) Interchannel error correlation associated with AIRS radiance observations: inference and impact in data assimilation. *J Appl Meteorol* 46:714–725
- Gaspari G, Cohn SE (1999) Construction of correlation functions in two and three dimensions. *Q J R Meteorol Soc* 125:723–757
- Gneiting T (1999) Correlation functions for atmospheric data analysis. *Q J R Meteorol Soc* 125:2449–2464
- Goldberg MD, Qu Y, McMillin LM, Wolf W, Zhou L, Divakarla M (2003) AIRS near-real-time products and algorithms in support of operational numerical weather prediction. *IEEE Trans Geosci Remote Sens* 41(2):379–389
- Hogan TF, Liu M, Ridout JA, Peng MS, Whitcomb TR, Ruston BC, Reynolds CA, Eckermann SD, Moskaitis JR, Baker NL, McCormack JP, Viner KC, McLay JG, Flatau MK, Xu L, Chen C, Chang SW (2014) The navy global environmental model. *Oceanography* 27(3):116–125
- Kuhl DD, Rosmond TE, Bishop CH, McLay J, Baker NL (2013) Comparison of hybrid ensemble/4DVar and 4Dvar within the NAVDAS-AR data assimilation framework. *Mon Weather Rev* 141:2740–2758
- Lahoz W (2010) Research satellites. In: Lahoz W, Khattatov B, Ménard R (eds) Data assimilation: making sense of observations. Springer, pp 301–321
- Langland RH, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus* 56A:189–201

- Liu J, Kalnay E (2008) Estimating observation impact without adjoint model in an ensemble Kalman filter. *Q J R Meteorol Soc* 134:1327–1335
- Liu J, Kalnay E, Miyoshi T, Cardinali C (2009) Analysis sensitivity calculation in an ensemble Kalman filter. *Q J R Meteorol Soc* 135:1842–1851
- Lorenc AC, Bowler NE, Clayton AM, Pring SR, Fairbairn D (2015) Comparison of hybrid-4DEnVar and hybrid-4DVar data assimilation methods for global NWP. *Mon Weather Rev* 143:212–229
- Lorenz EN, Emanuel KA (1998) Optimal sites for supplementary weather observations: simulation with a small model. *J Atmos Sci* 55:399–414
- Lupu C, Cardinali C, McNally AP (2015) Adjoint-based forecast sensitivity applied to observation error variances tuning. *Q J R Meteorol Soc* 141:3157–3165. doi:[10.1002/qj.2599](https://doi.org/10.1002/qj.2599)
- McNally AP, Watts PD, Smith JA, Engelen R, Kelly GA, Thépaut JN, Matricardi M (2006) The assimilation of AIRS radiance data at ECMWF. *Q J R Meteorol Soc* 132:935–957
- Pauley PM (2003) Supperobbing satellite winds for NAVDAS. Technical report. NRL/MR/7530-03-8670, Naval Research Laboratory, Monterey, CA, p 102
- Rosmond T, Xu L (2006) Development of NAVDAS-AR: non-linear formulation and outer loop tests. *Tellus* 58A:45–58
- Stewart LM, Dance SL, Nichols NK (2013) Data assimilation with correlated observation errors: experiments with a 1-D shallow water model. *Tellus A* 65:19546
- Stewart LM, Dance SL, Nichols NK, Eyre JR, Cameron J (2014) Estimating interchannel observation-error correlations for IASI radiance data in the Met Office system. *Q J R Meteorol Soc* 140:1236–1244
- Talagrand O (1999) A posteriori evaluation and verification of the analysis and assimilation algorithms. In: proceedings of workshop on diagnosis of data assimilation systems. ECMWF, Reading, UK, pp 17–28
- Thépaut JN, Andersson E (2010) The global observing system. In: Lahoz W, Khattatov B, Ménard R (eds) Data assimilation: making sense of observations. Springer, pp 263–281
- Todling R (2015) A complementary note to ‘A lag-1 smoother approach to system-error estimation’: the intrinsic limitations of residual diagnostics. *Q J R Meteorol Soc* 141:2917–2922
- Waller JA, Dance SL, Lawless AS, Nichols NK (2014) Estimating correlated observation error statistics using an ensemble transform Kalman filter. *Tellus A* 66:23294
- Wang X, Parrish D, Kleist D, Whitaker J (2013) GSI 3DVar-based ensemble-variational hybrid data assimilation for NCEP global forecast system: single-resolution experiments. *Mon Weather Rev* 141:4098–4117
- Weston PP, Bell W, Eyre JR (2014) Accounting for correlated error in the assimilation of high-resolution sounder data. *Q J R Meteorol Soc* 140:2420–2429
- Xu L, Rosmond T, Daley R (2005) Development of NAVDAS-AR: formulation and initial tests of the linear problem. *Tellus* 57A:546–559

GPS PWV Assimilation with the JMA Nonhydrostatic 4DVAR and Cloud Resolving Ensemble Forecast for the 2008 August Tokyo Metropolitan Area Local Heavy Rainfalls

Kazuo Saito, Yoshinori Shoji, Seiji Origuchi and Le Duc

Abstract On 5th August 2008, scattering local heavy rainfalls occurred at various places over the Tokyo metropolitan area, and five drainage workers were claimed by an abrupt increase of water level. The Japan Meteorological Agency (JMA) operational mesoscale model of the day failed to predict occurrence of the local heavy rainfalls, which were brought about by deep convective cells developed on the unstable atmospheric condition without strong synoptic/orographic forcings. A 11-member mesoscale ensemble prediction with a horizontal resolution of 10 km was conducted using the operational mesoscale analysis of JMA and perturbations of the JMA global one-week ensemble prediction system as the initial condition and the initial and lateral boundary perturbations, but the intense rains exceeding 20 mm/3 h were hardly predicted. A downscaling ensemble forecast experiment with a horizontal resolution of 2 km was conducted using the 6 h forecast of the 10 km ensemble as the initial and boundary conditions. Scattered intense rains were predicted in some ensemble members, but their locations and distribution were insufficient. The total precipitable water vapor (PWV) observed by the GNSS Earth Observation Network System (GEONET) of Geospatial Information Authority of Japan showed that the JMA mesoscale analysis given by the hydrostatic Meso-4DVAR underestimated water vapor over the Tokyo metropolitan area. To modify the initial condition, a reanalysis data assimilation experiment was conducted with the JMA's nonhydrostatic 4DVAR (JNoVA), where PWV data from GEONET were assimilated 2.5 days with 3-h data assimilation cycles. The 2 km downscale ensemble run from the JNoVA analysis properly predicted the areas of

K. Saito (✉) · Y. Shoji · S. Origuchi · L. Duc
Meteorological Research Institute, Tsukuba, Japan
e-mail: ksaito@mri-jma.go.jp

L. Duc
Japan Agency for Marine-Earth Science and Technology,
Yokohama, Japan

scattering local heavy rains. Threat scores and ROC area skill scores suggest that even in the ensemble prediction, accuracy of initial condition is critical to numerically predict small scale convective rains. Fractions skill scores indicated the value of the cloud resolving ensemble forecast for such the unforced convective rain case.

1 Introduction

Accuracy of quantitative precipitation forecast (QPF) in mesoscale numerical weather prediction (NWP) has been remarkably improved recently by virtue of progress of NWP technology, increase of observation data and the computer power, however, absolute forecast accuracy of the heavy rainfalls that lead to the disaster is still insufficient (Saito 2012). Although the precipitation that occurs along with the orographic and/or synoptic scale forcing has been becoming predictable, the accurate forecast of local heavy rainfalls without large scale forcing is still very difficult. Since the temporal and spatial scales of local heavy rain is small, a high resolution numerical model that can express complex physics inside the deep convective cloud commensurate with the phenomena is required. Besides, since such the phenomena occur in the unstable air condition, small differences of the initial condition sometimes cause a large difference in the forecast, it is necessary to determine the initial value with high accuracy.

It is well-known that the accuracy of the water vapor field in the lower level atmosphere is particularly important for the forecast of heavy rainfalls. The Global Navigation Satellite System (GNSS) as typified by the U.S Global Positioning System (GPS) is a relatively new observation tool to estimate water vapor in the atmosphere. It gives information on water vapors along the slant pass between the GNSS satellites and ground receivers, and the total precipitable water vapor (PWV) in high accuracy (Bevis et al. 1994). Shoji et al. (2009) conducted data assimilation experiments of the PWV derived from the nationwide ground GPS network (GNSS Earth Observation Network: GEONET) for a heavy rainfall event on 28th July 2008 in western Japan. They assimilated GEONET and global (East Asia) GPS PWV data to improve the initial condition of the numerical model, and demonstrated that underestimation of low level humidity and the positional error of the low-level convergence zone over the Sea of Japan in the analysis were reduced by the PWV data and the forecast of the observed rainfalls in western Japan was significantly improved. The precipitation system was developed along the convergence zone over the Sea of Japan, and the operational mesoscale model (MSM) of the Japan Meteorological Agency (JMA) with a horizontal resolution of 5 km was used. Shoji et al. (2011) conducted mesoscale data assimilation experiment of Myanmar cyclone Nargis using GPS-derived PWV over the Bay of Bengal, and showed the availability of the PWV data to improve the intensity forecast of the tropical cyclone. For both

studies, the JMA's hydrostatic mesoscale 4 dimensional variational assimilation system (Meso 4D-Var; Ishikawa and Koizumi 2002) were used for data assimilation.

As mentioned previously, the forecast of local heavy rainfalls without large scale forcing is difficult. Kawabata et al. (2007) was the first to apply a cloud resolving 4D-Var to reproduce a local convective heavy rainfall developed over the Tokyo metropolitan area with the GPS-derived PWV assimilation. With Kawabata et al. (2011), they demonstrated the availability of high resolution data assimilation and GPS-derived PWV data for the prediction of mesoscale convective systems and the associated local heavy rainfalls, but the lead times were about 1 h for both cases.

In general, deterministic prediction exceeding the time scale of the phenomena (less than a few hours in the local heavy rainfall) is essentially impossible, thus for prediction with a certain amount of lead time, the ensemble forecast is necessary. Recently Duc et al. (2013) conducted verification of high resolution ensemble forecasts using the JMA nonhydrostatic model with horizontal resolutions of 10 and 2 km. They verified 15 rainfall cases in July 2010 and indicated that the cloud resolving resolution (2 km) is necessary to predict intense rains even in the ensemble forecast. Their study was a statistical verification for several precipitation events and data assimilation trial for specific case was not conducted.

A typical example of convective rains that occur in the unstable atmosphere with weak large-scale forcing is a local heavy rainfall event in the Tokyo metropolitan area on 5th August, 2008. As mentioned in the next section, several convection cells developed sporadically in the area of Tokyo and its surroundings. A tragic accident where five underground drainage workers in sewer construction in Toshima, Tokyo were killed has occurred, and it became a trigger that the term “guerrilla heavy rain” was socially used. Ogura (2009), citing this case, pointed out the necessity to distribute a probability diagram of heavy rainfall for a certain threshold 6–12 h before their occurrence.

In this study, we conduct GPS-derived PWV assimilation experiments and ensemble forecast experiments using the JMA nonhydrostatic model and its 4D-VAR data assimilation system (JNoVA). We show the importance to improve water vapor field in the numerical model initial condition, and discuss the probability prediction of such local heavy rainfalls based on the ensemble forecast. In Sect. 2, the local heavy rainfall event on 5th August 2008 is presented as a target case. In Sect. 3, forecasts from operational mesoscale analysis of JMA are shown for both deterministic and ensemble forecast as the reference. In Sect. 4, we compare PWV of the JMA mesoscale analysis and GEONET, and conduct data assimilation of PWV using JNoVA. Forecast results from the GPS assimilated analysis and ensemble prediction are shown. Verification of the forecasts and the probabilistic prediction of the local heavy rains are shown, with the discussion of predictability of the intense convective rains based on the fraction skill score which consider the positional lags. Summary and concluding remarks are given in Sect. 6.

2 Tokyo Metropolitan Area Local Heavy Rainfalls on 5 August 2008

The summer of 2008 of Japan was characterized by several local heavy rainfalls. One of the events occurred in the Tokyo metropolitan area on 5th August 2008 along with the unstable atmospheric condition. Figure 1 shows the total precipitation amount on the day in Tokyo, where 134 mm was recorded at Toshima, and 111.5 mm at Otemachi. In Toshima-ward, an accident occurred where five underground drainage workers in the sewer construction were flowed by an abrupt increase of the water level around 1215 JST. JMA issued a heavy rain flood warning at 1233 JST, but it was 18 min after the accident.

Figure 2 is the surface weather map at 0900 JST, 5th August. A stationary front is seen in the southern part of the Kanto Plain and a low pressure area was located on the far south coast of Japan, but there were no clear disturbance in the Tokyo metropolitan area.

Figure 3a shows accumulated 3 h precipitation from 1200 to 1500 JST by the JMA precipitation analysis based on radar and surface rain-gauge networks. Several intense rainfall cells were scattered over a wide areas from the southern Kanto Plain (Tokyo metropolitan area) to Shizuoka Prefecture. Ishihara (2013) performed the analysis based on the reflection intensity data obtained by the 10 min interval meteorological radars of JMA, and reported that 179 convection cells developed in the Tokyo metropolitan area and its surroundings (about 140 km square) from 0900

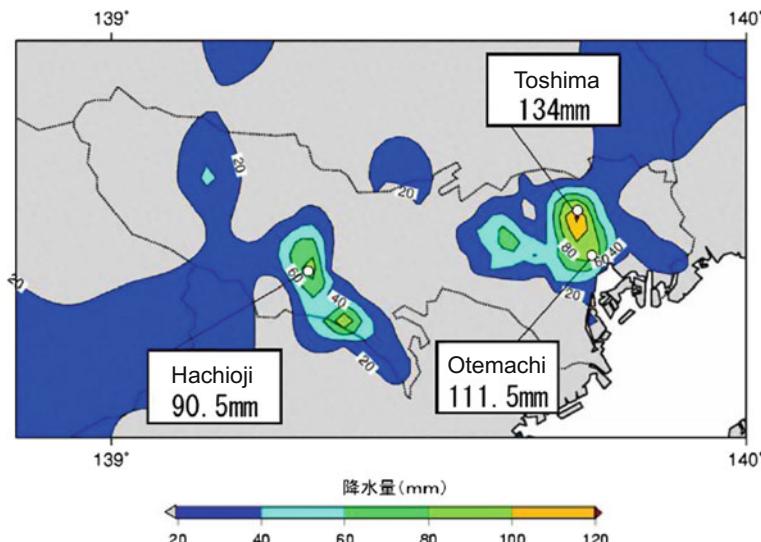


Fig. 1 Total precipitation amount on 5th August 2008. Reproduced from Tokyo Regional Headquarters (2008)

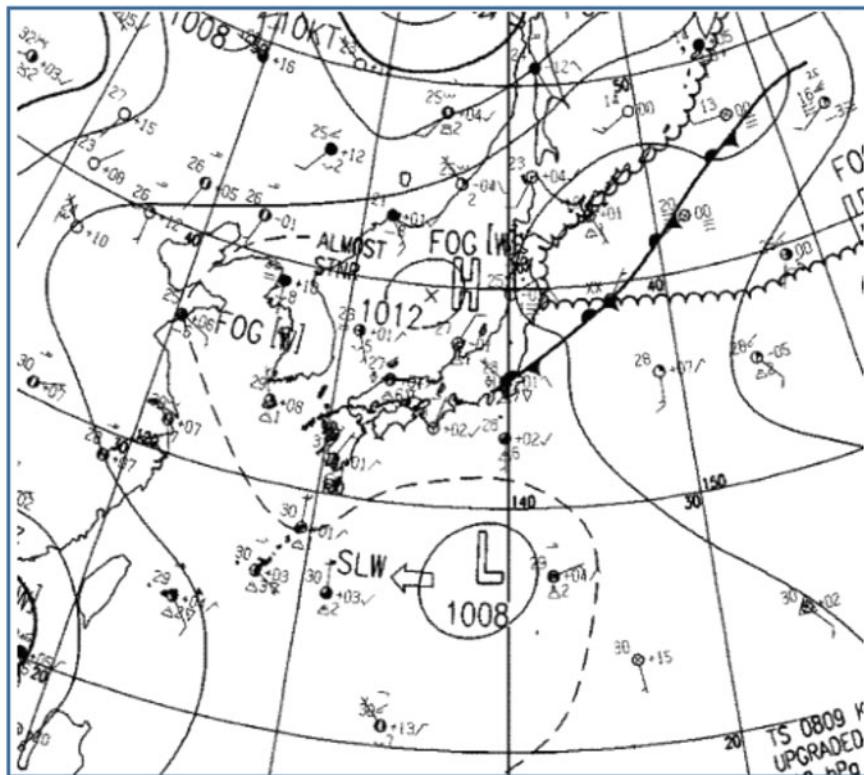


Fig. 2 Surface weather map at 0900 JST, 5th August 2008

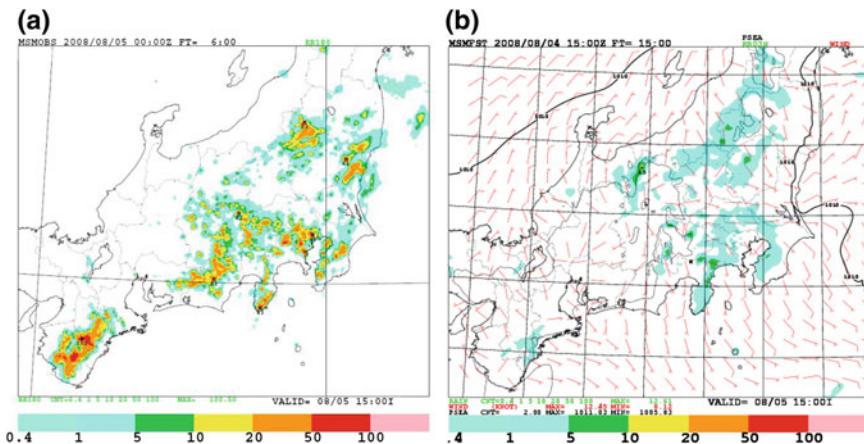


Fig. 3 a Accumulated 3 h precipitation from 1200 to 1500 JST on August 5, 2008 by JMA precipitation analysis. b Same as in (a) but 18 h forecast of the Mesoscale model (MSM) of JMA. Initial time is 1500 UTC (2400JST), 4th August

JST to 1800 JST. Note that almost no rains appeared in the northern part of the Kanto Plain. Figure 3b shows the corresponding 3 h precipitation from 1200 to 1500 JST by the 18 h forecast of the JMA operational mesoscale model (MSM), whose initial time was 1500 UTC (2400 JST), 4th August 4. Weak rain areas are forecasted in some places in the south of the Kanto Plain and the Joetsu mountainous region (north of Kanto), but most of them were less than 5 mm/3 h and the model failed to predict the occurrence of strong rains in Tokyo and its surrounding areas.

3 Forecast from Operational Mesoscale Analysis of JMA

3.1 Downscale Forecast

First, we conducted simple downscale experiments with the JMA nonhydrostatic model (NHM; Saito et al. 2006, 2007; Saito 2012) from the JMA operational mesoscale analysis (MA). Specifications of the models and analyses employed in this study are listed in Table 1. Here, specifications of the operational systems are as of August 2008.

Figure 4a shows 3-h precipitation from 1200 to 1500 JST on August 5 by NHM with a horizontal resolution of 10 km (NHM10). Initial and boundary condition is the same that in the operational MSM forecast, i.e., the mesoscale analysis (MA) at 1500 UTC (2400 JST), 4th August and the global model (GSM) forecast of JMA. Physical processes (3-ice bulk cloud microphysics and Kain-Fritsch convective parameterization) and vertical levels of NHM10 are the same as in MSM, except for the horizontal resolution. In Fig. 4a, weak rains over 10 mm/3 h is seen in Kanagawa Prefecture (west of Tokyo), which is slightly different from the MSM forecast (Fig. 3b), but intense precipitation is not predicted.

Figure 4b shows 12 h forecast of the 2 km resolution NHM (NHM2) with the initial time of 1800 UTC, August 4th (0300 JST, 5th). In this experiment, 6 h forecast of NHM10 and the subsequent forecasts of NHM10 were used for initial and boundary conditions of NHM2, respectively. Computational domain of NHM2 is the region of the 1600x1100 km which covers southern part of Japan, and is the same that used in the JMA operational local forecast model (LFM) from November 2010 to August 2012. Convective parameterization is not used in NHM2. Precipitations shown in Fig. 4b are divided into small areas, and intense convection cells are scattered. These characteristic features of predicted rains by NHM2 is closer to observation (Fig. 3a) than MSM and NHM10, however, most of the strong rainfalls are limited over the area west of Tokyo, probably due to the influence of the “parent model” NHM10, which predicted moderate rain only in the west of Tokyo.

Table 1 Specifications of the models and analyses used in this study (JMA model and analyses are as of August 2008)

	JMA global spectral model	JMA one week ensemble prediction	JMA mesoscale model	JMA nonhydrostatic model	2 km resolution JMA nonhydrostatic model	JMA mesoscale analysis (hydrostatic)	Nonhydrostatic 4DVAR
Abbreviation	GSM	WEP	MSM	NHM10	NHM2	MA	JNvOA
Horizontal resolution	20 km	60 km	5 km	10 km	2 km	10 km ^a	5 km ^b
Levels	60	60	50	50	60	40	50
Domain	global	global	Japan and its surrounding areas (3600 × 2880 km)	Japan and its surrounding areas (3600 × 2880 km)	Southern part of Japan (1600 × 1100 km)	Japan and its surrounding areas (3600 × 2880 km)	Japan and its surrounding areas (3600 × 2880 km)
Dynamics	Hydrostatic Spectral	Hydrostatic Spectral	Nonhydrostatic	Nonhydrostatic	Nonhydrostatic	Hydrostatic Spectral	Nonhydrostatic
Precipitation	Large scale condensation	Large scale condensation	Bulk method 3-ice cloud microphysics	Bulk method 3-ice cloud microphysics	Bulk method 3-ice cloud microphysics	Large scale condensation	Bulk method 3-ice cloud microphysics ^c
Cumulus convection	Arakawa-Schubert	Arakawa-Schubert	Kain-Fritsch	Kain-Fritsch	non	Moist convective adjustment	Kain-Fritsch ^d

^a 20 km for the inner model in data assimilation

^b 15 km for the inner mode

^c Large scale condensation for the inner model

^d No cumulus convection for the inner model

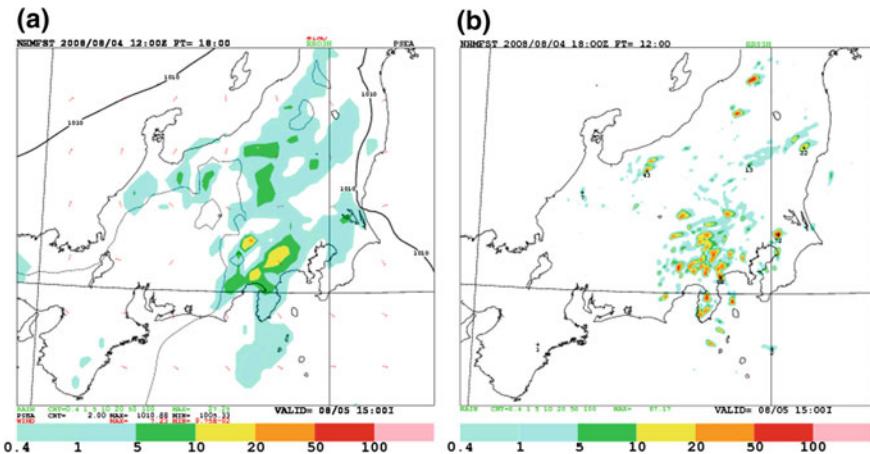


Fig. 4 **a** Same as in Fig. 3b but by the forecast of 10 km resolution NHM. **b** Same as (a) but 12 h forecast of 2 km resolution NHM with the initial time of 1800 UTC, August 4th (0300 JST, August 5th)

3.2 Ensemble Forecast

To see whether the positional lags and intensity errors of the precipitation by the downscaling experiments can be corrected by considering the uncertainty of the initial value or boundary values, we conducted the mesoscale ensemble prediction experiments using the downscale experiment described in the previous section as the control run.

First, as a 10 km resolution mesoscale ensemble forecast was conducted where the initial and boundary perturbations of NHM10 were given by the JMA one-week ensemble prediction system (EPS). The same procedure were tested at MRI in the comparison of mesoscale ensemble prediction systems conducted in the WWRP (world weather research program) Beijing Olympic 2008 Research and Development Projects (B08RDP; Duan et al. 2012). The ensemble using the JMA one-week EPS corresponds to the ‘WEP’ system, and the detailed procedures such as the normalization of perturbations are described in Saito et al. (2010, 2011). As a cloud resolving ensemble forecast, 2 km down scaling was conducted with NHM2 using the 6-h forecast of the 10 km mesoscale ensemble prediction as the initial and boundary conditions. Duc et al. (2013) conducted similar downscale ensemble experiments with NHM10 and NHM2 targeting on 15 rainfall cases in central Japan in July 2010 and made verifications.

Upper panels of Fig. 5 shows 3-h precipitation from 1200 JST to 1500 JST by each ensemble member of the 10 km mesoscale ensemble prediction. Although some of the members (p02, p05, m01, etc.) express the somewhat strong rainfalls in the Tokyo metropolitan area, no intense rains are predicted in Shizuoka prefecture

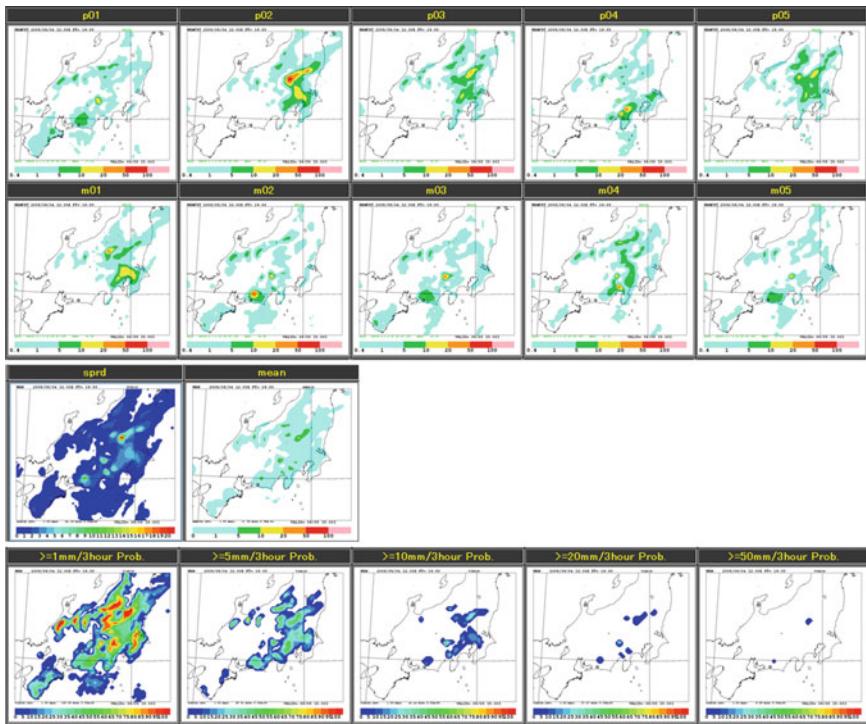


Fig. 5 *Top* Three-hour precipitation from 1200 JST to 1500 JST by each ensemble member of the 10 km mesoscale ensemble. Initial time is 1200 UTC (2100 JST) August 4. *Middle* Ensemble spread (left) and the ensemble mean (right). *Bottom* Probability forecast of 3-h precipitation for different thresholds. From the left, 1, 5, 10, 20, and 50 mm

in such the members. Middle panels of Fig. 5 show ensemble spread (left) and the ensemble mean (right). Reflecting that there are no members forecasting strong rains, no large ensemble spread is seen in the Tokyo metropolitan area. The rainfall intensity in the ensemble mean is weak. Lower panels of Fig. 5 show the probability forecast of the 3 h precipitation for different thresholds based on the NHM10 ensemble forecast. Probability of occurrence of precipitation is shown over the Tokyo metropolitan area up to 10 mm threshold, while occurrence is not shown for rainfall intensity larger than 20 mm. Note that rains are predicted over northern part of the Kanto plain, where rains were not observed actually. These results suggest that ensemble prediction with the horizontal resolution of 10 km is not useful in the viewpoint of disaster prevention for local heavy rainfall case in this study. As suggested in the operational MSM forecast (Fig. 3b), a similar result is expected in the ensemble forecast even with a horizontal resolution of 5 km if the cumulus parameterization is employed.

As the cloud resolving experiment, downscaling ensemble forecast with a horizontal resolution of 2 km was carried out. The relationship of the 10 km and the

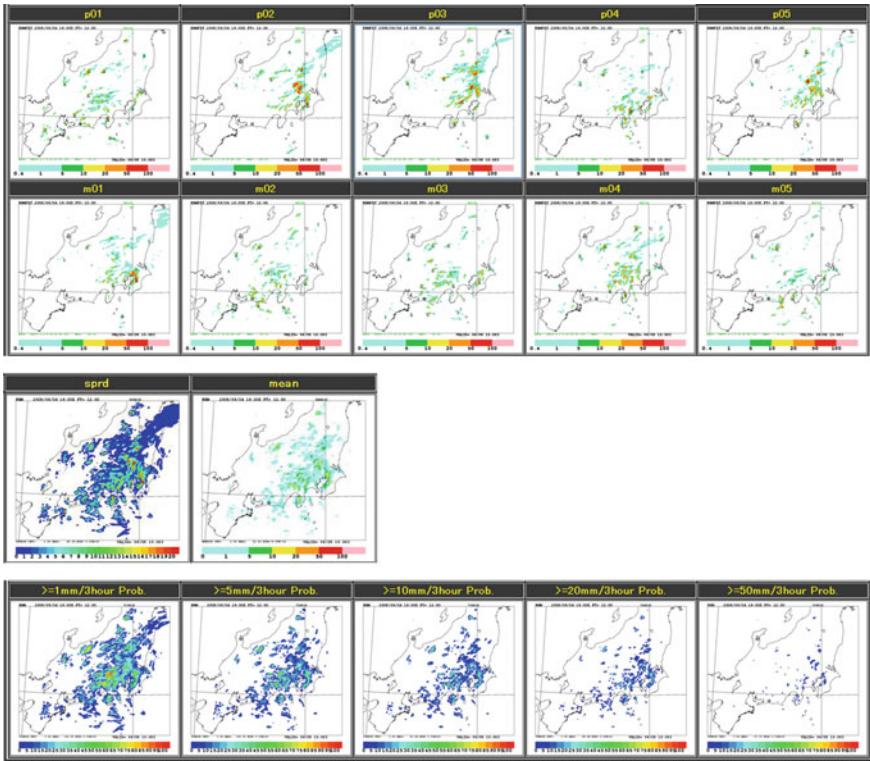


Fig. 6 Same as in Fig. 5 but forecast by 2 km cloud resolving ensemble prediction. Initial time is 1800 UTC, 4th (0300 JST, 5th) August 2008

2 km ensemble forecasts is shown in Fig. 8, which will be described later (Sect. 4.2). Figure 6 shows the 3 h precipitation of each member, ensemble spread and the ensemble mean, and the probability prediction of 3 h rainfalls for the different thresholds by the 12 h forecast of the 2 km ensemble forecast. Reflecting the forecast characteristics of NHM2 shown in the Fig. 4b, scattered strong convection cells are seen in the forecast of each ensemble member (the upper panels). In some members (e.g., p02, p05, and m01), very intense rainfalls more than 50 mm are seen in the Tokyo metropolitan area, corresponding to the members in the 10 km ensemble that predicted somewhat strong precipitations in the same area. Ensemble spread becomes larger over a wide area from the Tokyo metropolitan area to Shizuoka prefecture, while the ensemble mean is still small (less than 20 mm in the maximum) because the strong precipitation cells of each member are predicted at different positions.

In the probability forecast distribution of different threshold values shown in the lower panels of Fig. 6, probability more than 20 % is seen for a threshold of 20 mm/3 h over the Tokyo metropolitan area, where the heavy rain was not predicted by the control run (Fig. 4b). This result suggest the usefulness of the cloud

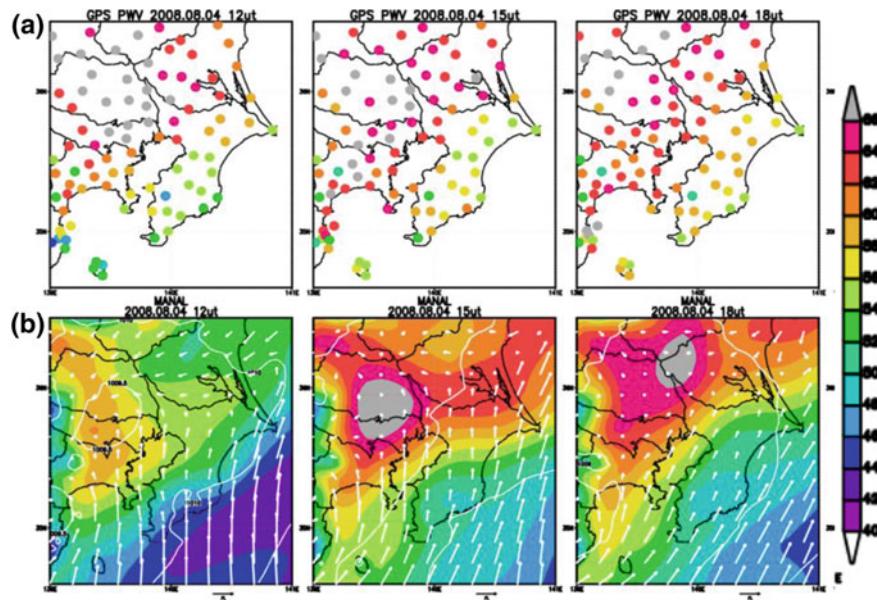


Fig. 7 **a** PWV observed by GEONET at 1200 (left), 1500 (middle) and 1800 (right) UTC, 4th (2100, 2400 JST 4th and 0300 JST 5th) August 2008. **b** Same as in (a) but PWV by the Meso4DVAR analysis

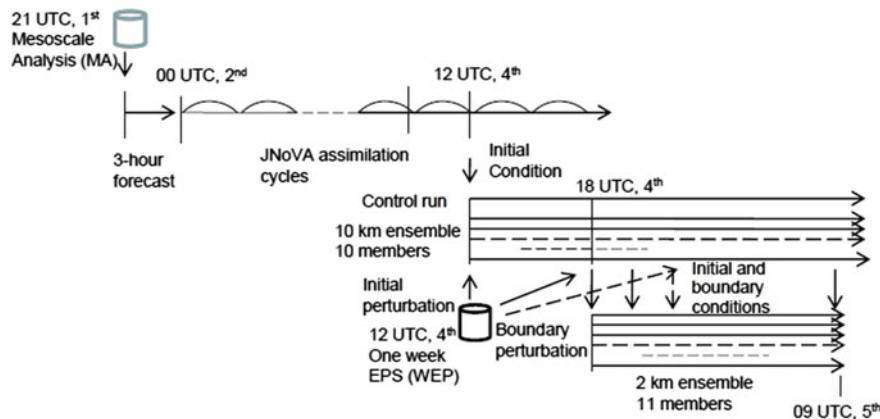


Fig. 8 Schematic chart of the JNoVA data assimilation cycles and the NHM ensemble forecast

resolving ensemble forecast to some extent, however, probability of very strong rains over 50 mm is less than 10 % at the maximum and the development is limited to a very narrow scattered areas. This means that the simple downscale ensemble prediction is not sufficient in terms of the disaster prevention.

We also conducted another downscale ensemble prediction with NHM10 and NHM2 using the mesoscale analysis of August 4 1500 UTC (2400 JST) as the initial condition of the control run of NHM10, but satisfactory results were not obtained because a fake strong rain area was predicted in the east of Tokyo (figure not shown).

4 Analysis and Assimilation of GPS PWV

4.1 Comparison of PWV of Mesoscale Analysis and GPS

Forecasts of NHM10 from the operational mesoscale analysis (MA) at 1200 UTC (2100 JST), August 4th could not reproduce the local heavy rains in the Tokyo metropolitan area, including the mesoscale ensemble prediction which uses it as the control run. Cloud resolving ensemble forecast by NHM2 showed some degree of usefulness to predict intense rains, but was still not sufficient for very strong rains. One of the causes of these insufficiency was in the initial field of water vapor given by MA. Figure 7 compares PWV observed by GEONET of the Geospatial Information Authority of Japan (upper) and MA (bottom) at 1200 (left), 1500 (middle) and 1800 (right) UTC, 4th (2100, 2400 JST 4th and 0300 JST 5th) August 2008. In the GEONET PWV of 1200 UTC (upper left), several GEONET stations observed a large PWV values more than 66 mm (gray) from Tokyo to Saitama Prefecture, but PWV of MA in this area at the same time (lower left) is quite underestimated. In the subsequent times (1500 and 1800 UTC), MA analyzed the large PWV areas to the east while PWVs over the Boso and the Miura peninsulas and coastal region facing to Pacific Ocean are underestimated.

4.2 Data Assimilation of GPS PWV

As mentioned in the above section, the operational mesoscale analysis of JMA underestimated PWV around the Tokyo metropolitan area at 1200 UTC (2100 JST) 4th August. To improve the initial field of the NHM forecast, a JMA non-hydrostatic model based four-dimensional variational method (JNoVA; Honda et al. 2005; Honda and Sawada 2009) was adopted to the assimilation experiment of GEONET PWV. Figure 8 illustrates a schematic chart of the JNoVA data assimilation cycles and the subsequent NHM ensemble forecasts. Three days assimilation cycles were performed from 0000 UTC (0900 JST) of August 2nd with 3 h assimilation windows up to 1800 UTC of 4th (0300 JST of 5th), August. The first guess field at the beginning of the data assimilation cycle was given by a 3 h forecast of 5 km NHM from 2100 UTC, 1st (0600 JST 2nd) August.

The following two types of initial value analyses were tested:

(1) JNoVA_noGPS:

Analysis by JNoVA where GEONET PWV data are not assimilated.

(2) JNoVAg2:

Those that assimilate GPS PWV in the analysis by JNoVA. Data only altitude difference is less than 200 m in the actual elevation and the model terrain and at observation points below 500 m elevation are used for assimilation. Observation error of GPS PWV is set to 3 mm. For more information about the real time analysis and assimilation details of GEONET PWV data, see Shoji (2009) and Shoji et al. (2009). The procedures are almost the same as the assimilation of GPS PWV data at JMA's operational system from April 2009, while the following points are different:

- i. the thinning interval in this study is set to 15 km compared to 30 km of the operational version.
- ii. the GPS PWV data in the precipitation area of more than 1.5 mm/h are not used in the operational JNoVA, but these data are assimilated in this study.

Actually, differences in the analysis results by above modifications was small.

Figure 9 shows the PWV from 1200 to 1800 UTC (2100 to 0300 JST) obtained by the two JNoVA experiments. Magnitudes of PWV in the JNoVA analyses are generally larger than those of MA even without GPS data (upper panels). At 1500

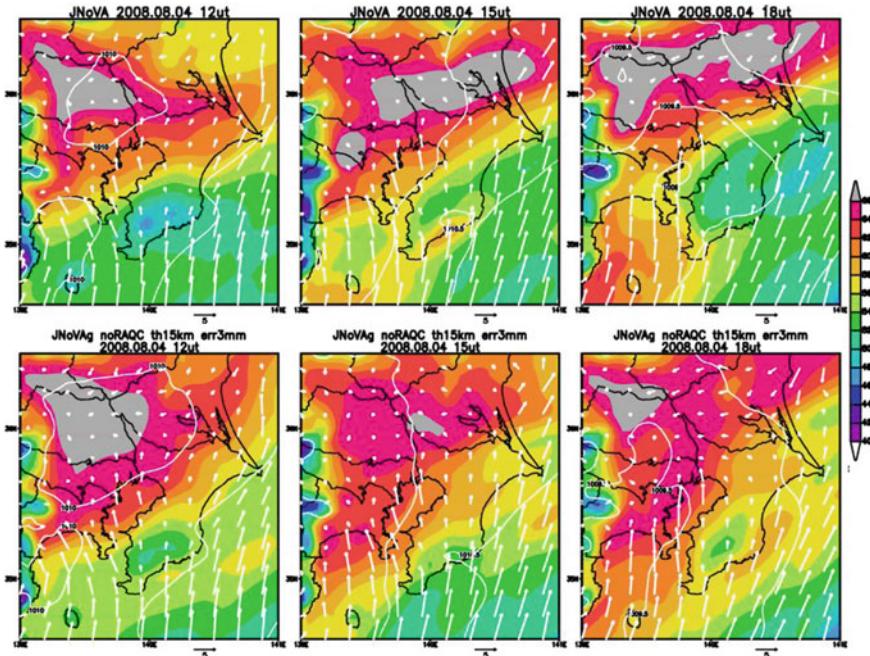


Fig. 9 Same as in Fig. 7 but PWV by the JNoVA analysis. *Upper* without GPS (JnoVA_noGPS), *Bottom* with GPS (JNoVAg2)

UTC (2400 JST) and 1800 UTC (0300JST), PWVs in JNoVA_NoGPS are rather overestimated compared to the PWV values in GPS observations. In JNoVAg2, PWV of analysis values are closer to the observed ones as a matter of course. A large PWV area is spread over the Saitama Prefecture, north of Tokyo, at 1200 UTC (2100 JST), while overestimation at 1500 and 1800 UTC (2400 and 0300 JST) is reduced. On the other hand, the water vapor increases in the south coastal area, the windward side of the Tokyo Metropolitan area.

4.3 Free Forecast from the GPS Assimilated Analysis and Ensemble Prediction

Forecast and ensemble experiments with NHM10 and NHM2 were conducted using the analysis by JNoVAg2 at 1200 UTC (2100 JST), 4th August (see Fig. 8 for design of experiments), similar to those using MA in the previous Sects. 3.1 and 3.2. Figure 10 shows the 3 h rainfall from 1200 to 1500 JST, 5th August by the

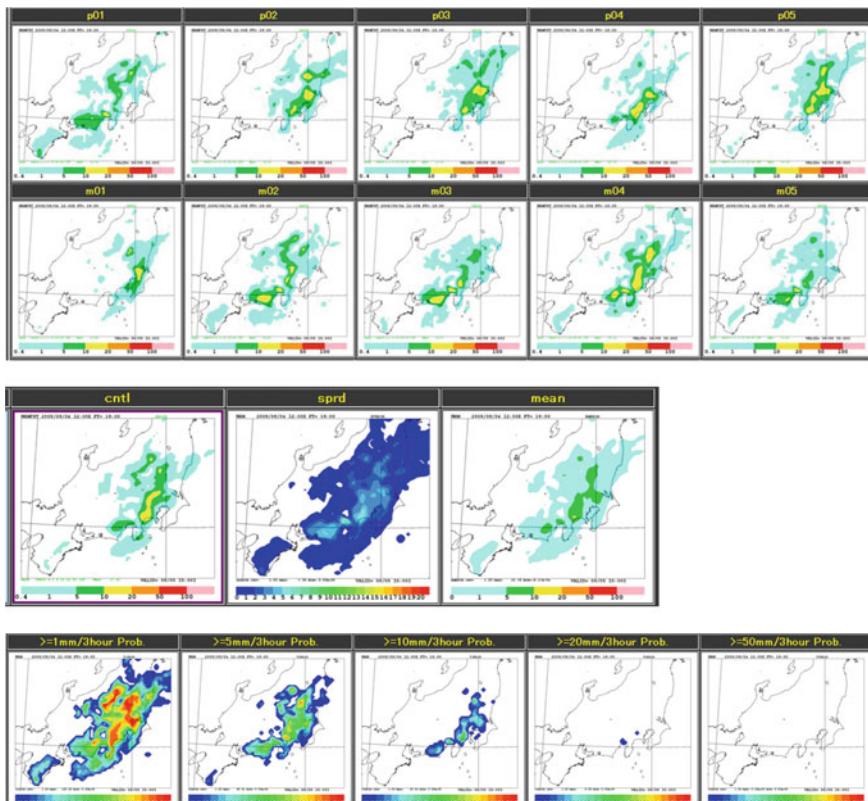


Fig. 10 Same as in Fig. 5, except ensemble forecast from JNOVA analysis with GPS PWV data (JNoVAg2). Control forecast without perturbation is added to left of the *middle* panels

control run and the ensemble forecast by NHM10. In the control forecast, which is shown in the middle left of Fig. 10, moderate strong precipitation ranging 10–20 mm/3 h is spread over more distinctively from the western Tokyo to the eastern part of Shizuoka Prefecture, compare to the corresponding forecast from MA (Fig. 4a). The similar intense rains are detected by many ensemble members which perturbed from the JNoVAg2 control run. The areas of the large ensemble spread correspondingly shifted to the east compared to the 10 km ensemble from MA (Fig. 5). In the figures of probability prediction (bottom), precipitation areas are well reproduced for the threshold of 10 mm/3 h but those more than 20 mm/3 h are not forecasted. This result suggests that the model resolution is intrinsically important for expression of the local convective intense rains appeared in this case.

Figure 11 shows the corresponding 3 h precipitation for the same valid times by the downscale forecast and the cloud resolving ensemble forecast by NHM2 using the JNoVAg2 analysis. Heavy rain area in the control run is mostly seen in the Kanagawa Prefecture, but most of the ensemble members predicted strong rains in

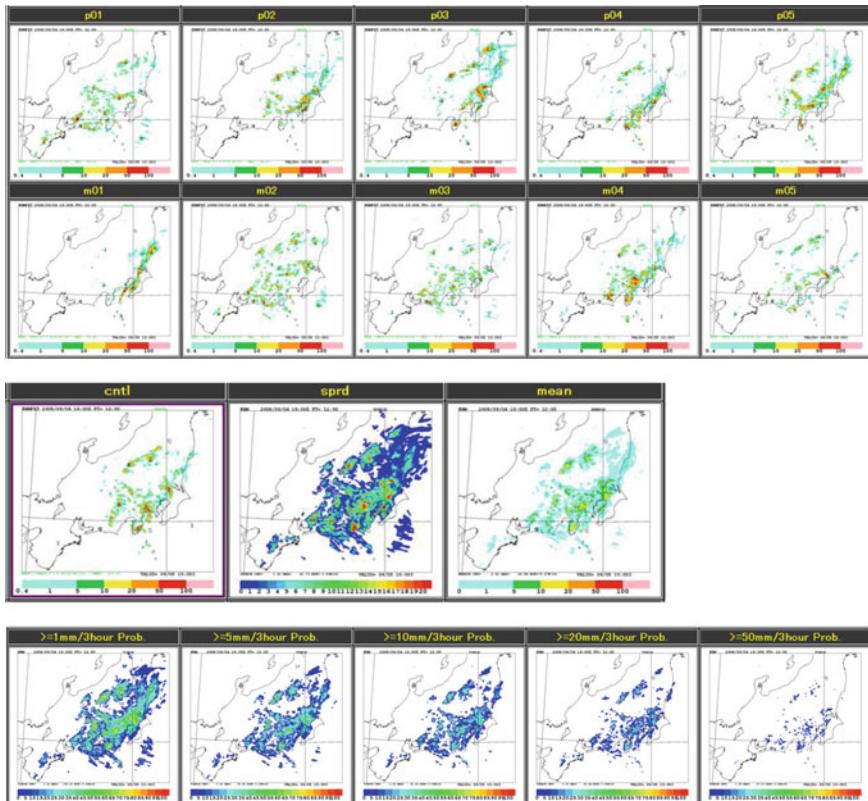


Fig. 11 Same as in Fig. 6, except ensemble forecast from JNOVA analysis with GPS PWV data (JNoVAg2). Control forecast without perturbation is added to left of the *middle* panels

the Tokyo metropolitan area, and strong rains are also found in the Shizuoka Prefecture in some members. A noteworthy characteristic feature in this ensemble forecast is that almost no precipitation is predicted by any members at northern Kanto where precipitation was not observed (Fig. 3a). This feature also reflected in the ensemble spread and the ensemble mean (middle panels of Fig. 11) and the probability forecast distribution diagrams (bottom) for the thresholds of 5, 10 and 20 mm/3 h. Positions of strong precipitation cells predicted by the individual ensemble members are slightly different each other. That is, convective precipitation of this day occurred without the specific forcing and the occurrence of convection was sporadic. Reflecting the spatial lags in the positions of individual cells, the probability values are generally less than 20 % for 20 mm/3 h, however, the areas where the probability of intense rain is predicted is very well corresponding to the locations where strong rains by convective cells were actually observed. Probability of 50 mm/3 h rains is sporadic and less than 10 %. The value is not high, but the risk of occurrence is certainly predicted from the Tokyo metropolitan area to the Shizuoka prefecture.

5 Verification

5.1 Verification of Ensemble Prediction

We performed verification of the ensemble forecasts by NHM10 and NHM2 from the JMA mesoscale analysis and nonhydrostatic 4DVAR analysis with GPS PWV described in the previous section. Figure 12a shows threat scores for 3 h precipitation by the control run and the ensemble mean using NHM10. Here, verification grid size is 5 km, and the average for 0600–1800 is indicated. The domain of verification is a central Honshu area that is indicated by Fig. 3. Compared to the forecast when the initial condition is given by the operational mesoscale analysis (Meso4dvar_ctl), the threat score of the forecast from the initial condition with GPS PWV (JNoVA_ctl) is much improved especially at the thresholds less than 10 mm/3 h. One of the interesting things in this figure is that, the ensemble means of the two ensembles (Meso4dvar_mean and JNoVA_mean) only improved the threat scores of the control forecasts for the weaker precipitation below 3 mm/3 h and for rains over 5 mm/3 h the score are even worse than those of the control runs. Threat scores for intense rains larger than 15 mm/3 h are extremely low in the ensemble means.

Figure 12b is the case for the 2 km NHM. In the forecast which assimilated the GPS PWV, the control forecast (JNoVA_ctl) showed its skill up to the threshold of 25 mm/3 h. The ensemble mean improve the control forecast (JNoVA_ctl) for weak and moderate rains with the strength of the threshold below 10 mm/3 h. Without GPS PWV, threat scores (Meso4dvar_ctl and Meso4dvar_mean) are not good even in the 2 km ensemble. This result suggest that the accuracy of the initial condition of the control run is primarily important even in the high resolution

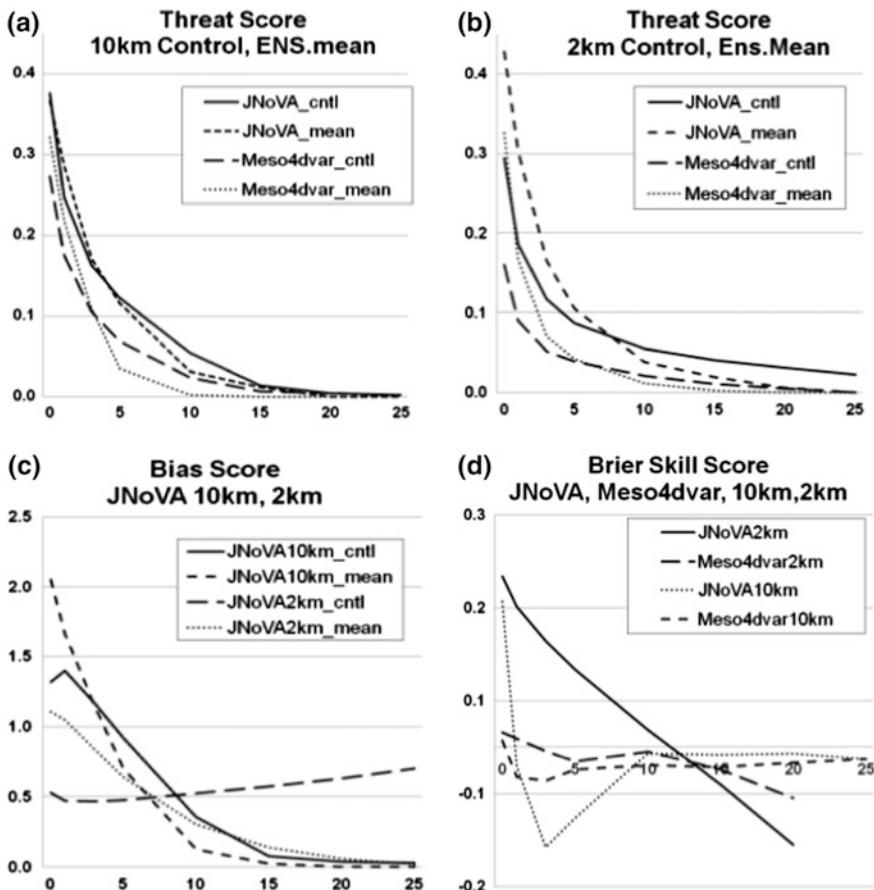


Fig. 12 **a** Threat scores by the forecast of 10 km forecast. Average of 0600 JST to 1800 JST of August 5th. **b** Same as in (a) but 2 km forecast. **c** Bias scores of 10 and 2 km forecasts. **d** Brier skill score of the 10 and 2 km ensemble forecasts

ensemble forecast, for such the cases of local convective rains where the result is sensitive to the slight difference in the initial value.

Figure 12c shows the bias scores by the NHM10 and NHM2 experiments with GPS PWV data. In the 10 km ensemble mean, frequency of rains exceeding 5 mm/3 h become smaller while weak rains less than 1 mm/3 h are overestimated, whereby the threat score (Fig. 12a) is also worsening. Bias scores of the 2 km control run gradually increase with the precipitation intensity, but underestimate weak rains below 5 mm/3 h, suggesting that the resolution of 2 km is not yet sufficient to predict the occurrence of small scale convection cells. In the 2 km ensemble, tendency of the ensemble mean to underestimate heavy rains is the same as in the 10 km ensemble, but the bias trend of the control run in 2 km acts to alleviate the tendency thus the bias of the 2 km ensemble (JNoVA2 km_mean)

becomes somewhat flat compared to the 10 km ensemble (JNoVA10km_mean). This compensation is considered to be a reason that threat scores of the 2 km ensemble mean overcomes those of the control run for the weak rains (Fig. 12b).

5.2 Verification of Probabilistic Forecast

In this section we verify scores which represent the accuracy of probability forecast. Brier score is the well-known mean square error of the probability estimation:

$$BS = \frac{1}{N} \sum_1^N (p_f - p_o)^2, \quad (1)$$

where p_o is the probability of an event being observed and p_f the probability of prediction that the event occurs. If the positional lags are not considered, p_o is 1 when the event occurs and 0 if there is no event.

Brier score is an index to show the accuracy of the probability forecast, however, in case that the event occurs only occasionally, there is a drawback that Brier score tends to be better when the predicted probability p_f is small. There is Brier skill score (BSS) as an indicator that fixes this shortcoming. BSS is an index showing the improvement rate of Brier score against the probabilistic prediction based on climatology (BS_{ref}), and is expressed by the following equation:

$$BSS = 1 - BS/BS_{ref}. \quad (2)$$

Figure 12d shows the Brier skill scores for the 10 km mesoscale ensemble forecast and the 2 km cloud resolving ensemble forecast. Here we took observed sample frequency during the verification period as BS_{ref} . Brier skill scores are negative except 2 km cloud resolving ensemble forecast that assimilates GPS PWV (JNoVA2 km). Even for the forecast for JNoVA2km, positive skills were not obtained for thresholds over 15 mm/3 h. This result means that if the positional lags are not allowed, even the probability of precipitation by the ensemble forecast, high-resolution verification may face the problem of ‘double penalty’ (Ebert 2008; Gilleland et al. 2009), and suggests the difficulty of applying the results of the numerical weather prediction to the forecast as it is. Incidentally, it is well-known that a phenomenon that Brier skills score becomes negative often happens when the observed sample frequency during the verification period is used as BS_{ref} . In fact, the observed sample frequency is obtained only after the verification period thus to use them as BS_{ref} is unfair for the forecast. We can expect BSS becomes positive if a longer-term statistic is used for BS_{ref} .

In calculating the Brier score, fraction skill score (FSS) has been proposed by Mittermaier and Roberts (2010), where a template of a certain size is considered and an indicator is computed on how the average probability of occurrence in the template is well expressed:

$$FSS = 1 - \frac{\frac{1}{n} \sum_1^n (p_{fcst} - p_{obs})^2}{\frac{1}{n} \sum_1^n (p_{fcst})^2 + \frac{1}{N} \sum_1^n (p_{obs})^2}, \quad (3)$$

where p_{obs} and p_{fcst} are the percentage of verification grids where the event was observed and predicted in the template, respectively.

Figure 13 illustrate FSSs of 10 and 2 km ensemble predictions for different rainfall intensity when the size of the template to compute the fraction is varied. Here the horizontal axis is the rainfall intensity, and the vertical axis represents the size of the template to calculate the fraction (in the case of 30 km, 3×3 grids for 10 km and 15×15 grids for 2 km). For weak rains below 5 mm/3 h, the scores by the 10 km ensemble (left) is better than the 2 km ensemble (right), which reflects the low bias scores in NHM2 seen in Fig. 12c. The performances of the two ensembles become reverse for intense rains more than 10 mm/3 h, and the 2 km cloud resolving ensemble greatly improves the forecast of heavy rains. The scores become better when the template size is larger, but the improvement is distinct up to positional lag of 30 km (template size of 60 km). This result is consistent with Duc et al. (2013) which verified downscale cloud resolving ensemble prediction over Japan for the period of July 2010. Although the resolution of the JNoVA analysis is 5 km, the data assimilation is conducted by the inner loop model with a horizontal grid space of 15 km (Table 1). The template size of 60 km is likely corresponding to the effective minimum resolution of a grid model (4–5 times of the grid spacing). High-resolution downscaling can express the phenomenon corresponding to its model resolution, but the uncertainty of the spatial and time scales corresponding to the resolution and frequency of the original observation data and assimilation remains unavoidable, thus the deterministic forecast for smaller spatiotemporal scales is not possible.

FSSs shown in Fig. 13 are not high even considering the positional lags. This result is in contrast with the case of the 2011 July Niigata and Fukushima heavy

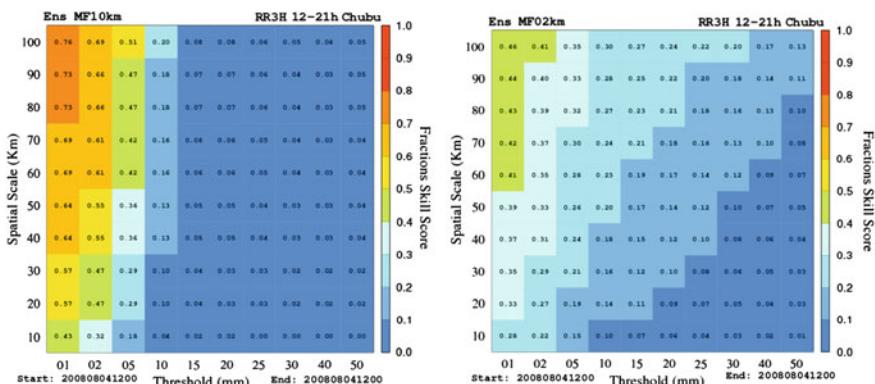


Fig. 13 Fractions Skill Score of 10 km ensemble forecast (left) and 2 km ensemble forecast (right). The horizontal axis is 3 h rainfall intensity, and the vertical axis is horizontal scale

rains (Saito et al. 2013) where the horizontal convergence associated with the large scale stationary front was clear.

6 Summary and Concluding Remarks

Targeting on the Tokyo metropolitan area local heavy rains of August 5th, 2008, we conducted downscale and ensemble forecast experiments with horizontal resolutions of 10 and 2 km using the JMA nonhydrostatic model. Initial time of the 10 km experiments were set at 1200 UTC (2100 JST), 4th August. Heavy rains in the Tokyo metropolitan area were not predicted by the control run of the 10 km forecast, and rainfalls in the metropolitan area was small even in the 2 km model. In the 10 km ensemble forecast, some members represented the moderate precipitation, but the maximum rainfall was less than 20 mm in three hours. Although the ensemble mean somewhat improved the control run in place and spread of the rainfalls, strong rains were not represented at all. In the 2 km cloud resolving ensemble, some members expressed strong rainfalls in the metropolitan area, but positions of the intense precipitation were shifted to the west compared to observation.

PWV of operational mesoscale analysis of JMA at the initial time was underestimated compared to the GEONET observations. To improve water vapor field in the initial condition, we conducted data assimilation experiments of GPS PWV using the JMA nonhydrostatic regional 4DVAR system. Water vapor field of the initial value was modified when GPS PWV was assimilated and precipitation forecast were significantly improved in the 10 km forecast and its ensemble, but strong precipitation over 20 mm/3 h were not predicted. The forecast of the strong precipitation was improved by the 2 km downscaling, and in the 2 km ensemble, occurrence of the strong rainfalls in the Tokyo metropolitan area and in Shizuoka prefecture was represented in most members. No precipitation area of the northern Kanto is also well represented, and the probability distribution diagram of 20 mm/3 h well reproduced the characteristic distributions of the observed spread of the strong precipitation cells.

From above results, it can be seen that heavy rain caused by convection cells in the field of weak external forcing cannot be expressed sufficiently by the low resolution, cumulus convection parameterized model. The 2 km cloud resolving model represents heavy rains, but the area of weak rains is reduced. The ensemble forecast reflects the tendency of the control run, thus the accuracy of the initial value is still important. Since it adds information about the quantitative forecast uncertainty, the value of the forecast significantly increases compared to the deterministic forecast only.

The event in this study were sporadic localized heavy rains by small scale convection cells that occurs in the field of weak large-scale forcing, whose deterministic prediction was very difficult. The situation is in contrast with the case of the 2011 July Niigata and Fukushima heavy rains where the horizontal convergence

by a stationary front was evident. Even in the case of this study, areas where strong rains occur were well reproduced by the 2 km ensemble forecast with the modified initial condition. This result suggests a certain predictability of small scale convective rains in terms of the probabilistic forecast. Although Brier skill scores were negative even in the 2 km ensemble with GPS PWV assimilation but fraction skill score clearly demonstrated advantage of the cloud resolving ensemble for intense rains. Spatial scale of the template size to compute the fraction may be affected by the original uncertainty in the analysis. Even in local heavy rainfall that occurs in a field of weak large-scale forcing, local surface convergence by the sea breeze sometimes triggers the occurrence of deep convection. Predictability for such the case observed in the Tokyo Metropolitan Area Convection Study (TOMACS; Nakatani et al. 2013) has been reported by Saito et al. (2014) and will be discussed in another paper.

Acknowledgements In the ensemble numerical experiments and verification, we are indebted to Masaru Kunii of the Meteorological Research Institute. We also thank Hiromu Seko of the Meteorological Research Institute and Tohru Kuroda of the Japan Agency for Marine Science and Technology for their help. A part of this study was supported by the HPCI Strategic Program for Innovative Research (SPIRE) field 3.

References

- Bevis M, Businger S, Chiswell S, Herring TA, Anthes RA, Rocken C, Ware RH (1994) GPS meteorology: mapping Zenith wet delays onto precipitable water. *J Appl Meteor* 33:379–386
- Duan Y, Gong J, Du J, Charron M, Chen J, Deng G, DiMego G, Hara M, Kunii M, Li X, Li Y, Saito K, Seko H, Wang Y, Wittmann C (2012) An overview of the Beijing 2008 olympics research and development project (B08RDP). *Bull Am Meteor Soc* 93:381–403
- Duc L, Saito K, Seko H (2013) Spatial-temporal fractions verification for high resolution ensemble forecasts. *Tellus* 65. doi:[10.3402/tellusa.v65i0.18171](https://doi.org/10.3402/tellusa.v65i0.18171)
- Ebert E (2008) Fuzzy verification of high resolution gridded forecasts: a review and proposed framework. *Meteor Appl* 15:51–64
- Honda Y, Nishijima M, Koizumi K, Ohta Y, Tamiya K, Kawabata T, Tsuyuki T (2005) A pre-operational variational data assimilation system for a non-hydrostatic model at the Japan meteorological agency: formulation and preliminary results. *Q J R Meteorol Soc* 131:3465–3475
- Honda Y, Sawada K (2009) Upgrade of the operational mesoscale 4D-Var system at the Japan meteorological agency, CAS/JSC WGNE Res Act Atmos Ocea Model 39:01.11–01.12
- Ishihara M (2013) Radar echo population of air-mass thunderstorms and nowcasting of thunderstorm-induced local heavy rainfalls. Part 1: statistical characteristics. *J Disaster Res* 8:56–68
- Ishikawa Y, Koizumi K (2002) One month cycle experiments of JMA mesoscale 4-dimensional variational data assimilation (4D-Var) system. *CAS/JSC Res Act Atmos Ocea Model* 32:01.26–01.27
- Gilleland E, Ahijevych D, Brown B, Casati B, Ebert E (2009) Intercomparison of spatial forecast verification methods. *Weather Forecast* 24:1416–1430
- Kawabata T, Seko H, Saito K, Kuroda T, Tamiya K, Tsuyuki T, Honda Y, Wakazuki Y (2007) An assimilation experiment of the nerima heavy rainfall with a cloud-resolving nonhydrostatic 4-dimensional variational data assimilation system. *J Meteor Soc Jpn* 85:255–276

- Kawabata T, Kuroda T, Seko H, Saito K (2011) A cloud-resolving 4D-Var assimilation experiment for a local heavy rainfall event in the Tokyo metropolitan area. *Mon Weather Rev* 139:1911–1931
- Mittermaier M, Roberts N (2010) Intercomparison of spatial forecast verification methods: identifying skillful spatial scales using the fractions skill score. *Weather Forecast* 25:343–354
- Nakatani T, Shoji Y, Misumi R, Saito K, Seino N, Seko H, Fujiyoshi Y, Nakamura I (2013) WWRP RDP science plan: Tokyo metropolitan area convection study for extreme weather resilient cities (TOMACS). WWRP report for Joint Scientific Committee, 26pp. http://www.wmo.int/pages/prog/arep/wwrp/new/documents/Doc4_6_TOMACS_RDP_proposal_20130704.pdf
- Ogura Y (2009) View, enjoy the weather (16) Let's eliminate the word 'guerrilla heavy rain'. *Tenki*, 56:555–563 (in Japanese)
- Saito K (2012) The Japan Meteorological Agency nonhydrostatic model and its application to operation and research. *Atmos Model Appl Intech* 85–110. doi:[10.5772/35368](https://doi.org/10.5772/35368)
- Saito K, Fujita T, Yamada Y, Ishida J, Kumagai Y, Aranami K, Ohmori S, Nagasawa R, Kumagai S, Muroi C, Kato T, Eito H, Yamazaki Y (2006) The operational JMA nonhydrostatic mesoscale model. *Mon Weather Rev* 134:1266–1298
- Saito K, Ishida J, Aranami K, Hara T, Segawa T, Narita M, Honda Y (2007) Nonhydrostatic atmospheric models and operational development at JMA. *J Meteor Soc Jpn* 85B:271–304. doi:[10.2151/jmsj.85B.271](https://doi.org/10.2151/jmsj.85B.271)
- Saito K, Kunii M, Hara M, Seko H, Hara T, Yamaguchi M, Miyoshi T, Wong W (2010) WWRP Beijing 2008 Olympics Forecast Demonstration/Research and Development Project (B08FDP/RDP). Tech. Rep. MRI, 62, 210pp. doi:[10.11483/mritechrepo.62](https://doi.org/10.11483/mritechrepo.62)
- Saito K, Hara M, Kunii M, Seko H, Yamaguchi M (2011) Comparison of initial perturbation methods for the mesoscale ensemble prediction system of the Meteorological Research Institute for the WWRP Beijing 2008 Olympics Research and Development Project (B08RDP). *Tellus* 63A:445–467
- Saito K, Origuchi S, Duc L, Kobayashi K (2013) Mesoscale ensemble forecast experiment of the Niigata-Fukushima heavy rainfall. *Tech. Rep. JMA*, 134, 10–184 (in Japanese, <http://www.jma.go.jp/jma/kishou/books/gizyutu/134/ALL.pdf>)
- Saito K, Kunii M, Araki K (2014) Cloud resolving simulation of a local heavy rainfall event on 26 August 2011 observed by the Tokyo Metropolitan Area Convection Study (TOMACS). *CAS/JSC WGNE Res Act Atmos Ocea Model* 44:5.05–5.06
- Shoji Y (2009) A study of near real-time water vapor analysis using a nationwide dense GPS network of Japan. *J Meteor Soc Jpn* 87:1–18
- Shoji Y, Kunii M, Saito K (2009) Assimilation of Nationwide and Global GPS PWV Data for a Heavy Rain Event on 28 July 2008 in Hokuriku and Kinki, Japan. *SOLA* 5:45–48
- Shoji Y, Kunii M, Saito K (2011) Mesoscale data assimilation of Myanmar cyclone Nargis. Part 2: assimilation of GPS derived precipitable water vapor. *J Meteor Soc Japan*, 89:67–88
- Tokyo Regional Headquarters (2008) Report on heavy rainfall of 5 August 2008. 10pp (http://www.jma-net.go.jp/tokyo/sub_index/bosai/disaster/20080805/20080805.pdf)

Validation and Operational Implementation of the Navy Coastal Ocean Model Four Dimensional Variational Data Assimilation System (NCOM 4DVAR) in the Okinawa Trough

Scott Smith, Hans Ngodock, Matthew Carrier, Jay Shriner,
Philip Muscarella and Innocent Souopgui

Abstract The Navy Coastal Ocean Model Four-Dimensional Variational Assimilation (NCOM 4DVAR) system is an analysis software package that is designed to supplement the current capability of the operational analysis/prediction system known as the Relocatable Navy Coupled Ocean Model (Relo NCOM) system. The present assimilation component of Relo NCOM employs the Navy Coupled Ocean Data Assimilation Three-Dimensional Variational Assimilation (NCODA 3DVAR) system to process and assimilate observations. The NCOM 4DVAR, on the other hand, uses a representer based 4DVAR method and has been found to improve the forecast-skill for several regional applications. This chapter presents the results of validation experiments performed in the Okinawa Trough. The analysis and resulting forecast skill of the two assimilation methods within Relo NCOM (NCOM 4DVAR and NCODA 3DVAR) are compared, and the operational implementation of NCOM 4DVAR is examined to verify that it satisfies operational constraints. The metrics used to validate the NCOM 4DVAR system include: computational efficiency, scalability, robustness, and the prediction accuracy of temperature, sea surface height, and sonic layer depth through NCOM 4DVAR and NCODA 3DVAR analyses. Forecast skill metrics are computed using surface observations of temperature, salinity and sea surface height, and profile observations from gliders and AXBTs (aerial expendable bathythermograph). Overall, the validation reveals that NCOM 4DVAR has lower root mean square errors for both analyses and forecasts than the operational NCODA 3DVAR system.

S. Smith (✉) · H. Ngodock · M. Carrier · J. Shriner · P. Muscarella
Stennis Space Center, Naval Research Laboratory, Bay St. Louis, MS, USA
e-mail: Scott.Smith@nrlssc.navy.mil

I. Souopgui
Stennis Space Center, University of Southern Mississippi, Hattiesburg, MS, USA

1 Introduction

The Navy Coastal Ocean Model four-dimensional variational (NCOM 4DVAR) system is a data assimilative nowcast/forecast ocean modeling and prediction system developed at the Naval Research Laboratory (NRL) for use at the Naval Oceanographic Office (NAVOCEANO) (Smith et al. 2015). This system is built to be used within the same framework as the Relocatable NCOM (Relo NCOM), which is the present operational ocean analysis and prediction tool that the Navy uses for non-global applications. The current data assimilation component of Relo NCOM uses the three-dimensional variational data assimilation method, performed by the Navy Coupled Ocean Data Assimilation system (NCODA 3DVAR, Smith et al. 2012). The newly developed NCOM 4DVAR system is designed to supplement NCODA 3DVAR allowing the user to select either system depending on the application at hand. Regardless of which assimilation option is selected, Relo NCOM uses the same forcing, initial and boundary conditions, and the same ocean model (NCOM) for its forecasting component.

Most ocean models have reduced accuracy and prediction skill at regional and coastal scales where the prediction of tracers, currents, and acoustic properties are important for search and rescue operations, hydrocarbon/chemical spill simulations, environmental prediction, and other Navy operations. While the currently operational NCODA 3DVAR may be ideal for global and large basin scales due to its computational efficiency, NCOM 4DVAR has improved analysis/forecasting capabilities and has shown that it can be operated at sufficiently high resolution in coastal and/or regional areas in a reasonable amount of time (Ngodock and Carrier 2014b). The NCOM 4DVAR is able to provide an improved analysis by accounting for observations at their actual collection times, rather than assuming the observations occur at the same time as in 3DVAR. Also in 4DVAR, observation corrections are temporally correlated and their influence is propagated throughout the entire assimilation window via the model dynamics. This allows more information to be extracted and utilized from sparse observations, thereby producing a more accurate and dynamically consistent analysis, which in turn increases the forecasting predictability skill. Another advantage that NCOM 4DVAR has is the capability to directly assimilate velocity (Carrier et al. 2014) and sea surface height (SSH) (Ngodock et al. 2015) observations without having to use synthetic observations. Synthetic observations consist of temperature and salinity profiles that are derived from SSH observations (Fox et al. 2002). The generation of synthetic observations is required in the NCODA 3DVAR assimilation system because there are no model dynamics or cross-covariances to correlate SSH observations to the other variables.

In this study, the resulting analyses and forecasts from three experiments are analyzed and compared. All experiments are performed using the operational implementation of Relo NCOM for the Okinawa Trough. The only difference between the experiments is that the first uses NCODA 3DVAR and the second and third experiments use NCOM 4DVAR. Two separate NCOM 4DVAR experiments are performed and presented that use different methods of assimilating sea surface

height (SSH) observations (described in Sect. 2.3). Every effort is made to keep the forcing, parameters, data, and data processing as similar as possible between the experiments, so that the primary aspect being compared is the assimilation method.

Section 2 describes the Relo NCOM system along with its major components that are relevant to the validation experiments in this study. Then in Sect. 3, the setup of the validation testing experiments for the Okinawa Trough are discussed, followed by the results in Sect. 4. Section 5 goes over some of the implications of applying the NCOM 4DVAR system operationally. Finally, some conclusions are provided in Sect. 6.

2 Components of Relo NCOM

The Relo NCOM system is a flexible data assimilation/forecasting system (Rowley 2010), with most model configuration parameters available for the user to define. The Relo NCOM system consists of a suite of scripts that efficiently handle the input and output data streams, NCODA data processing, the data assimilation, and NCOM forecasts. It also performs the preparation of a new domain, which includes interpolating and setting up the initial and boundary conditions and surface forcing. The initial and boundary conditions are extracted from a larger model, such as the global Hybrid Coordinate Ocean Model (HYCOM) (Metzger et al. 2014). The surface forcing fields can come from the Navy Operational Global Atmospheric Prediction System (NOGAPS, Rosmond 1992; Rosmond et al. 2002); Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS; Hodur 1997) products generated at the Fleet Numerical Meteorology and Oceanography Command (FNMOC); from COAMPS raw output; or now from the Navy Global Environmental Model (NAVGEN, Hogan et al. 2014). In most cases, atmospheric model wind stresses, radiation fluxes, atmospheric pressure, temperature, and humidity are prepared for the NCOM model, and bulk flux formulae are used in NCOM to calculate surface heat fluxes (Rowley 2010).

2.1 *Navy Coastal Ocean Model (NCOM)*

The Navy Coastal Ocean Model (NCOM, Martin 2000) is the ocean forecasting component of Relo NCOM. NCOM was developed primarily from two existing ocean circulation models, the Princeton Ocean Model (POM) (Blumberg and Mellor 1983; 1987) and the Sigma/Z-level Model (SZM) (Martin et al. 1998). NCOM has a free-surface and is based on the primitive equations and hydrostatic, Boussinesq, and incompressible approximations. Turbulent mixing is parameterized by the Mellor-Yamada Level-2.5 (MYL2.5) turbulence closure parameterization (Mellor and Yamada 1982) for vertical diffusion and the Smagorinsky scheme (Smagorinsky 1963) for horizontal diffusion (Carrier et al. 2014). The vertical

mixing enhancement scheme of Large et al. (1994) is used for parameterization of unresolved mixing processes occurring at near-critical Richardson numbers. A source term is included in the model equations to allow for river input and runoff inflows (Barron et al. 2007).

As in the POM, NCOM employs a staggered Arakawa C grid with an orthogonal-curvilinear horizontal grid orientation. Spatial finite differences are mostly second-order centered, but higher-order spatial differences are optional. NCOM features a leapfrog temporal scheme with an Asselin filter to suppress time splitting. Most terms are handled explicitly in time, but surface wave propagation and vertical diffusion are solved implicitly (Martin 2000). In the vertical, NCOM can be configured with terrain-following free-sigma or fixed sigma, or constant z -level surfaces or their combination (Barron et al. 2006). Typically, one of two types of combinations is used: the first is a hybrid sigma and z -level combination with sigma coordinates applied from the surface down to a designated depth (100–200 m depending on where the shelf break is located), and z -levels below this specified depth. The second vertical grid choice typically used is the general vertical coordinate (GVC) grid consisting of a three-tiered structure: (1) a near-surface “free” sigma grid that expands and contracts with the movement of the free surface, (2) a “fixed” sigma, and (3) a z -level grid allowing for “partial” bottom (Martin et al. 2008).

2.2 *Navy Coupled Ocean Data Assimilation 3D Variational Analysis (NCODA 3DVAR) System*

NRL developed and implemented an ocean data analysis component of COAMPS called the Navy Coupled Ocean Data Assimilation System (NCODA; Cummings 2005). The version of NCODA used operationally and in this study employs the 3DVAR method and is capable of processing observations from a large number of different platforms. These include, but are not limited to: satellite sea surface temperature (SST), SSH/altimetry, satellite microwave-derived sea ice concentration, and in situ surface and profile data from ships, drifters, fixed buoys, profiling floats, XBTs (expendable bathythermographs), AXBTs (aerial expendable bathythermographs), CTDs (conductivity, temperature, and depth), and gliders. The observational data are prepared and processed through the NCODA automated data quality control system (NCODA-QC) which identifies spurious observations compared against climatological or model fields and associated variability information (Cummings 2011). Observations that satisfy the quality control are then passed into another NCODA module called NCODA-PREP where they are combined with the previous forecast fields to produce the initial innovations. Observation and forecast errors, and correlation scales are also computed in NCODA-PREP.

The NCODA 3DVAR module reads in the innovations and error covariance information and uses a conjugate gradient routine to minimize a 3D variational cost function to determine the analysis increments in observation space. These increments are then mapped back to the model space using the background error

covariances resulting in a set of corrections corresponding to the NCOM forecast fields (Smith et al. 2012). The NCODA 3DVAR system is currently being used operationally at NAVOCEANO in the Relo NCOM, global HYCOM, and COAMPS.

2.3 *Navy Coastal Ocean Model Four Dimensional Variational System (NCOM 4DVAR)*

The NCOM 4DVAR system operates within the framework of Relo NCOM. The same scripts that are used to operate Relo NCOM with NCODA 3DVAR are used to operate the NCOM 4DVAR, with a few additional parameters for the NCOM model adjoint and the specification of the assimilation window. NCOM 4DVAR uses the same data that is processed by NCODA-QC and it also uses NCODA-PREP to process these observations for the specified domain. NCODA-PREP had to be slightly modified to account for the temporal distribution of the observations and to create time dependent innovations that are required for the NCOM 4DVAR. It should be noted that an observation density-reduction option has been added to the NCOM 4DVAR to ensure that no two observations fall within a correlation scale distance of one another, as too many correlated observations can adversely affect the conditioning of the minimization.

The analysis component of NCOM 4DVAR is a variational assimilation system based on the indirect representer method as described by Bennett (1992, 2002) and Chua and Bennett (2001) and uses the tangent linearization (TL) of the NCOM code and its adjoint. The NCOM 4DVAR system is described in detail by Ngodock and Carrier (2014a), and a full derivation of the representer method can be found in Chua and Bennett (2001). Therefore, only an overview is provided here.

The representer method aims to find an optimal analysis solution as the linear combination of a first guess (i.e., prior model solution) and a finite number of representer functions, one per datum,

$$\hat{u}(x, t) = u_F(x, t) + \sum_{m=1}^M \hat{\beta}_m r_m(x, t), \quad (1)$$

where $\hat{u}(x, t)$ is the analysis solution, $u_F(x, t)$ is the prior forecast, $r_m(x, t)$ is the representer function for the m th observation, and $\hat{\beta}_m$ is the m th representer coefficient. The representer coefficients are found by solving the linear system,

$$(\mathbf{R} + \mathbf{O})\beta = \mathbf{y} - \mathbf{H}u_F, \quad (2)$$

where \mathbf{O} is the observation error covariance, \mathbf{y} is the observation vector, and \mathbf{H} is the linear observation operator that maps the model fields to the observation locations. \mathbf{R} is the representer matrix defined as,

$$\mathbf{R} = \mathbf{HMBM}^T \mathbf{H}^T, \quad (3)$$

where \mathbf{M} is the TL of NCOM, \mathbf{M}^T is the adjoint of NCOM, and \mathbf{B} is the model error covariance. Since the matrix $\mathbf{R} + \mathbf{O}$ is symmetric and positive definite, Eq. (2) can be solved for β iteratively using a linear solver, such as the conjugate gradient method. From Eqs. (2) and (3), β_m can be found for each representer by integrating the adjoint and TL models over some number of minimization steps until convergence.

In the NCOM 4DVAR, β_m is found with a pre-conditioned conjugate gradient solver. The preconditioner here follows from Courtier (1997), where β is redefined as $\mathbf{u} = \sqrt{\mathbf{O}}\beta$ in the minimization step such that Eq. (2) can be expressed as,

$$\left(\sqrt{\mathbf{O}^{-1}} \mathbf{R} \sqrt{\mathbf{O}^{-1}} + \mathbf{I} \right) \mathbf{u} = \sqrt{\mathbf{O}^{-1}} (\mathbf{y} - \mathbf{H}\mathbf{u}_F) \quad (4)$$

This transformation ensures that there is a lower bound of 1 for the eigenvalues, which insures that the condition number will remain reasonably small and allow the conjugate gradient solver to converge relatively quickly. Once β is determined, Eq. (1) is then used to compute the analysis.

The background and model error covariance in NCOM 4DVAR is univariate and follows the work of Weaver and Courtier (2001) and Carrier and Ngodock (2010). This is deemed acceptable as the application of the TL and adjoint models in the minimization and final sweep provide multivariate balance constraints through the linearized dynamics. It has been shown (Yu et al. 2012) that omitting linear balance constraints does not lead to a significant degradation of the final solution in terms of the fit to observations. The univariate error covariance can be decomposed into a correlation matrix and associated error variance such that,

$$\mathbf{B} = \Sigma \mathbf{C} \Sigma, \quad (5)$$

where Σ is a diagonal matrix consisting of the standard deviations of the background error and \mathbf{C} is a symmetric matrix of background error correlations. In NCOM 4DVAR, the error standard deviations of the background are used at the initialization of the TL model only, whereas the model error (also contained in the matrix Σ) is used when the adjoint forces the TL model during integration (i.e., as the TL model integrates forward in time). This allows the weak constraint method to correct for the initial condition error while also adjusting the forward model trajectory based on the specification of the model error. The error correlation, for both the model and the background errors, is not directly calculated and stored in NCOM 4DVAR; rather, the effect of the correlation matrix acting on an input vector is modeled by the solution of a diffusion equation (Weaver and Courtier 2001; Yaremchuk et al. 2013; Carrier and Ngodock 2010; Ngodock 2005).

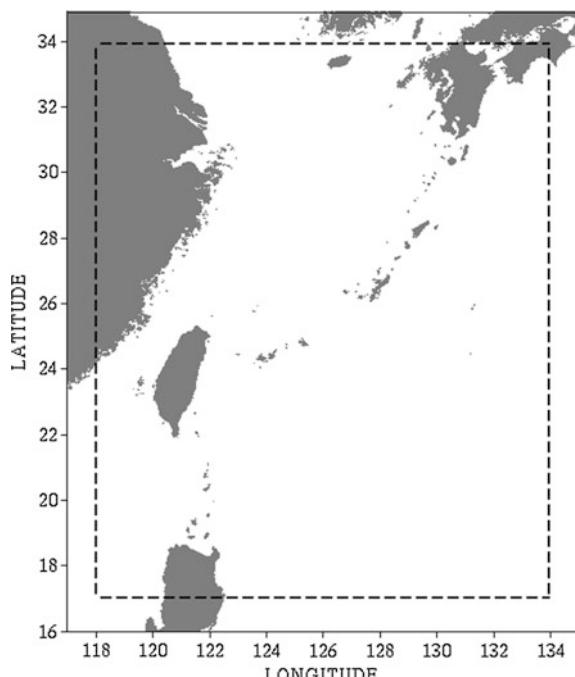
The NCOM 4DVAR includes two methods for assimilating SSH. The first is by creating synthetic profiles of temperature and salinity (T and S) in the same way as

NCODA 3DVAR (Smith et al. 2012). The second option is to assimilate SSH observations directly. Direct assimilation of SSH is not feasible with the NCODA 3DVAR system because it creates gravity waves. A method has been devised for NCOM 4DVAR to overcome this issue by assimilating SSH observations only into the baroclinic mode of the model. NCOM 4DVAR has an internal routine that checks and adjusts the barotropic mode to ensure that it is in balance with the baroclinic mode. Therefore, by the time the SSH observation information reaches the barotropic mode, it is in dynamic balance with the model and does not produce gravity waves. A more detailed description of this method is provided in Ngodock et al. (2015).

3 Validation Test Description: Okinawa Trough

The study region encompasses both the Okinawa Trough and Ryukyu Islands of Japan, from 17°N to 34°N and 118°E to 134°E (Fig. 1). The Okinawa Trough region is highly dynamic in nature; it has a complex geometry, sharp bathymetry gradient, a strong Kuroshio current, large barotropic and internal tides, significant river input, and frequent typhoon passage. All of these features provide an excellent testing ground to evaluate the predictive capability of the NCOM 4DVAR assimilation system. The Okinawa Trough is located between Taiwan and southern Japan

Fig. 1 The Okinawa Trough model domain, with 3 km horizontal resolution. The study region encompassed both the Okinawa Trough and Ryukyu Islands of Japan, from 17°N to 34°N and 118°E to 134°E (dashed lines)



and is a seabed feature of the East China Sea; it is an active, initial back-arc rifting basin which formed behind the Ryukyu arc-trench system in the western Pacific Ocean. A large portion of the domain is more than 1,000 m deep with a maximum depth of 2,716 m.

All of the Relo NCOM Okinawa Trough experiments that are compared in this study are 12 months long (Jan 1, 2007–Dec 31, 2007). Each of these experiments uses surface boundary conditions from the global 0.5° NOGAPS (Rosemond et al. 2002) and lateral boundary conditions from a 6 km Relo NCOM Western Pacific domain. This Western Pacific Relo NCOM is performed operationally at NAVO-CEANO and receives its lateral boundary conditions from the global NCOM (Barron et al. 2006 and 2007). Each experiment employs NCOM configured with 50 layers in the vertical including 25 free-sigma layers extending to a depth of 116 m with constant z-levels extending down to a maximum of 5500 m.

The following three experiments will be used in this comparison: (1) a standard Relo NCOM using the operational implementation of NCODA 3DVAR (EXP1); (2) the NCOM 4DVAR where the SSH observations are assimilated via synthetic profiles of temperature and salinity generated by the Modular Ocean Data Assimilation System (MODAS, Fox et al. 2002) (EXP2); and (3) NCOM 4DVAR assimilating SSH observations through direct assimilation of the along-track measurements (EXP3). The standard implementation of NCODA 3DVAR (EXP1) uses the MODAS synthetic profiles to assimilate SSH observations, as is the case with EXP2.

The 4DVAR assimilation of along-track SSH (EXP3) is included in this comparison because it is a relatively new technique and has outperformed the other two methods in previous tests. In order to run EXP3, an estimated mean SSH field is needed to transform the observations from height anomalies into the SSH form of the ocean model. Since a sufficiently long enough time period of Relo NCOM does not exist for this domain, a 5-year mean SSH field from the global HYbrid Coordinate Ocean Model (HYCOM) is interpolated to the observation locations and added prior to the inclusion of the data within the assimilation. Another reason to include the direct assimilation of SSH in this study is to identify any model-drift that may be present in the cycling forecast from the 4DVAR analysis. There is a possibility that the assimilation of along-track SSH may produce unrealistic corrections to the thermodynamic state of the model. This is not a concern when synthetic profiles are assimilated, as the generation of these profiles uses climatology to constrain the temperature and salinity profiles. This climatological constraint does not exist when SSH observations are assimilated directly. On the other hand, the 4DVAR does constrain the adjustments to the temperature and salinity by the background around which the adjoint and TL models are linearized. The objective is to determine if this constraint is sufficient to prevent unrealistic adjustments to the thermodynamic structure and, therefore, prevent the model solution from drifting far from reality.

The observational data used in these experiments come from several sources: subsurface in situ and profile temperature (T) and salinity (S) observations were collected from XBTs and Argo Floats, SST observations are collected from

NOAA's GAC and LAC satellites, and SSH observations are from altimeter data obtained from the ENVISAT, GFO, and Jason-1 satellites. The SSH data are processed through the ALtimeter Processing System (ALPS; Jacobs et al. 2002), which is available from the Altimetry Data Fusion Center (ADFC) at the Navy Oceanographic Office (NAVOCEANO). A collection of additional glider and AXBT observations are also provided and used in this study.

At 3 kilometers (km) resolution, the Okinawa Trough domain has a spatial size of 535 by 628 grid points and 50 layers; this corresponds to a total of 16,799,000 grid points. Due to the computational cost of NCOM 4DVAR, which involves solving the adjoint and TL models several times within the minimization driver, the total time to run the assimilation for a model grid of this size is operationally prohibitive.

To reduce the computational time it is necessary to run the NCOM 4DVAR assimilation on a reduced resolution grid. For the 4DVAR experiments (EXP2 and EXP3), the model grid is coarsened by interpolating the 3 km model background to a 6 km analysis grid that covers the same region and vertical structure as the original configuration. This is deemed acceptable as the static spatial covariance scales employed by the NCOM 4DVAR are based on the Rossby radius of deformation, which is approximately 40 km for this region. Once the assimilation is complete on the reduced-grid, the analysis increments are projected back to the original 3 km resolution and added to the full-resolution background state to produce the analysis. A series of experiments conducted during the early testing phase for the NCOM 4DVAR in the Okinawa Trough confirmed that a forecast run at 3 km initialized by a 6 km analysis yields a nearly identical solution as one run from a 3 km analysis. This result, coupled with the fact that the computational cost of the analysis is greatly reduced by the use of the coarse-resolution analysis, justifies this method.

4 Validation of NCOM 4DVAR

The results of the validation testing are broken up into 4 subsections. The first is the analysis and 24-h forecast errors of the different experiments as a function of time throughout the 12-month time period. The remaining 3 subsections focus on subsurface predictability. These statistics are only computed over a 3-month time period (Aug–Oct, 2007), because the vast majority of profile observations are collected during this time. It is also important to examine the predictability and persistence of extended forecasts out to 96 h, and it is prohibitive to perform these extended forecasts for the entire year. In the second and third subsections, the time average predictability of the subsurface temperature and salinity is compared with all assimilated profile data (Sect. 4.2); and then with non-assimilated AXBT and glider profile data for independent verification (Sect. 4.3). Finally, in Sect. 4.4, the predictability of sonic layer depth (SLD) is analyzed.

4.1 Time Distribution of Errors

The first comparison performed on these experiments is to examine the seasonality and errors of the analysis and the corresponding 24-h forecast that is generated by NCOM. To do this, a normalized error metric is computed as a function of time over the 12 month time-period of the experiments. This evaluation can also help determine if any model drift is present in the solution; this would manifest itself as an increasing 24-h forecast error with time as the model solution slowly drifts from reality. The error metric employed for this is a normalized mean absolute error that will be referred to as the J_{fit} measure,

$$J_{fit} = \frac{1}{M} \sum_{m=1}^M \frac{|y_m - H_m x|}{\sigma_m}, \quad (6)$$

where M is the total number of observations; y_m , H_m , and σ_m are the observation, observation operator, and observation error, respectively, associated with the m th observation; and \mathbf{x} is the model state (either the forecast or analysis). Equation 6 indicates that if the forecast or analysis fits the collective observations within their corresponding prescribed observation errors, the J_{fit} value will be at or below one. If the J_{fit} value is well below the value of one, then this may indicate that the solution is over-fitting the observations, and the prescribed model errors may need to be reduced.

Figure 2 shows the J_{fit} normalized error of the analysis (red) and the 24-h forecast (blue) for both 4DAR NCOM experiments. In these figures, the dashed black line represents the overall observation error. The normalized errors (J_{fit}) are computed relative to all of the observations that are assimilated; or in the case of the 24-h forecast it is all of the observations that will be assimilated in the next cycle. These observations include temperature, salinity and in the case of EXP3, SSH. For both 4DVAR experiments (Fig. 2a, b) the analysis fits within the observation error for the majority of the time period and the 24-h forecasts are generally within 2 standard deviations of the observation error.

The results in this figure also show that the 4DVAR assimilation of synthetic observations (EXP2) is outperforming the 4DVAR assimilating SSH directly (EXP3). This, however, does not necessarily imply that EXP2 is better, because the normalized error metric (J_{fit}) computed for these two experiments uses different observations and observation errors. In EXP2 (Fig. 2a) SSH observations are converted to synthetic temperature and salinity observations, and it is these synthetic observations that are assimilated and used in the J_{fit} value, along with the remaining in situ temperature and salinity observations. The synthetic observations have a relatively high observation error, higher than the real observations of temperature and salinity that are assimilated directly in EXP3 (Fig. 2b). Therefore, the solution in EXP2 will fit within a lower percentage of the observation error.

Figure 3 displays comparisons of SSH normalized error of the 24-h forecasts generated from NCODA 3DVAR (EXP1) and the two 4DVAR NCOM

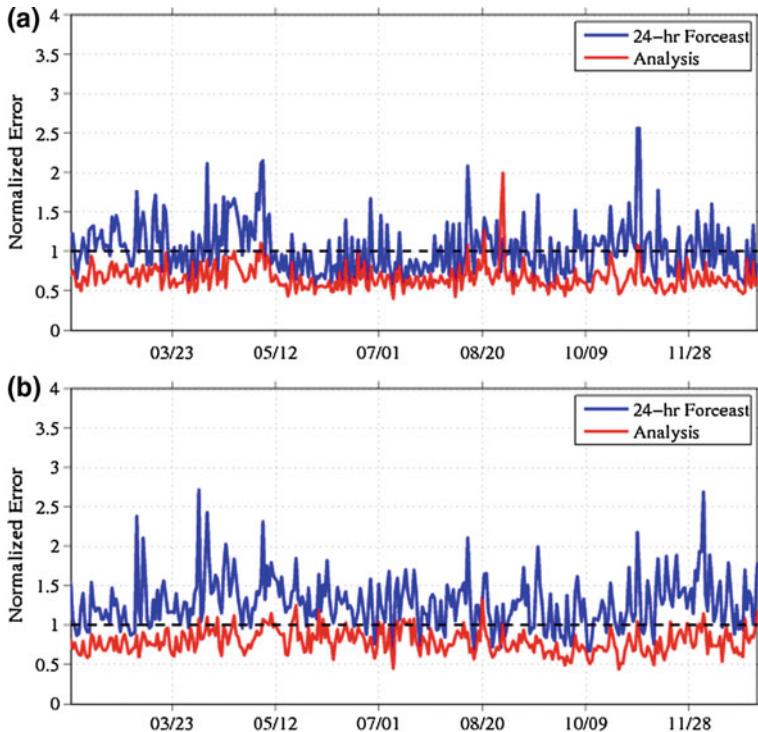


Fig. 2 Normalized error of the analysis (red) and 24-h forecast (blue) from the 4DVAR experiment assimilating **a** SSH via MODAS synthetics (EXP2) and **b** SSH directly (EXP3). The normalized errors are computed using Eq. 6 and are relative to all of the assimilated observations during the year-long Okinawa Trough experiments

experiments (EXP2 and EXP3). In these error metrics, the SSH forecasts are compared to a SSH map product created by the ALtimeter Processing System (ALPS) (Jacobs et al. 2002). Along-track SSH observations are high resolution in the along-track direction, but sparse in the cross-track direction. This makes comparisons with models difficult as the structure of mesoscale eddies cannot be entirely resolved using instantaneous SSH observations. The ALPS SSH product is a 2D optimal interpolation of sea surface height anomalies (SSHA) from multiple altimetry sources using characteristic covariance information regarding the scale of typical ocean eddies, propagation speeds, and time scales. A 5-year HYCOM mean SSH field is added to the ALPS SSH, in the same manner as the along-track SSH observations.

Figure 3a displays a comparison of the 24-h SSH forecast error between EXP1 (black) and EXP2 (red) for the 12 months of the experiments using the J_{fit} error metric in Eq. 6. In this comparison, EXP2 exhibits similar SSH forecast error as EXP1, albeit lower from January through May, and higher from May through September. This result is not surprising, as both forecasts are generated from

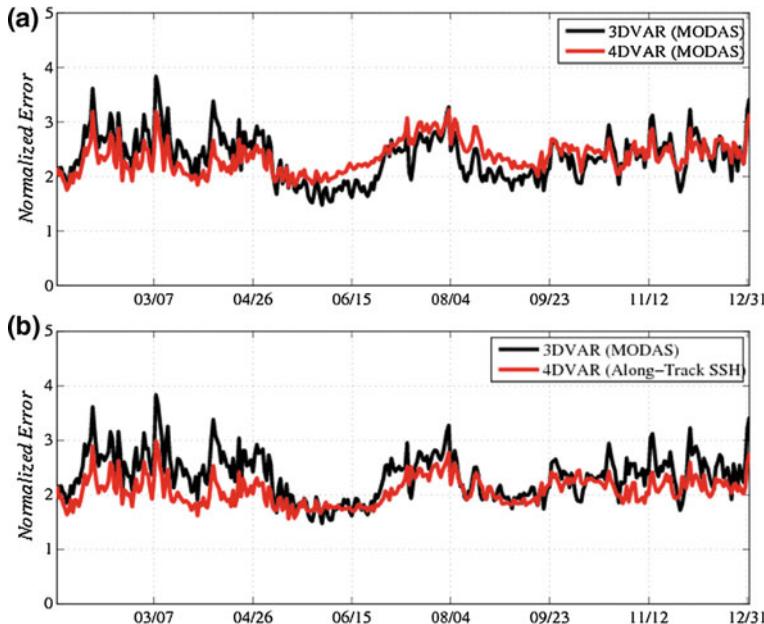


Fig. 3 Comparison of the 24-h SSH forecast error resulting from the year-long 3DVAR experiment (EXP1, black) and the two 4DVAR experiments: **a** assimilation of SSH via MODAS synthetics (EXP2, red) and **b** assimilation of SSH directly (EXP3, red). The normalized errors are computed using Eq. 6 and are relative to the SSH maps from ALPS

analyses that use synthetic profiles from MODAS to constrain the mass field. Since MODAS synthetics are based on climatology, they are generally more isotropic and slowly varying; therefore, limiting the advantage of 4DVAR over 3DVAR.

Figure 3b compares EXP3 and EXP1 24-h SSH forecast errors. In the case of EXP3, the 4DVAR analysis assimilates the along-track SSH observations directly. As such, the EXP3 24-h SSH forecast exhibits lower error than EXP1 generally throughout the entire 12-month experiment. This indicates that directly assimilating SSH, rather than through derived synthetic profiles of temperature and salinity, yields a superior SSH forecast. This is consistent with theory, as the observation errors for synthetic profiles are relatively high (Ngodock et al. 2015).

4.2 Profile Distribution Errors

It is important for the subsurface thermodynamic characteristics to be captured by the model, thus the first comparison presented is the RMS errors computed as a function of depth, rather than time. This error metric, calculated for the

forecasts generated from NCODA 3DVAR and NCOM 4DVAR, presents a comparison of the model layer-by-layer error relative to available profile observations.

Figures 4 and 5 show the 24-h forecast layer-by-layer RMS error value comparison between EXP1 and the 2 4DVAR experiments, EXP2 and EXP3, respectively. These error statistics are calculated for temperature (left panel) and salinity (right panel) relative to profile observations from three out of the 12 months of the experiments (August–October). In these figures, the value (N) in the left panel is the total number of profiles used to compute these statistics during the 3-month time period. Each profile consisted of both temperature and salinity observations down to a particular depth, so the total number of temperature and salinity observations used in these comparisons is the same. NCODA-QC calculates synthetic salinity profiles using MODAS for profile observations of just temperature (such as AXBTs). Then, layer-by-layer, RMS error values are computed for each experiment using forecast-observation comparisons during the entire 3-month period. It should be noted that not all of the profiles used for these comparisons went below 1400 m and many were confined to the upper 100 m. The results shown in these figures reveal that both EXP2 and EXP3 outperform EXP1 in predicting temperature, especially within the depth range of 100–600 m. Whereas, the systems are pretty similar at predicting salinity, except that EXP2 does not have the increased error near 350 m that is in EXP1.

Figure 6 is an overlay of all the error profiles in Figs. 4 and 5 for comparison, including their corresponding 96-h forecasts (dashed lines). As expected, the error

Fig. 4 Comparison of 24-h forecast RMS profile errors between EXP2 (red) and EXP1 (black) for temperature profiles (left panel) and salinity profiles (right panel). These are from 3-months (August–October). The value N is the total number of profile observations used in these statistics

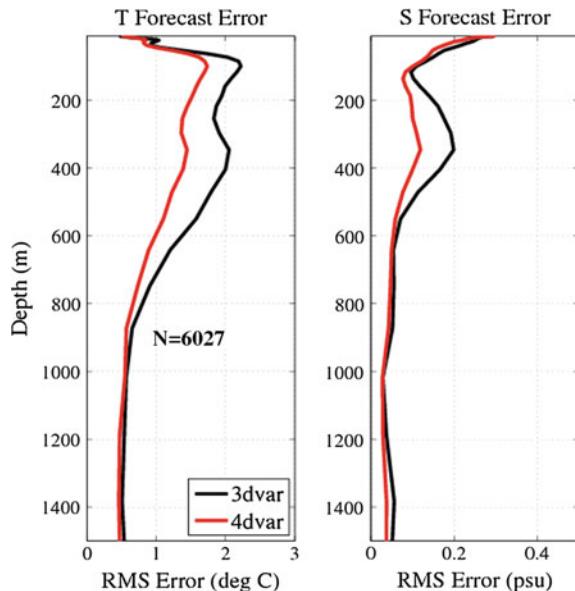


Fig. 5 Comparison of 24-h forecast RMS profile errors between EXP3 (red) and EXP1 (black) for temperature profiles (left panel) and salinity profiles (right panel). These are from 3-months (August–October). The value N is the total number of profile observations used in these statistics

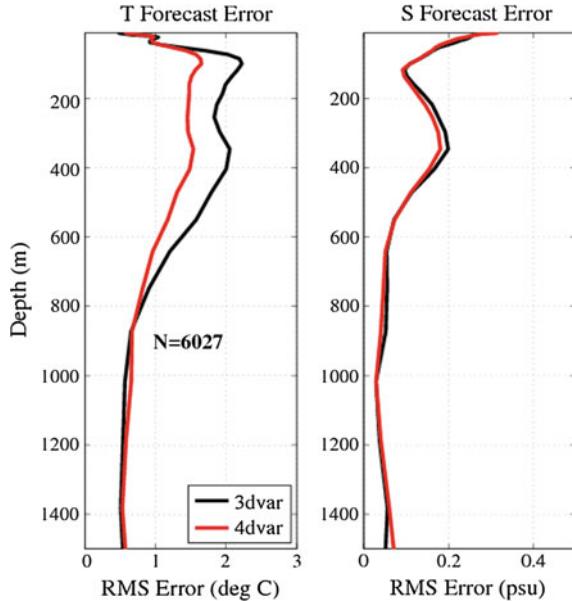
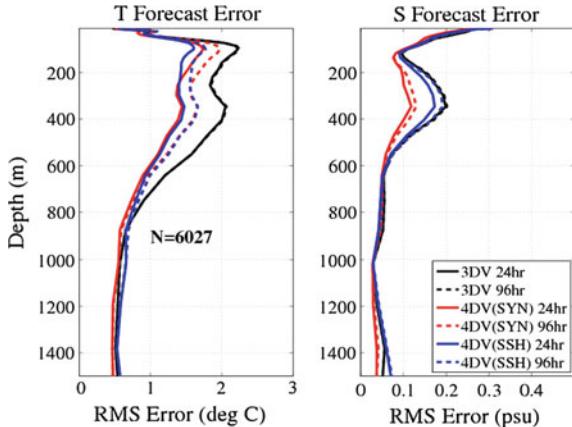


Fig. 6 Comparison of 24-h (solid) and 96-h (dashed) forecast RMS profile errors between EXP1 (black), EXP2 (red), and EXP3 SSH (blue) for temperature profiles (left panel) and salinity profiles (right panel). These statistics are computed over 3-months (August–October). The value N is the total number of profile observations used in these statistics



characteristics grow from the 24-h forecast to the 96-h for both 4DVAR systems. However, the gains provided by the 4DVAR analyses do not degrade much over the period of 96-h and the forecasts generated from the 4DVAR analyses continue to demonstrate skill over the forecasts generated by NCODA 3DVAR. It is interesting to point out that the 96-h forecasts of EXP2 and EXP3 have the same, or better, skill than the 24-h forecast of EXP1.

4.3 Profile Errors Relative to Independent Data

In addition to the 12-month experiments, a series of smaller 3-month runs using the NCOM 4DVAR (with direct SSH and synthetic assimilation) and the NCODA 3DVAR experiments were performed with some data types withheld for independent forecast evaluation. For these comparisons, EXP1 and EXP3 are compared with (1) all withheld glider data (Fig. 7) and (2) all withheld AXBT data (Fig. 8).

Figure 7 illustrates the layer-by-layer J_{fit} values (Eq. 6) for the EXP3 24-h forecast (red) versus the EXP1 24-h forecast (black) for temperature (left panel) and salinity (middle panel) computed against withheld glider observations. It should be noted that the withheld glider observations were also processed through NCODA-PREP. Therefore, the observation counts in the right panel of this figure are the processed glider observations binned into the NCODA analysis layers in time increments of 3 h. Clearly, the forecast using EXP3 outperformed EXP1 according to this independent data comparison, for both temperature and salinity through all model layers. Figure 8 shows the same comparison, but using withheld AXBT data. There is no salinity AXBT data, therefore there is no panel for salinity. Just as in the glider comparison, EXP3 outperforms EXP1 when compared to this independent data set.

Overall, the results from these experiments indicate that the NCOM 4DVAR analysis system, when assimilating SSH observations directly or through synthetic profiles of temperature and salinity, fits the assimilated observations within the

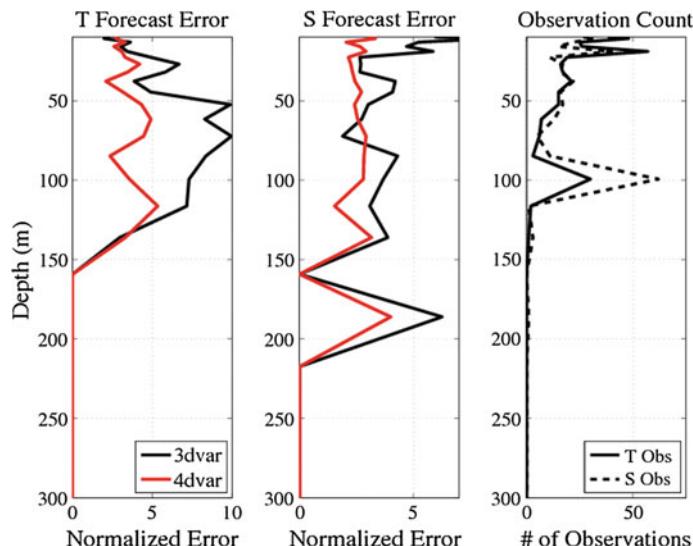
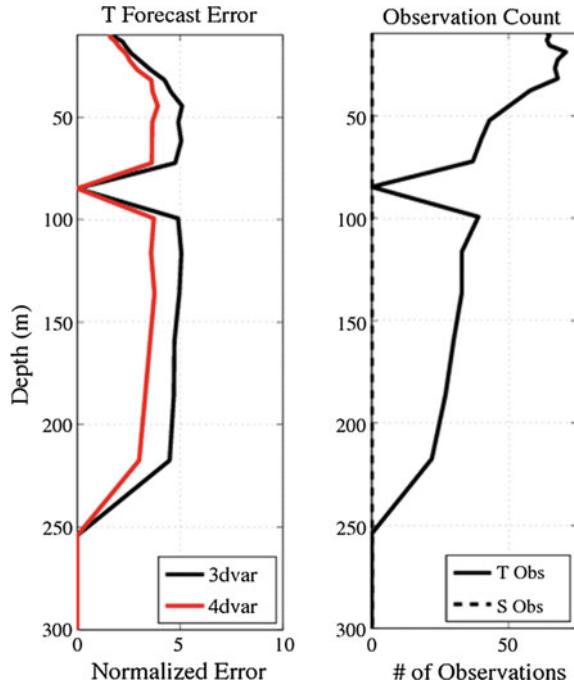


Fig. 7 The layer-by-layer J_{fit} error values (Eq. 6) for the EXP3 24-h forecast (red) versus the EXP1 24-h forecast (black) for temperature (left panel) and salinity (middle panel) computed against withheld glider observations (right panel)

Fig. 8 The layer-by-layer J_{fit} error values for the EXP3 24-h forecast (red) versus the EXP1 24-h forecast (black) for temperature (left panel) computed against withheld AXBT data (right panel). There is not a panel for salinity, because there are no salinity AXBT data



prescribed observation error. Further, the resulting forecasts generated from the NCOM 4DVAR analyses perform equally or better than the forecasts generated from the NCODA 3DVAR analyses, for both subsurface temperature and salinity, and also for model sea surface height.

4.4 Sonic Layer Depth Prediction

Sonic Layer Depth (SLD) is the depth at which sound speed is maximum in the upper water column. SLD is an important quantity to the Navy, because it is the upper boundary of the SOFAR (sound fixing and ranging) channel in which acoustic signals at certain frequencies can become trapped. Therefore, the prediction skill of SLD is one of the more important metrics that the Navy uses in determining the quality of a prediction system. For the comparison in this section, SLD was calculated using NRL's ProfParam software (Helber et al. 2008) for all of the glider and AXBT profile data (collected during 1 August 2007 through 31 October 2007). The analyses, and 24, 48, 72, and 96-h forecasts of EXP1, EXP2, and EXP3 were interpolated to these observation locations and times.

Table 1 provides the overall statistics of SLD prediction skill of each of the three experiments over the 3-month time period. In this table, N is the total number of SLD observations computed from glider and AXBT profiles. The mean difference

Table 1 Sonic Layer Depth (SLD) prediction errors of the NCODA 3DVAR and NCOM 4DVAR analyses, along with their ensuing 24, 48, 72, and 96-h forecasts. Errors are relative to the SLD computed from all AXBT and glider profile observations during Aug–Oct 2007. The experiments with the best correlation are in bold

	N	RMS error (m)	Correlation coefficient	Mean diff (m)
<i>Analysis</i>				
EXP1	5579	22.59	0.46	-9.07
EXP2	5579	18.07	0.65	-1.79
EXP3	5579	17.85	0.65	-2.02
<i>NCOM 24 h forecast</i>				
EXP1	5600	21.28	0.52	-7.96
EXP2	5600	19.14	0.61	-2.68
EXP3	5600	18.68	0.63	-2.84
<i>NCOM 48 h forecast</i>				
EXP1	5602	20.41	0.55	-7.14
EXP2	5602	19.71	0.59	-3.43
EXP3	5602	19.27	0.60	-3.58
<i>NCOM 72 h forecast</i>				
EXP1	5531	20.25	0.55	-6.75
EXP2	5531	19.82	0.58	-3.51
EXP3	5531	19.54	0.58	-3.76
<i>NCOM 96 h forecast</i>				
EXP1	5469	20.02	0.55	-6.18
EXP2	5469	19.83	0.57	-3.53
EXP3	5469	19.60	0.58	-4.05

(bias) between the SLDs calculated from the prediction system and data (Model SLD–Data SLD) reveal that the analysis and forecast systems are consistently predicting a shallower SLD than the data in all experiments. The RMS errors from these differences, along with their correlation coefficient, demonstrate that both versions of 4DVAR perform better than the NCODA 3DVAR at predicting SLD for the analysis and all forecast lengths.

Figure 9 displays 2D histograms of occurrence counts between matching observation (x-axis) and model (y-axis) SLDs. Here, the model SLD values are interpolated to each observation location and compared to the observed SLD. The SLD matchups are binned in 5 m resolution cells and the number of occurrences of each matchup within each cell is indicated by the colorbar. This is done for the analysis (top row), 24-h forecast (middle row), and 96-h forecast (bottom row) for EXP1 (left-most column), EXP2 (middle column), and EXP3 (right-most column). Note that the color bar is in log scale. Also, there are no observation counts shallower than 10 m (for both data and model), because the ProfParam software used to calculate SLDs does not allow for a SLD below 10 m. Therefore, more occurrence counts concentrated near the diagonal black line signifies that the model is doing well at predicting SLD.

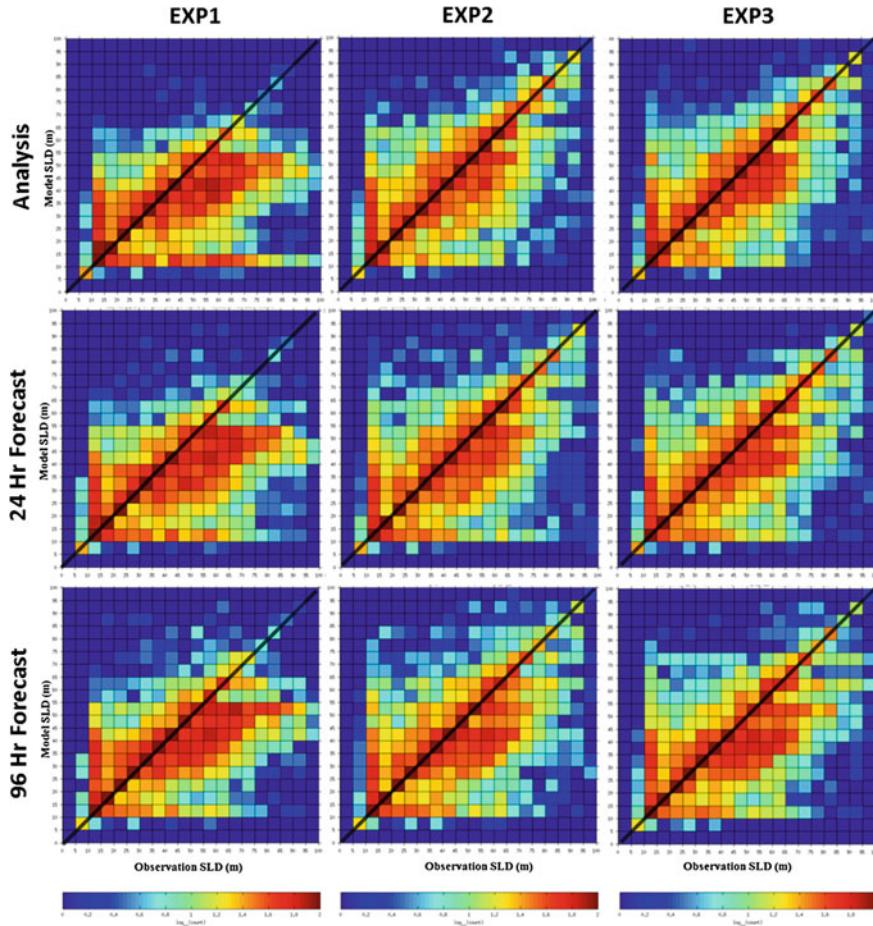


Fig. 9 Okinawa Trough 2D histograms of SLD (m) of NCODA 3DVAR (left), NCOM 4DVAR with synthetic SSH assimilation (middle), and NCOM 4DVAR with direct SSH assimilation (right) analyses (top), 24-h forecasts (middle) and 96-h forecasts (bottom) relative to SLD computed from profile observations during the 3-month time period of 1 August to 31 October 2007. The *diagonal black line* denotes the locations on each histogram where the modelled SLD matches the observed and the *color bar* denotes the number of counts on a log scale

In the EXP1 analysis histogram, there is an unusual band of modeled SLD counts between 10-15 m depth, and there is an overall significant bias towards the model under-predicting SLD relative to the observations (there are more red squares below and to the right of the diagonal black line). In the EXP3 analysis this bias is significantly reduced; and in the EXP2 analysis, one can barely notice the bias. As the forecast proceeds from 24 to 96-h, there is a clear trend of the shallow modelled SLD bias becoming more pronounced in the 4DVAR and the overall SLD prediction capability of 4DVAR moving towards that of EXP1. However, it can be

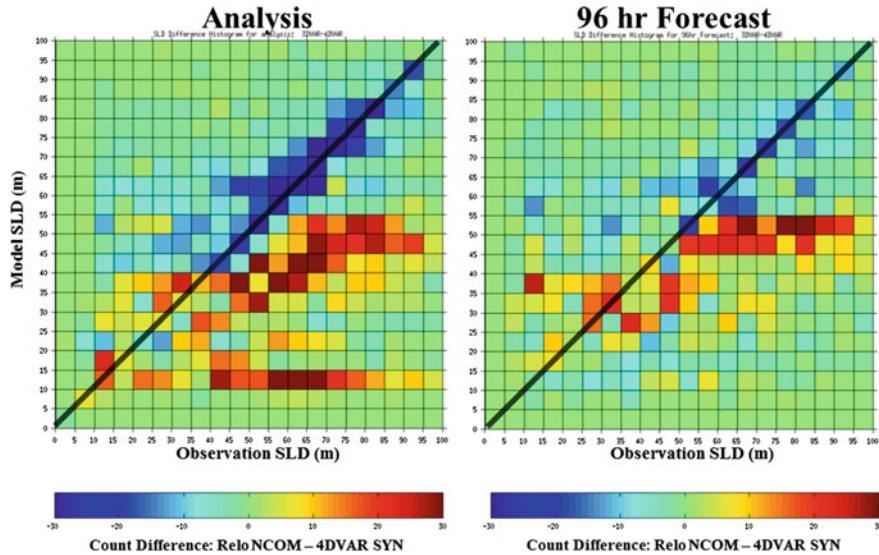


Fig. 10 2D histograms showing the difference in SLD counts between EXP1 and EXP2 analyses (left) and 96-h forecasts (right). *Blue (red) squares* signify that the NCOM 4DVAR has more (less) SLD combination counts than NCODA 3DVAR

seen that EXP2 does better than EXP3 at predicting SLD, and both 4DVAR systems perform significantly better than EXP1 (even after 96-h of forecasts).

To better visualize this improvement, Fig. 10 shows the difference in counts between EXP1 and EXP2 for both the analysis and the 96-h forecast. In this figure, a blue box signifies that EXP2 has more counts at that particular SLD comparison. The analysis histogram (left panel of Fig. 10) has mostly blue boxes along and near the diagonal, and red boxes below and to the right; it is clear that the 4DVAR is doing better and that the NCODA 3DVAR has a significant shallow SLD bias. This improvement persists throughout the 96 h of forecast (right panel of Fig. 10).

5 Operational Implementation of the NCOM-4DVAR

The majority of the NCOM 4DVAR experiments were performed at the DoD Supercomputing Resource Center (DSRC) where the average wall clock time for an analysis/forecast cycle was 70 min. Occasionally, if there were a significant number of observations during a cycle, the conjugate gradient solver would take up to ten iterations to converge, and hence take up to 1.5 h to complete a cycle. Most of the cycles, however, required an hour or less. This puts the time it takes to perform the NCOM 4DVAR for a relatively large domain within the operational constraints of NAVOCEANO.

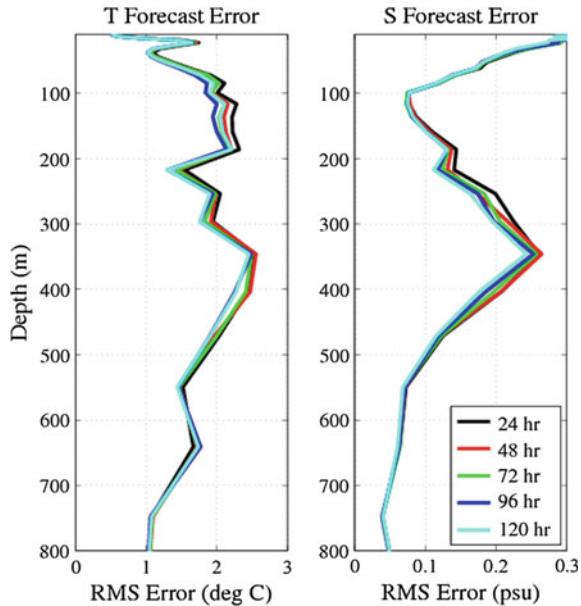
Although there is room for improvement, much effort was put into the parallelization of the NCOM 4DVAR software such that it can operate on as many processors as efficiently possible. The software is parallelized by splitting the horizontal domain into separate tiles that are each assigned to a processor (CPU). Periodically, throughout the operation of the code, information has to be transferred amongst the tiles. Therefore, as the number of CPUs is increased, so does the amount of data that needs to be transferred amongst the tiles, and there comes a limit when adding more CPUs for a particular domain only marginally decreases the total wallclock time (this is the software's scalability). The core of the NCOM 4DVAR software scales relatively efficiently. Through a number of tests, it was determined that the optimal CPU tile size for the NCOM 4DVAR is about 20×20 grid points, therefore, 192 CPUs was the optimal number for the Okinawa Trough domain.

To achieve the goal of performing an analysis/forecast cycle in about an hour, a software module was created to allow the analysis and the forecast to operate at different grid resolutions. As discussed in Sect. 3, this module is used to interpolate the high resolution forecast to a coarser resolution to be used as the background for the 4DVAR analyses. After the analysis, the module is used to interpolate the coarse analysis to the high resolution grid for the initial conditions of the forecast. In the Okinawa Trough experiments, the 4DVAR analysis is run at a 6 km resolution and the forecast is run at 3 km resolution. A number of experiments were performed testing the impact of reducing the resolution of the analysis component of the system and it was determined that the impact on forecast skill was negligible, but the reduction in computation time was tremendous.

Even though the NCOM 4DVAR takes significantly longer and requires more resources to operate than the 3DVAR (the average wall clock time to perform an analysis/forecast cycle with the 3DVAR is 5 min on 12 CPUs) the ability to correlate observations with the dynamics over multiple days significantly improves the analysis and ensuing forecast skill. Also, innovations are computed and applied in NCOM 4DVAR at the actual observation time. Whereas, the assimilation window for 3DVAR is only one day and regardless of when the observations are recorded, their innovations are all applied at a single analysis time.

The optimal assimilation window length for the NCOM 4DVAR can vary depending on the region, grid resolution, and the observations being assimilated. A longer assimilation window allows the observations more time to propagate throughout the domain via the model dynamics, which therefore should improve the model covariance and produce a better analysis. However, increasing the assimilation window increases the computational time and allows more time for small errors that may arise from the TL approximation to potentially grow. Figure 11 shows the impact the length of the assimilation window has on the predictability of the 24-h forecast. In this figure, experiments were performed on the Okinawa Trough domain during August 2007 for assimilation windows ranging from one to five days. The RMS errors of the 24-h forecast of temperature and salinity follow the correct pattern and generally decrease as the assimilation window is increased.

Fig. 11 Comparison of assimilation window lengths for the NCOM 4DVAR in the Okinawa Trough. RMS errors are computed for the 24-h forecasts of temperature (left) and salinity (right) during August 2007 using assimilation windows ranging from 24 to 120-h



The improvement resulting from increasing the assimilation window, however, is minor and doesn't warrant the extra computational cost. This is why just a 3-day assimilation window was used in the experiments presented in this chapter.

6 Conclusions

In this chapter, the accuracy of NCOM 4DVAR was compared to the operational 3DVAR-based Relo NCOM analysis/prediction system. Year-long experiments were performed for the Okinawa Trough domain. For three of the months in these experiments, the forecasts were extended out to 96 h so that the long-term predictability can be examined. A number of different types of observations were used in both the assimilation and validation: SST observations from satellites, subsurface temperature and salinity profile observations from ARGO floats, AXBTs and gliders, and SSH observations from altimeters. In some of the experiments, portions of the profile data were removed from the assimilation and used as an independent validation data set. The overall results from these experiments indicate that the NCOM 4DVAR analysis system, when assimilating SSH observations directly or through synthetic profiles of temperature and salinity, fits the assimilated observations within the prescribed observation error. Further, the resulting forecasts generated from the NCOM 4DVAR perform equally or better than the forecasts generated from the NCODA 3DVAR for subsurface temperature and salinity, model sea surface height, and sonic layer depth. Finally, it is demonstrated that

despite the computational requirements of the NCOM 4DVAR exceeding those of NCODA 3DVAR, they are within the operational constraints.

Acknowledgements The authors were supported by the NRL 6.4 NCOM 4DVAR Rapid Transition Project (projects 4727-04 and 4727-14), which was managed by both Space and Naval Warfare Systems Command under program element 063207N and the Office of Naval Research under program element 0602435N. Numerical simulations were performed at the DoD Supercomputing Resource Center (DSRC) with grants of computer time from the HPCMP Variational Assimilation High Performance Computing (HPC) subproject.

References

- Barron CN, Kara AB, Martin PJ, Rhodes RC, Smedstad LF (2006) Formulation, implementation, and examination of vertical coordinate choices in the global Navy Coastal Ocean Model (NCOM). *Ocean Modell* 11:347–375. doi:[10.1016/j.ocemod.2005.01.004](https://doi.org/10.1016/j.ocemod.2005.01.004)
- Barron CN, Kara AB, Rhodes RC, Rowley C, Shriner JF (2007) Validation test report for the 1/8° Global Navy Coastal Ocean Model Nowcast/Forecast System. NRL Tech Report NRL/MR/7320-07-9019. Naval Research Laboratory, Stennis Space Center, MS
- Bennet AF (1992) Inverse methods in physical oceanography. Cambridge University Press, New York 347 pp
- Bennet AF (2002) Inverse modeling of the ocean and atmosphere. Cambridge University Press, New York 234 pp
- Blumberg AF, Mellor GL (1983) Diagnostic and prognostic numerical circulation studies of the South Atlantic Bight. *J Geophys Res* 88:4579–4592
- Blumberg AF, Mellor GL (1987) A description of a three-dimensional coastal ocean circulation model. In: Heaps N (ed) Three-dimensional coastal ocean models. American Union, New York, N.Y., 208 pp
- Carrier MJ, Ngodock H (2010) Background-error correlation model based on implicit solution of a diffusion equation. *Ocean Modell* 35:45–53
- Carrier M, Ngodock H, Smith S, Jacobs G, Muscarella P, Ozgokmen T, Haus B, Lipphardt B (2014) Impact of assimilating ocean velocity observations inferred from Lagrangian drifter data using the NCOM 4DVAR. *Mon Weather Rev* 142:1509–1524
- Chua BS, Bennett AF (2001) An inverse ocean modeling system. *Ocean Modell* 3:137–165
- Courtier P (1997) Dual formulation of four-dimensional variational assimilation. *Q J R Meteorol Soc* 123:2449–2461
- Cummings JA (2005) Operational multivariate ocean data assimilation. *Q J R Meteorol Soc* 131:3583–3604
- Cummings JA (2011) Ocean data quality control. In: Schiller A, Brassington G (eds) Operational oceanography in the 21st century. Springer, pp 91–122. doi:[10.1007/978-94-007-0332-2_4](https://doi.org/10.1007/978-94-007-0332-2_4)
- Fox DN, Barron CN, Carnes MR, Booda M, Peggion G, Gurley JV (2002) The modular ocean data assimilation system. *Oceanography* 15:22–28
- Helber RW, Barron CN, Carnes MR, Zingarelli RA (2008) Evaluating the sonic layer depth relative to the mixed layer depth. *J Geophys Res* 113. doi:[10.1029/2007JC004595](https://doi.org/10.1029/2007JC004595)
- Hodur RM (1997) The naval research laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). *Mon Weather Rev* 125:1414–1430
- Hogan TF, Liu M, Ridout JA, Peng MS, Whitcomb TR, Ruston BC, Reynolds CA, Eckermann SD, Moskaitis JR, Baker NL, McCormack JP, Viner KC, McLay JG, Flatau MK, Xu L, Chen C, Chang SW (2014) The navy global environmental model. *Oceanography* 27(3):116–125
- Jacobs GA, Barron CN, Fox DN, Whitmer KR, Klingenberg S, May D, Blaha JP (2002) Operational altimeter sea level products. *Oceanography* 15(1):13–21

- Large WG, McWilliams JC, Doney SC (1994) Oceanic vertical mixing: a review and a model with a nonlocal boundary layer parameterization. *Rev Geophys* 32:363–403
- Martin PJ, Peggion G, Yip KJ (1998) A comparison of several coastal ocean models. NRL Report NRL/FR/7322-97-9692, Naval Research Laboratory, Stennis Space Center, MS
- Martin P (2000) Description of the Navy Coastal Ocean Model Version 1.0. NRL report NRL/FR/7322-00-9961, Naval Research Laboratory, Stennis Space Center, MS
- Martin PJ, Barron CN, Smedstad LF, Wallcraft AJ, Rhodes RC, Campbell TJ, Rowley C, Carroll SN (2008) Software design description for the Navy Coastal Ocean Model Version 4.0. NRL Report NRL/MR/7320-08-9149, Naval Research Laboratory, Stennis Space Center, MS
- Mellor GL, Yamada T (1982) Development of a turbulence closure model for geophysical fluid problems. *Rev Geophys Space Phys* 20:851–875
- Metzger EJ, Smedstad OM, Thoppil PG, Hurlburt HE, Cummings JA, Wallcraft AJ, Zamudio L, Franklin DS, Posey PG, Phelps MW, Hogan PJ, Bub FL, Dehaan CJ (2014) US Navy operational global ocean and arctic ice prediction systems. *Oceanography* 27(3). <http://dx.doi.org/10.5670/oceanog.2014.66>
- Ngodock HE (2005) Efficient implementation of covariance multiplication for data assimilation with the representer method. *Ocean Model* 8(3):237–251
- Ngodock HE, Carrier M (2014a) A 4DVAR system for the Navy Coastal Ocean Model Part I: System description and assimilation of synthetic observations in Monterey Bay. *Mon Weather Rev* 142. doi:[10.1175/MWR-D-13-00221](https://doi.org/10.1175/MWR-D-13-00221)
- Ngodock HE, Carrier M (2014b) A 4DVAR system for the Navy Coastal Ocean Model Part II: strong and weak constraints assimilation experiments with real observations in the Monterey Bay. *Mon Weather Rev* 142. doi:[10.1175/MWR-D-13-00220](https://doi.org/10.1175/MWR-D-13-00220)
- Ngodock HE, Carrier M, Souopgui I, Smith S, Martin P, Muscarella P, Jacobs G (2015) On the direct assimilation of along-track sea surface height observations into a free-surface ocean model using a weak constraints four dimensional variational (4DVAR) method. *Q J R Meteorol Soc* (in press). doi:[10.1002/qj.2721](https://doi.org/10.1002/qj.2721)
- Rosmond TE (1992) The design and testing of the Navy Operational Global Atmospheric Prediction System. *Weather Forecast* 7:262–272
- Rosmond TE, Teixeria J, Peng M, Hogan TF, Pauley R (2002) Navy Operational Global Atmospheric Prediction System (NOGAPS): forcing for ocean models. *Oceanography* 15:99–106
- Rowley C (2010) Validation test report for the Relo system. NRL Report NRL/MR/7320-10-9216, Naval Research Laboratory, Stennis Space Center, MS
- Smagorinsky J (1963) General circulation experiments with the primitive equations. I: The basic experiment. *Mon Weather Rev* 91:99–164
- Smith SR, Cummings JA, Rowley C, Chu P, Shriner J, Helber R, Spence P, Carroll S, Smedstad OM (2012) Validation test report for the Navy Coupled Ocean Data Assimilation 3D Variational Analysis (NCODA-VAR) System, Version 3.43. NRL Memorandum Report NRL/MR/7320-11-9363, Naval Research Laboratory, Stennis Space Center, MS
- Smith SR, Carrier MJ, Ngodock HE, Shriner J, Muscarella P (2015) Validation testing report for the Navy Coastal Ocean Model Four-Dimensional Variational Assimilation (NCOM 4DVAR) System, Version 1.0. NRL Memorandum Report NRL/MR/732-14-9574, Naval Research Laboratory, Stennis Space Center, MS
- Weaver A, Courtier P (2001) Correlation modeling on the sphere using a generalized diffusion Equation. *Q J R Meteorol Soc* 127:1815–1846
- Yaremchuk M, Carrier M, Smith S, Jacobs G (2013) Background error correlation modeling with diffusion operators. In: Park SK, Xu L (eds) *Data assimilation for atmospheric, oceanic and hydrologic applications*, vol 2. Springer, Berlin, Heidelberg. doi:[10.1007/978-3-642-35088-7_15](https://doi.org/10.1007/978-3-642-35088-7_15)
- Yu P, Kurapov AL, Egbert GD, Allen JS, Kosro AP (2012) Variational assimilation of HF radar surface currents in a coastal ocean model off Oregon. *Ocean Model* 49–50:86–104

Stratospheric and Mesospheric Data Assimilation: The Role of Middle Atmospheric Dynamics

Saroja Polavarapu and Manuel Pulido

Abstract The middle atmosphere refers to the stratosphere and mesosphere and features dynamics and circulations that are fundamentally different from those of the troposphere. The large-scale meridional circulations in the middle atmosphere operate on seasonal and longer time scales and are largely forced by the breaking of upward propagating waves. The winter stratosphere is dominated by large-scale waves and a polar vortex which confines constituents and which is sometimes punctuated by stratospheric sudden warmings. In contrast, the summer stratosphere is quiescent. Meanwhile, the meridional circulation in the mesosphere is mainly driven by the breaking of a broad spectrum of gravity waves that have propagated upward from the troposphere. These facets of middle atmosphere dynamics have implications for, and pose unique challenges to, data assimilation systems whose models encompass this region of the atmosphere. In this work, we provide an overview of middle atmosphere data assimilation in the context of the dynamics of this region. The purpose is to demonstrate how the dynamics can be used to explain the behavior of data assimilation systems in the middle atmosphere, and also to identify challenges in assimilating measurements from this region of the atmosphere. There are two overarching themes. Firstly, we consider the vertical propagation of information through waves, resolved and parameterized, and background error covariances. Secondly, we delve into the dynamical sources of model errors and techniques for their estimation.

S. Polavarapu (✉)

Environment and Climate Change Canada, Toronto, ON, Canada

e-mail: Saroja.Polavarapu@canada.ca

M. Pulido

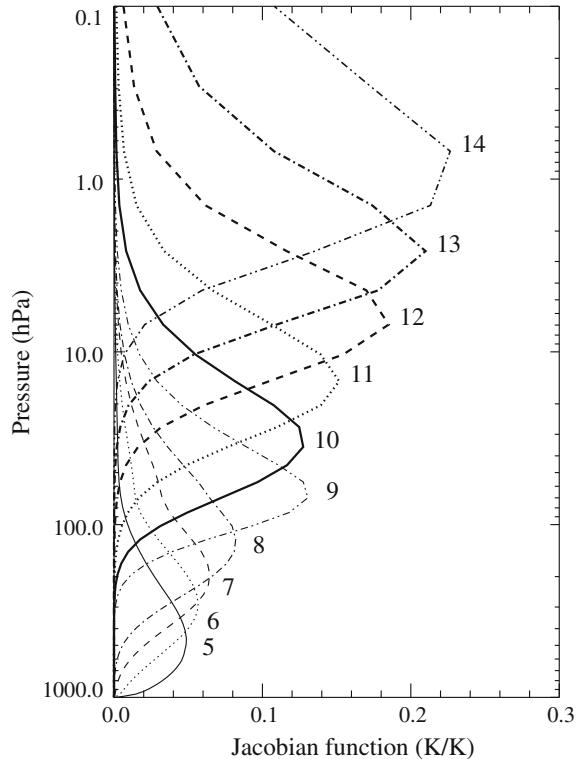
Department of Physics, FACENA,

Universidad Nacional del Nordeste and CONICET, Corrientes, Argentina

1 Introduction

The past decade has seen operational weather forecasting centers raise their model lids to the middle or upper mesosphere. The NASA's Global Modeling and Assimilation Office (GMAO) raised its model lid height to 80 km in January 2004. The European Centre for Medium-Range Weather Forecasts (ECMWF) did the same in February 2006, as did the Met Office in November 2009. The Canadian Meteorological Centre (CMC)'s model lid moved to 0.1 hPa (roughly 65 km) in June 2009 (Charron et al. 2012). Given that the primary focus of an operational weather forecasting centre is on producing real time forecasts for the troposphere, the extra computational expense invested in raising the model lid to such heights merits an explanation. There are two main motivating factors for making this change. On the one hand, a precise model representation of the stratosphere is expected to increase predictive skill of extended range (10 days to subseasonal) forecasts (Tripathi et al. 2014; Gerber et al. 2012; Charlton et al. 2004, 2005b) but more importantly there is a significant number of nadir satellite observations that are sensitive to the middle atmosphere, Fig. 1 shows the normalized weighting functions from the Advanced Microwave Sounding Unit (AMSU)-A instrument.

Fig. 1 AMSU-A weighting functions on the 43 RTTOV levels computed using the US standard atmosphere. Each function is identified by a channel number on the right. Only weighting functions for channels 5 to 14 are shown. Figure courtesy of D. Shawn Turner



Several channels (12–14) exhibit peak sensitivity to upper stratospheric temperature and many others (6–11) have peak or significant sensitivity to lower or mid stratosphere temperature. In order to assimilate these channels, a model would need a good background forecast up to 0.1 hPa, so its sponge layer should begin above this level. This then implies a model lid in the middle mesosphere. Thus the stratosphere and most of the mesosphere are now part of the weather forecasting domain.

Since weather forecast models form the basis of major reanalysis efforts such as ERA-Interim (Dee et al. 2011), JRA-55 (Kobayashi et al. 2015) or MERRA (Rienecker et al. 2011), there are now long time series of analyses of the middle atmosphere which serve the climate community. These reference datasets are widely used for assessing and validating climate models, for understanding measurements and driving chemistry transport models (CTMs). Thus a whole community is inspecting assimilation products in the middle atmosphere, providing feedback on successes and deficiencies. In particular, middle atmosphere dynamicists, climatologists and chemists have noted differences in the various reanalysis products and established an international effort (Stratospheric Processes and their Role in Climate (SPARC) Ranalysis Intercomparison Project or S-RIP) to systematically compare the products and provide guidance to climate scientists as to where and when they are reliable as well as to data assimilators (<http://www.sparc-climate.org/activities/reanalysis/>). The very existence of S-RIP points to the value placed on middle atmosphere analyses by climate scientists.

There are yet other reasons for assimilating observations of the middle atmosphere, such as driving models with chemistry to quantify stratospheric ozone loss (Shepherd et al. 2014), assessing changes in the transport of constituents (and the Brewer-Dobson circulation) (Hegglin et al. 2014), better understanding large-scale dynamic events such as sudden stratospheric warmings and their potential impacts or responses under climate change scenarios and providing lower boundary conditions for space weather models. Finally, the stratosphere has a memory that can be exploited for improving the skill of seasonal forecasts (Stockdale et al. 2015; Sigmond et al. 2013; Marshall and Scaife 2009; Boer and Hamilton 2008) providing yet another motivation for the assimilation of observations into models capable of depicting the middle atmosphere.

Middle atmosphere data assimilation has therefore become increasingly relevant to operational and research atmospheric data assimilation efforts. While the same general techniques of data assimilation are applied to all regions of the atmosphere simultaneously, there are reasons to afford middle atmosphere data assimilation separate consideration. Firstly, the dynamics of the middle atmosphere differ from those of the troposphere and, as will be shown below, this has important implications for understanding the behavior of data assimilation systems. Secondly, the degree to which a forecast model captures these dynamics provides insight into the nature of model errors in this region of the atmosphere. Finally, the observing system targeting this region of the atmosphere poses unique challenges. While radiosondes have formed the backbone of tropospheric assimilation systems (whether directly or through the anchoring of bias correction schemes applied to

satellite measurements), they only reach as high as the middle stratosphere, leaving much of the middle atmosphere devoid of in situ measurements. Operational satellites sensing the middle atmosphere are primarily nadir sounders such as AMSU (Advanced Microwave Sounding Unit). Since these instruments are sensitive to broad layers of the atmosphere, the vertical structure of the stratosphere and mesosphere cannot be well resolved. Fortunately, with the routine assimilation of Global Positioning System (GPS) Radio Occultation (RO) measurements, the vertical structure of the lower to mid stratosphere can be better resolved (Cardinali and Healy 2014; Healy and Thépaut 2006). However, above 35 km, the data become noisier and only weakly constrain the full state despite effectively constraining the bias of satellite measurements. Thus, the vertical structure of the atmosphere above the mid stratosphere remains difficult to estimate.

The goal of this chapter is to present a subjective overview of middle atmosphere data assimilation, from the perspective of middle atmosphere dynamics but targeted to a data assimilation audience. The intent is not to be exhaustive, but rather, illustrative, while maintaining a focus on identifying particular issues and challenges in the middle atmosphere of coupled troposphere-middle atmosphere data assimilation systems. Constituent assimilation is an enormous topic that is relevant for the middle atmosphere, but one that is not considered here. The distribution of constituents is determined by atmospheric transport and mixing as well as chemical reactions. An overview of transport in the middle atmosphere is provided by Andrews et al. (1987) and Shepherd (2007) while Monge-Sanz et al. (2013) discuss the role of assimilation techniques on constituent transport in the context of chemistry transport models. The assimilation of constituents also has the potential to improve wind analyses in the middle atmosphere (e.g. Allen et al. 2014, 2015; Milewski and Bourqui 2011, 2013; Semane et al. 2009). Recent reviews of chemical data assimilation are provided by Bocquet et al. (2015) and Sandu and Chai (2011).

The chapter is organized as follows. In Sect. 2 a brief overview of middle atmosphere dynamics is provided while in Sect. 3 we look at middle atmosphere data assimilation through the lens of middle atmosphere dynamics. Some additional challenges and issues of middle atmosphere data assimilation not previously mentioned are discussed in Sect. 4 and a summary of the chapter is presented in Sect. 5.

2 Brief Overview of Middle Atmosphere Dynamics

In order to understand how forecasting and assimilation systems respond to perturbations such as analysis increments, it is necessary to introduce a few concepts about middle atmosphere dynamics. A very brief introduction of the dynamic features that have some impact on data assimilation is given here, but more detailed accounts are available in textbooks (e.g. Andrews et al. 1987; Vallis 2006) and articles (e.g. Shepherd 2000, 2002, 2007; McLandress 1998; Smith 2004).

The middle atmosphere refers to the stratosphere and mesosphere and extends from roughly 10 to 80 km above the Earth's surface. Temperature increases with height in the stratosphere due to the absorption of ultraviolet radiation by ozone and decreases with height in the mesosphere as the ozone concentration drops off. The stratosphere is statically stable and the climatological winds are to a first approximation zonal. If we consider the two-dimensional, steady, geostrophic and hydrostatic equations, in the absence of a momentum source, the atmosphere would be in radiative equilibrium balance with outgoing terrestrial radiation balancing incoming solar radiation. This means a cold dark winter pole and a warm sunlit summer pole. Through thermal wind balance, zonal winds increase with height. However, as shown in Fig. 2a, radiative-equilibrium temperature calculations yield temperatures near the winter pole (dashed lines) that are far too cold compared to observed values (solid lines), and zonal wind speeds that are much too strong (compare dashed and solid lines in Fig. 2b).

The addition of a simple Rayleigh friction term as a forcing term in the zonal momentum equation which is linearly proportional to zonal wind with the proportionality constant increasing with height (McLandress, 1988) results in temperatures and zonal wind speeds which are closer to observations (compare dotted and solid lines in Fig. 2) than the radiative equilibrium solution (dashed lines). The conclusion is that some kind of momentum source/sink is needed to explain the observed zonal mean temperatures and winds, as was first hypothesized by Leovy (1964).

The origin of this momentum source/sink is breaking waves, which exert a drag or forcing on the zonal mean flow and drive a mean meridional circulation. In the winter stratosphere, large-scale quasi-stationary Rossby waves forced by topography and land-sea contrasts are able to propagate upward through the stratospheric

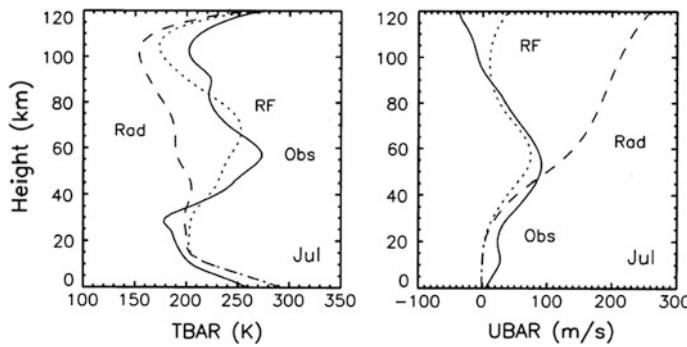


Fig. 2 Zonal mean temperature at 90 S (left) and zonal wind field at 40 S (right) during austral winter. Observations from the COSPAR International Reference Atmosphere (CIRA) are shown as solid lines. (COSPAR = COmmittee on SPAce Research). The profiles obtained from radiative equilibrium are shown in dashed curves. Assuming a momentum force that is negative in the winter hemisphere and positive in the summer hemisphere yields a better fit to observations (dotted curves). (Reprinted from McLandress (1998) with permission from Elsevier.)

westerlies where they increase in amplitude as density decreases. Eventually they break, impart their (negative) momentum to the zonal mean flow, exerting a drag on the wintertime westerlies. This creates poleward motion through a Coriolis torque and by continuity, descent (and warming through adiabatic compression) over the winter pole. Thus, large-scale waves drive this thermally-indirect circulation, called the Brewer-Dobson circulation. The Brewer-Dobson circulation is important not only for explaining stratospheric temperature distributions, but also for transporting constituents, as is apparent in the accumulation of ozone over the winter pole in Fig. 3. The conditions for vertical propagation of quasi-stationary Rossby waves (see Andrews et al. 1987, Chap. 4.5 or Vallis 2006, Chap. 13.3) in the case of a constant wind (U) are that $U > 0$ (eastward) and U remains below a critical value (U_c). Thus, these waves cannot propagate into the stratosphere in summer when zonal winds are easterly. Furthermore, in the winter when they can propagate vertically, large-scale waves (wavenumbers 1 to 3) are favoured because the critical wind speed (U_c) decreases rapidly with increasing wavenumber. Thus the winter stratosphere is dominated by waves having large horizontal scales. Due to the fact that quasi-stationary Rossby waves are unable to propagate in easterly zonal wind, the summer stratosphere is characterized by an absence of these waves and so by temperatures closer to radiative equilibrium.

The stratospheric jets also act to filter much smaller-scale waves (i.e., gravity waves) which would otherwise propagate up to the mesosphere. In winter when

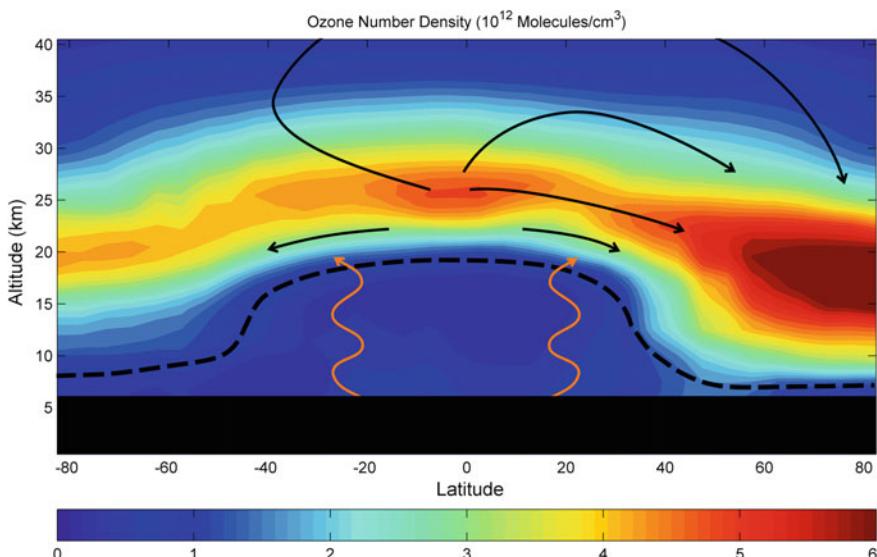


Fig. 3 Cartoon of the Brewer-Dobson circulation. Meridional circulation is indicated by *black* arrows. The tropopause is indicated by a heavy *dashed line*. The ozone distribution for 15 Feb.–31 March 2004 from OSIRIS is shown in colours with values indicated by a colour bar on the *bottom*. After Shaw and Shepherd (2008)

stratospheric winds are westerly and increasing with height, gravity waves with eastward phase speeds may reach their critical level (where the zonal phase velocity equals the zonal wind) in the stratosphere. This removal or “filtering” of eastward propagating waves at their critical levels leads to predominantly westward propagating gravity waves reaching the mesosphere (assuming a westward-eastward isotropic launch gravity wave spectrum). When those waves break in the mesosphere they deposit westward momentum (Lindzen 1981). Similarly, in the summer hemisphere, easterly winds filter westward propagating gravity waves at their critical levels, so that gravity waves which break in the mesosphere deposit eastward momentum. In the mesosphere, this deceleration of the westerlies in the winter hemisphere and deceleration of the easterlies in the summer hemisphere caused by gravity wave momentum deposition create a poleward motion in the winter hemisphere, but equatorward motion in the summer hemisphere. By continuity, there is descent over the winter pole and ascent over the summer pole. Thus small scale gravity waves are responsible for driving the pole-to-pole Murgatroyd-Singleton circulation in the mesosphere seen in the upper part of Fig. 3.

Since the spectrum of Rossby waves propagating upward in the winter stratosphere is dominated by the largest modes, as synoptic scale waves are filtered in the lower stratosphere, Rossby waves are expected to be well represented in global numerical models. On the other hand, gravity waves are small-scale waves which propagate almost vertically. Therefore, global numerical models are typically not able to directly represent these waves. In order to incorporate the momentum forcing produced by small-scale gravity waves in global numerical models, the drag exerted by the upward propagation and breaking of small-scale gravity waves on the zonal mean flow is accounted for through “gravity wave drag” parameterizations. These parameterizations are divided in two groups: parameterizations of waves of orographic origin and non-orographic gravity wave drag schemes. The generation characteristics of orographically generated waves are well known and have been shown to have a large impact in the lower stratosphere and in particular on the Brewer-Dobson circulation (Li et al. 2008; McLandress and Shepherd 2009). Gravity waves from other sources such as fronts, convection, and geostrophic adjustment are modelled through non-orographic gravity wave parameterizations (e.g. Hines 1997; Scinocca 2003) and are expected to have a large impact on the upper stratosphere and the mesosphere.

3 Impact of Middle Atmosphere Dynamics on Data Assimilation

The fact that the middle atmosphere is largely driven by waves propagating up from the troposphere has implications for data assimilation. The fundamental difference in stratospheric dynamics between winter and summer also impacts data assimilation results and inputs (such as background error covariances). Finally, the

importance of gravity waves (that are generated in the troposphere and propagate upward) to the mesospheric circulation means that these signals (which are frequently treated as noise in the troposphere) might need to be better simulated or estimated in the troposphere and lower stratosphere. In this section, we explore how middle atmosphere dynamics impact the inputs and results of data assimilation systems.

3.1 *Upward Propagation of Information*

3.1.1 Propagation of Information and Errors Through Resolved Waves

Figure 4 shows averaged analyzed temperature vertical profiles (over the global domain) from a set of data assimilation experiments using a 3D-variational (3D-Var) system for which only the strength of an externally applied filter is varied. As a result of changing the strength of the filter, a large impact on mesospheric temperatures is found. Specifically, the stronger the filter, the colder the global mean mesopause temperature. A difference of 20 K at 90 km is seen between experiments. These results were surprising because the Canadian Middle Atmosphere Model (CMAM) (Scinocca et al. 2008) which was employed in the data assimilation system extends to about 95 km but the observations were inserted only below about 45 km. Thus the filter was targeting imbalance arising from increments below 45 km. Yet below 45 km, the temperature profile averaged over all coincident measurement locations was virtually identical regardless of which filter was employed. This is because the averaging over all profiles smooths whatever degree of noise is present in the profiles obtained with different filters. However, the waves defined by the increments in the troposphere (whether real or spurious) propagate up to the mesosphere where they break and create a drag (momentum forcing) which impacts the zonal mean flow. If they do not break by 80 km they encounter the model sponge layer which is designed to absorb such waves thus preventing their reflection from the model lid at 95 km. The sponge layer is a numerical device acting on departures from the zonal mean flow (Shepherd et al. 1996) that is intended to mimic the fact that in the absence of the model lid these waves would reach higher altitudes where they would break, deposit momentum and create turbulent dissipation generating heating. Eventually the radiation to space balances the wave-generated heating resulting in the temperature profiles seen in Fig. 4. Thus although the heating is artificially created here through the enforced dissipation of these upward propagating waves, the same process would occur in the real atmosphere though at higher altitudes because of molecular viscosity. Indeed Lübken et al. (2002) show that gravity waves are an important source of heating in the mesopause region. Thus, in Fig. 4, a stronger, more dissipative filter results in smaller wave momentum flux reaching the mesosphere and less enforced wave breaking and resultant heating in the sponge layer. This hypothesis was confirmed

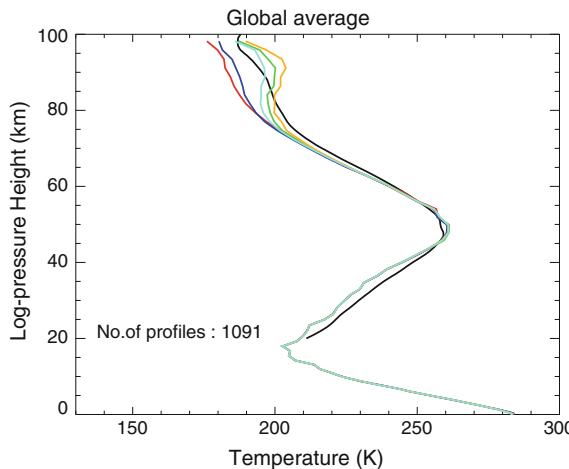


Fig. 4 Average of CMAM-DAS temperature profiles sampled at SABER locations over the globe during 25 January 2002. The temperatures are from analyses obtained from assimilation experiments which were identical except for the externally applied filter. In all cases, observations were assimilated below 45 km only. The colours are black (SABER data), cyan (DF with 12-h cutoff), yellow (DF with 6-h cutoff), green (IAU with 6-h cutoff), blue (IAU with 4-h cutoff), and red (IAU with constant coefficients). Filter strength increases as follows: yellow-green-cyan-blue-red. (Reprinted from Sankey et al. (2007) with permission from the Crown.)

by comparing the temperature variance of time series of analyses from the various experiments (Sankey et al. 2007). The stronger the filter, the smaller the variance. Thus resolved waves in the troposphere and stratosphere can propagate up to the mesosphere and impact the zonal mean or even the global mean flow. The implication is that tropospheric tuning of data assimilation systems can have large impacts on mesospheric analyses. On the other hand, the sensitivity of the mesosphere can also be used to tune assimilation parameters (such as filter strength, as was done in Sankey et al. 2007).

Nezlin et al. (2009) demonstrated that even without observations above 45 km, the large scale dynamics (up to wavenumber 10) in the mesosphere could be improved. They also showed that the quality of mesospheric analyses was sensitive to the accuracy of observations taken below 45 km. Both of these facts attest to the vertical propagation of information. (Here we use the term “information” to describe that part of the true atmospheric signal that a given model can resolve.) The observations constrain the atmospheric signal at a given height, and then the dynamics of the model propagates this observational information upward producing an impact at heights where no observations had been assimilated. Since the middle atmosphere is largely forced by upward propagating waves, both information and errors propagate vertically through waves in data assimilation systems. While Nezlin et al. (2009) demonstrated that the vertical propagation of information in a 3D-Var system (the CMAM-Data Assimilation System (DAS)) is theoretically possible through the use of simulated observations, Xu et al. (2011a, b)

demonstrated that CMAM-DAS mesospheric winds do indeed compare well to independent measurements on long time scales. This confirms that vertical propagation of information from the troposphere to the mesosphere actually occurs in assimilation systems since no observations in the mesosphere had been assimilated. The same effect is seen in the intraseasonal variability of mesospheric zonal-mean temperature and constituents (carbon monoxide) in a set-up where the CMAM is nudged towards the ERA-Interim reanalysis in the troposphere and stratosphere, and the model is seen to agree well with MLS (Microwave Limb Sounder) observations in the mesosphere (McLandress et al. 2013).

Even without mesospheric observations, the migrating diurnal and semi-diurnal tidal signals in the mesosphere can be captured (Sankey et al. 2007; Wang et al. 2011; Xu et al. 2011a, b; Hoppel et al. 2013). Since these signals are generated by the absorption of solar radiation by water vapour in the troposphere and by ozone in the stratosphere, observations from the troposphere and stratosphere constrain these signals well. Thus agreement of a model's mesospheric tidal amplitudes with observations indicates that the vertical propagation of the signal into the mesosphere is at least partially captured by the model. Of course, assimilation of mesospheric observations, greatly improves agreement with observations (Hoppel et al. 2013).

3.1.2 Propagation of Information Through Background Error Covariances

The impact of background error covariances on analysis increments is most easily illustrated through experiments in which a single observation is assimilated. To illustrate the vertical structure of analysis increments, a one-dimensional variational assimilation is performed using only AMSU channel 11. The two orders of magnitude increase with height in forecast error variance seen in the bottom left panel of Fig. 5 largely reflects the increasing amplitude of gravity waves from the stratosphere to the mesosphere. As a result, spurious analysis increments in the mesosphere can be produced (top left panel) when the large forecast error variances are combined with small but nonzero correlations in the wings of the weighting function. For example, the analysis increment at 0.01 hPa in the top left panel of Fig. 5 appears at a height where the weighting function for an observation at 10 hPa is virtually zero (i.e. the wings of the weighting function) (top right panel) because the correlation function (bottom right panel, black curve) is not exactly zero at 0.01 hPa. Setting such tiny correlations (which are due to statistical noise) to exactly zero removes much of the spurious mesospheric analysis increments (dashed lines in top left panel). In fact, removing such spurious increments in the mesosphere is imperative when an assimilation system assimilates no mesospheric observations and is therefore unable to damp such errors. In the mesosphere, such spurious increments may be persistent (because of the presence of model and/or observation biases) and can actually lead to physically nonsensical results after only a few weeks of assimilation. Thus information propagated to the mesosphere

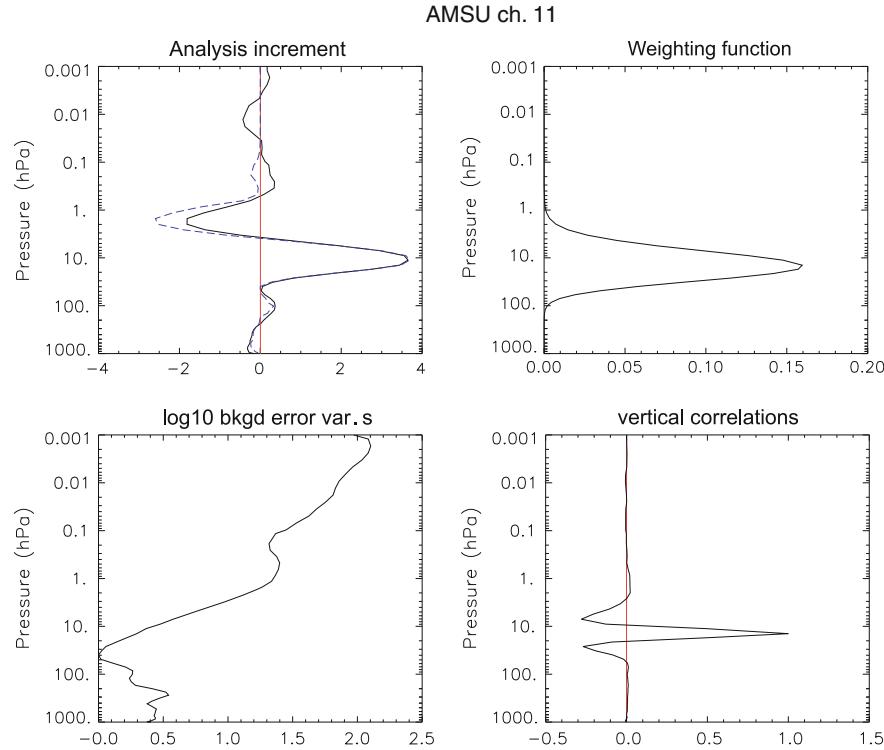


Fig. 5 A 1-D assimilation of AMSU channel 11. *Top left*: Temperature analysis increments obtained when vertical correlations are unmodified (solid) or modified so that near zero values are exactly zero (dashed). *Top right*: weighting function for AMSU-A channel 11. *Bottom left*: \log_{10} of temperature background error variance used with the CMAM-DAS. *Bottom right*: A sample vertical correlation function with peak amplitude near 10 hPa. (Reprinted from Polavarapu et al. (2005) with permission from the Crown.)

through background error covariances is not necessarily desirable. Similarly, erroneous small scale vertical structures in background error covariances cannot be damped by measurements if the observing system is lacking in detailed vertical information. This is the case in the upper stratosphere where nadir temperature sounders are the dominant source of information.

Since observations of the middle atmosphere are predominantly from satellite-based radiance measurements, the problem of vertical localization of covariances needed for some ensemble-based data assimilation techniques in radiance space is worth noting. The issue is that these types of measurements are related to model quantities integrated in space so that the concept of localization is unclear. Yet the localization of error covariances is important for practical application of ensemble Kalman filters to comprehensive and complex meteorological models. Such localization is frequently done in observation space for computational expediency. The problem identified by Campbell et al. (2010) is that location and

distance are ill defined quantities in radiance space so that observation-space localization applied to the usual case of channels with overlapping sensitivities cannot achieve the correct Kalman gain even when observations are known to be perfect. Ideally, the vertical localization must therefore be at least as broad as the weighting functions but not so broad that the suppression of spurious correlations due to sampling errors become ineffective. On the other hand, localization in model space does not suffer from this problem.

3.1.3 Propagation of Information Through Gravity Wave Drag Schemes

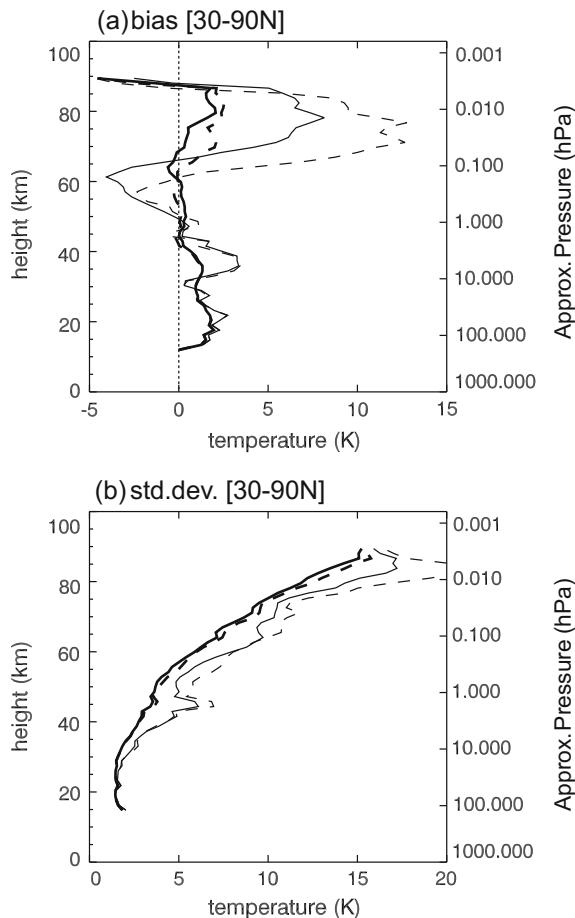
Gravity wave drag (GWD) schemes can also propagate information from the troposphere and stratosphere to the mesosphere. GWD schemes parameterize (represent simplifications of) the processes of gravity wave generation in the troposphere, vertical propagation and nonlinear saturation. The output of such a scheme is a drag or forcing term for the momentum equations. GWD schemes are needed in climate models because the lack of representation of small-scale gravity waves leads to insufficient forcing of the meridional circulation and insufficient downwelling (and warming) over the winter pole (as well as insufficient upwelling and cooling over the summer pole). Thus, without a GWD scheme, climate models can suffer from the “cold pole” problem, which is particularly evident in the southern hemisphere where there are fewer forced planetary waves (Austin et al. 2003).

Apart from insufficient horizontal model resolution, there are a number of factors that contribute to the poor representation of small-scale gravity waves in numerical models. (a) There is insufficient vertical resolution particularly in the middle atmosphere where most of the models decrease their vertical resolution. An exception is the Kanto model which due to its high vertical resolution is able to represent a large part of the gravity wave spectrum and thus the gravity wave forcing produced in the middle atmosphere. It is thus is able to represent a rather realistic meridional circulation (Watanabe et al. 2008). On the other hand, an inconsistent ratio of vertical to horizontal resolution (i.e. high horizontal resolution and low vertical resolution) may produce spurious gravity waves (Fox and Lindzen 1993, Iga et al. 2007). (b) In general Eulerian and semi-Lagrangian schemes are diffusive so that they do not represent well waves that are under-resolved (waves with wavelengths larger but close to the model resolution). (c) Finally, time update schemes also affect the resolved waves. Thus, gravity wave drag parameterizations must represent the unresolved part of the spectrum of gravity waves as well as compensating for model deficiencies when reproducing the resolved gravity wave spectrum.

GWD schemes can also propagate information vertically in data assimilation systems (Ren et al. 2008). The parameterized subgrid-scale gravity waves in GWD schemes are affected by observations of tropospheric and stratospheric winds, which in turn affect the forcing produced by the GWD schemes in the mesosphere (McLandress et al. 2013). The benefit of a GWD scheme on mesospheric analyses

was demonstrated by Ren et al. (2011). Background or 6-h forecasts were closer to independent observations of mesospheric temperature (from SABER retrievals) when a GWD scheme was used (Fig. 6). The benefit was quite large if no mesospheric observations were assimilated, but still apparent even if they were assimilated. Since mesospheric analyses obtained with a model using a GWD scheme but with no mesospheric observations were close to independent measurements, it is evident that GWD is able to propagate useful information to the mesosphere. At ECMWF, the same GWD scheme used in Ren et al. (2011) was implemented operationally, and shown to improve the bias in temperature at the stratopause at the winter pole in 5-day forecasts (Orr et al. 2010).

Fig. 6 Fit of 6 h temperature forecasts to SABER observations during 1–14 February 2006 using the CMAM data assimilation system. Assimilation cycles were run with parameterized gravity wave drag scheme (*solid lines*) or without it (*dashed lines*). Results from experiments in which SABER temperatures were assimilated are shown in bold while those from experiments that did not assimilate those measurements are shown in thin curves *blue*. Panel (a) shows the bias from these 4 experiments while panel (b) shows the standard deviation for the northern hemisphere high latitudes. (Reprinted from Ren et al. (2011) with permission from the Crown.)



3.2 *Understanding Forecast Improvements*

The winter polar stratosphere is dominated by westerly winds that increase with height and define a polar vortex—polar night jet. In the Northern Hemisphere, this vortex is occasionally disrupted by stratospheric sudden warming (SSW) events during which temperatures can rise dramatically (by 50 K in one week). Simultaneously, the climatological westerly winds weaken and may even become easterly. Mesospheric coolings can also occur in conjunction with stratospheric warmings. Since SSW events are primarily driven by planetary waves propagating up from the troposphere, such events involve vertical coupling from the troposphere to the mesosphere. Baldwin and Dunkerton (2001) showed that the dominant mode of slowly varying wintertime variability called the Northern Annular Mode (or NAM) has a spatial structure which is similar from the surface to over 50 km altitude, thus indicating a coupling of the troposphere and stratosphere. (At the surface the pattern is sometimes called the Arctic Oscillation.) The NAM pattern at 10 hPa is a disk of similarly signed values around the pole with oppositely signed values in a ring or annulus around this. A projection of the geopotential height onto this pattern indicates the relative strength of the polar vortex. A strongly positive projection indicates a stronger than normal polar vortex, while a strongly negative projection indicates a weaker than normal vortex. Moreover, when time series of strongly positive or negative NAM events are composited, the vertical coupling becomes apparent. Specifically, a large stratospheric event, such as an SSW, will appear in the mid-stratosphere (10 hPa is often used as a reference level) about ten days prior to its appearance at the surface. Once the NAM signal appears in the troposphere (300 hPa), the same sign of the NAM index persists in the troposphere for around 60 days. During this time, the troposphere is characterized by a particular climatology. For instance, during a strong vortex event, cool winds would flow over eastern Canada, North Atlantic storms would bring rain and mild temperatures to northern Europe and drought conditions would prevail in the Mediterranean (Thompson and Wallace 2001). Thus, the stratospheric modulation of tropospheric climate suggests a predictive skill which can be exploited on the week to seasonal timescales (e.g. Douville 2009). Charlton et al. (2004, 2005b) also showed that stratospheric initial conditions can impact tropospheric forecast skill on the 10–15 day timescale. Various mechanisms have been proposed to explain the stratospheric modulation of tropospheric climate on the week to seasonal timescale but there is no consensus yet as to which is the most important one (Charlton et al. 2005a; Tripathi et al. 2014).

On shorter (medium range weather forecasting) time scales, middle atmosphere dynamics are still useful for understanding forecast improvements. When the Canadian Meteorological Centre raised the lid of its operational forecast model from 10 to 0.1 hPa, most (over 80 %) of the improvement in forecast skill (of both stratosphere and troposphere) was achieved without new measurements in the upper stratosphere (AMSU-A ch. 11–14 and GPS RO between 30–40 km) (Charron et al. 2012). This means that an improved modeling of the stratosphere is sufficient to

obtain improved upper stratospheric analysis (where no new data were assimilated). Moreover, the improvement was greatest in the winter for both hemispheres. Thus improvement depended more on season (when the stratosphere was dynamically active) than on hemisphere (or observation distribution). Furthermore, when additional observations in the upper stratosphere were assimilated, they were beneficial in winter but not in summer. These results are understandable in the context of middle atmosphere dynamics. Just as tropospheric observations are most useful when dynamic activity (such as baroclinic wave development) is occurring, stratospheric observations are most beneficial when the stratosphere is dynamically active (in winter).

Because of the prevalence of gravity waves and divergent motions in the mesosphere (Koshyk et al. 1999), and the sparse observational coverage, the assimilation of mesospheric observations brings further challenges. By comparing forecasts started from analyses in the middle atmosphere (100–0.1 hPa) with those started from climatology in the middle atmosphere, Hoppel et al. (2008) found that the assimilation of middle atmospheric observations were indeed beneficial for winter high latitudes up to the 10-day forecast lead time. In the summer, when the stratosphere is quiescent and dominated by zonal mean flow, persistence or climatology works reasonably well so the benefit of assimilation is not as apparent. On the other hand, mesospheric observations also help to improve the depiction and forecasts of certain planetary waves in the mesosphere as well as reducing biases in zonal mean fields stemming from model errors (Hoppel et al. 2013).

3.3 *Model Error*

3.3.1 **Bias Estimation**

Since not all the resolved waves will be correctly analysed because the observing system can detect only certain spatial scales, and some of the resolved waves are forced in the models by parameterization schemes which are imperfect (e.g. deep convection), and the launch spectrum in GWD schemes is largely unknown, we should expect errors in the meridional circulation. Errors in the forcing of a meridional circulation should then lead to latitudinally varying biases. Thus, we should expect bias in zonal mean fields in stratospheric forecasts. Observations (such as those from nadir sounders) also have biases and require a pre-assimilation bias-correction procedure, so the challenge is to separate these two sources of biases. Moreover, observation bias correction schemes often rely on an assumption of unbiased forecasts—which is clearly invalid in the stratosphere. Dee and Uppala (2009) noted that improvement in the stratospheric bias of ERA-interim over ERA-40 was achieved through the introduction of variational bias correction (Derber and Wu 1998). In this procedure, bias correction parameters are added to the control vector so that all observations—including those which are not corrected, such as radiosondes—are used to determine their values. This then forces a

consistency among observations which are being bias corrected (e.g. the same instrument on different platforms). Of course, even with variational bias correction, the bias so-determined could be due to either a bias in observations or observation operators or to a bias in the model forecast. Since the bias correction is applied to the observation, only the former type of bias is desired. Thus care must be taken to ensure that the recovered bias is truly due to the observations. To some extent, the anchoring of the assimilation system by uncorrected observations (such as radiosondes) reduces the likelihood that model bias will be detected. However, in the upper stratosphere and mesosphere where few uncorrected observations exist, the danger of correcting for model bias is considerable. Thus Dee and Uppala (2009) chose to leave the top peaking channel (SSU channel 3 or AMSU-A channel 14) uncorrected in the ERA-interim, in order to anchor the system. This resulted in a reduced warm bias near the model top. Since a warm bias had independently been attributed to the model forecast (McNally 2004) the results were positive. So although variational bias correction has proven to be a valuable tool for reanalyses as well as operational assimilation systems, the problem of separating model and measurement bias in the upper stratosphere and the mesosphere remains (Hoppel et al. 2013). Leaving a certain instrument uncorrected still creates difficulty when it is present on multiple platforms, or when the observing system changes (e.g. when the top peaking channel changed from SSU ch. 3 to AMSU-A ch. 14). Furthermore, whatever bias exists in the uncorrected measurement will appear in the analyses. Several distinct temporal inhomogeneities in global-mean ERA-Interim temperatures were identified by McLandress et al. (2014).

3.3.2 Unresolved Gravity Wave Drag Estimation

The breaking of small-scale gravity waves in the upper stratosphere and mesosphere plays an important role in driving the meridional circulation and thus the impact of these small-scale waves can be detected in large-scale observations such as nadir or AMSU-A satellite measurements. Therefore, the systematic biases found in the model compared to observations may be associated to a large extent with those small-scale gravity wave breaking processes that are not resolved in the model but have a large global impact. Data assimilation techniques which are used to produce analyses can also be used to help diagnose this systematic model error. For example, McLandress et al. (2012) used the time averaged zonal mean zonal wind analysis increments to identify missing gravity wave drag in the southern hemisphere, while Pulido (2014) employed systematic differences of potential vorticity between analyses and forecasts to determine momentum forcing via potential vorticity inversion. In addition, data assimilation techniques can also be used to estimate the missing momentum forcing in the model as a product of the assimilation. Since this missing forcing may be associated to a large extent with gravity wave drag due to unresolved waves, this information can then be used to constrain gravity wave drag parameterization schemes. For example, Pulido and Thuburn (2005) applied a 4D variational assimilation (4D-Var) technique to

estimate the missing momentum forcing in the model instead of estimating the model state. The optimal momentum forcing is the one whose model state evolution is associated with the minimum of the cost function.

One helpful aspect of middle atmosphere data assimilation is that the only two processes that are parameterized in models at that height range are radiation and the dissipation of small-scale gravity waves. Since the physics of radiation is well known and only has a large impact on long (seasonal) time scales, the missing zonal momentum forcing at these heights on shorter time scales may be mainly attributed to small-scale gravity waves (Pulido and Thuburn 2006). The optimal momentum forcing estimated with 4D-Var resembles that expected from the filtering of an isotropic gravity wave spectrum (Lindzen 1981), with large deceleration centres at high latitudes during the winter and summer (Pulido and Thuburn 2008). On the other hand, it is less evident that the estimated forcing at high latitudes during equinox is associated with an isotropic gravity wave spectrum.

3.3.3 Parameter Estimation

Models have a large number of parameters that are not directly observable. Currently, climate modelers infer the values of such unknown parameters manually by comparing the climatology of model integrations with the observed climatology. These inferred parameters may then change if resolution, parameterizations or other parameter values are changed in the model. Data assimilation provides an objective approach to the estimation of unknown model parameters (Ruiz et al. 2013). Online parameter estimation techniques based on the ensemble Kalman filter or 4DVar usually define an augmented state which is composed of the model state and also the parameters to be estimated. However, the parameters are not directly constrained by observations as is the model state. Instead, the parameters are constrained through background error correlations between the parameters and model state variables.

Gravity wave drag parameters related to the launch wave spectrum and to the saturation and breaking properties of the waves (which determine the gravity wave drag vertical profile) are poorly known from observations. Various techniques have been employed to estimate reasonable values for such parameters. Watanabe (2008) used the results of a high resolution global simulation to determine the characteristics of wave momentum fluxes and to estimate the launch wave spectrum parameters in the Hines parameterization scheme. Some efforts have also been devoted to using high resolution satellite observations (e.g. Atmospheric InfraRed Sounder or AIRS, High Resolution Dynamics Limb Sounder or HIRDLS) to constrain the launch gravity wave spectrum (Alexander et al. 2010). In addition, inverse techniques based on data assimilation have been used to estimate gravity wave parameters. Pulido et al. (2012) proposed an offline data assimilation technique based on a genetic algorithm that uses the estimated missing momentum forcing to constrain launch momentum flux and saturation parameters. Tandeo et al. (2015) proposed an ensemble Kalman filter (EnKF) coupled to an expectation

maximization algorithm to estimate parameters from an orographic gravity wave drag scheme and also to estimate initial background error covariances which are essential for convergence of the filter in a perfect model experiment. They show that in the presence of model error the filter may converge to different optimal parameters depending on the choice of the first guess value. In general, the estimation of parameters using data assimilation techniques faces a number of unique challenges which require further development or refinement. These challenges are associated with the highly nonlinear nature of the model state response to parameters changes in the context of current assimilation techniques (EnKF and 4DVar) which are based on the linear-Gaussian assumption. To deal with this limitation, some potential solutions have been proposed such as the combination of an offline genetic algorithm with a 4DVar technique (Pulido et al. 2012), the use of a hybrid EnKF-particle filter, with EnKF for the state variables and a particle filter applied to the parameters (Santitissadeekorn and Jones 2015) and the afore-mentioned EnKF combined with an Expectation-Maximization algorithm (Tandeo et al. 2015). A second major challenge is parameter estimation in the presence of model error. In this regard, a recent study found a positive impact particularly when parameter estimation was combined with model error treatment approaches (Ruiz and Pulido 2015). A third challenge is the interaction between different parameterizations. While this issue is particularly important for tropospheric data assimilation (e.g. interactions between a planetary boundary layer scheme and a convective scheme), it is not as relevant for middle atmosphere data assimilation, as noted earlier.

4 Discussion

While in Sect. 3, we have already noted some challenges in middle atmosphere data assimilation in the context of the topics discussed earlier, in this section we focus on issues that have not yet been raised or fully considered.

A characteristic of the observing system for the middle atmosphere is the difficulty of obtaining in situ measurements apart from radiosondes and aircraft observations. Thus, as noted earlier, the vertical structure above the lower stratosphere is not well observed from operational satellites since nadir sounders are sensitive to temperatures over thick layers. On the other hand, GPS radio occultation measurements have been very beneficial not only for their information content and vertical resolution (e.g. Cardinali and Healy 2014; Healy and Thépaut 2006) but also for their very low bias which allows them to serve as anchors in bias correction schemes for satellite observations (Cucurull et al. 2014). The constraint of GPS RO data on analyses, however, diminishes as the data become noisier in the upper stratosphere, even as their effectiveness for anchoring the bias correction of satellite data extends throughout the stratosphere (e.g. Cucurull et al. 2014, Charron et al. 2012). Limb sounders such as MIPAS (Michelson Interferometer for Passive Atmospheric Sounding) on Envisat (Environmental satellite), Atmospheric Chemistry Experiment (ACE) aboard SciSat, Microwave Limb Sounder (MLS) on

EOS AURA and Sounding of the Atmosphere Using Broadband Emission Radiometry (SABER) aboard TIMED also provide (or did provide) useful information on vertical structure, but these instruments have been on research not operational satellites. As a result, these observations are useful for reanalyses and non-operational assimilation systems but not in operational systems unless they are available in real time (as in the case of MIPAS). It may require the research data assimilation community to make a case that such observations are important for middle atmosphere data assimilation in order for them to be considered for operational delivery. Moreover, concern has been expressed over the lack of plans for new limb sounders in the future (Errera et al. 2015). In the tropics where simple dynamical balances between mass and wind fields are lacking, wind observations are vital but sparse. The Atmospheric Dynamics Mission (ADM-Aeolus) will measure winds globally up to 30 km using an active sensor which will help to better constrain the tropical lower stratosphere but in the mesosphere where unbalanced motions are important, the absence of wind measurements remains an issue. Furthermore, ADM-Aeolus will only measure line-of-sight winds, not vector winds, so a reasonable first guess is required in order to use the information. Apart from ADM-Aeolus, there is a potential for constituent assimilation to improve wind analyses in the middle atmosphere (e.g. Allen et al. 2014, 2015; Milewski and Bourqui 2011, 2013; Semane et al. 2009). Particularly noteworthy is the potential for improving tropical wind analyses using combined ozone and height measurements in an ensemble Kalman filter (Allen et al. 2015). The only real time observations of the mesosphere are from the SSMIS instrument, however Hoppel et al. (2013) note that apart from the F19 and F20 deployments of the Defense Meteorological Satellite Program (DMSP), there are no other plans for upper atmospheric sounding channels on other satellite sensors raising the possibility of an unconstrained mesosphere analysis in the future. Given that the mesosphere provides the upper boundary for the stratosphere, negative impact on the quality of meteorological forecasts from a paucity of mesospheric measurements is plausible.

The usual filtering of or the imposition of balance constraints on the initial state in the data assimilation cycle (Daley 1991) to eliminate gravity waves should be avoided to keep relevant information for the middle atmosphere. On the other hand, most data assimilation techniques which are intermittent (e.g. EnKF, 3DVar) produce temporal discontinuities in the state variables when observations are assimilated, therefore the generation of spurious gravity waves is inevitable with these techniques (e.g. Sankey et al. 2007; Allen et al. 2015). A promising way to avoid this issue is the use of an incremental analysis update (IAU) approach (Bloom et al., 1996) or nudging techniques (Lei et al. 2012). The idea behind these approaches is to distribute the forcing toward observations along the whole assimilation window, in the first case through a uniform forcing term (i.e. the analysis increment), or in the second case through a linear damping forcing term (nudging term) that pushes the model toward the analyses. These approaches give a smooth evolution of the model state so that they avoid the spurious generation of gravity waves due to spin-up processes and in turn avoid the need to apply external filters. The smooth model state evolution then permits cleaner momentum budget studies. The use of

IAU was also found helpful for capturing mesospheric tides (Sankey et al. 2007; Wang et al. 2011). Since the migrating diurnal tide is already captured by general circulation models, its signal in the mesosphere can also be captured even without mesospheric observations, if care is taken when filtering analysis increments. The incremental analysis update approach was employed in combination with 3D variational assimilation to produce Modern-Era Retrospective Analysis for Research and Applications (MERRA) reanalyses by the NASA-GEOS data assimilation system (Rienecker et al. 2011). A 4D extension of the IAU procedure was developed by Lorenc et al. (2015) and is used in the hybrid ensemble variational assimilation scheme which is used for deterministic medium range weather forecasts at Environment Canada (Buehner et al. 2015).

Section 3.1.1 highlighted the fact that information propagates vertically with resolved and unresolved waves, but it is worth remarking that by the same mechanisms, errors can also propagate vertically. For example, Nezlin et al. (2009) show that increasing the observation error applied to simulated tropospheric observations deteriorates the quality of stratospheric and mesospheric analyses. Thus the tuning of assimilation schemes for tropospheric forecast quality may have unintended impacts on the analyses of the middle atmosphere (e.g. Sankey et al. 2007). On the other hand, the upscale propagation of errors and concomitant loss of predictability characteristic of the troposphere is not as severe in the middle atmosphere (Ngan and Eperon 2012). Moreover, increased resolution and better resolved gravity waves may actually help to improve predictability on longer time scales. This may sound counter intuitive given the flat kinetic energy spectrum in the mesosphere due to large amplitude gravity waves (Koshyk et al. 1999), but the hypothesis of Ngan and Eperon (2012) is that predictability can be increased if the gravity waves are better resolved because the upscale error cascade is slower for these waves than for balanced modes. Whether the results of this theoretical study are borne out in operational data assimilation systems is yet to be determined.

The impact of the stratosphere on tropospheric forecasts has primarily been associated with certain “extreme” dynamical events in the stratosphere such as sudden warmings. However, there remain many questions associated with the stratospheric influence on tropospheric forecasts, such as whether the influence extends beyond such extreme events and how far in advance extreme events can be predicted. These are some of the questions that are being addressed by a new international collaboration within SPARC called the Stratospheric Network on Assessment of Predictability (SNAP, <http://www.sparc-climate.org/activities/assessing-predictability/>). Operational weather centers have seen improvements in tropospheric forecasts when raising their model lids (e.g. Charron et al. 2012), however such changes are made simultaneously with other assimilation system changes. Thus another goal of SNAP is to try to isolate the extent to which accurate stratospheric forecasts contribute to tropospheric predictability.

5 Summary

Information can be propagated vertically in data assimilation systems through covariances, vertically propagating waves, and gravity wave drag schemes. As a result, very large scales in the mesosphere can be improved even without assimilating any mesospheric measurements. The fact that the middle atmosphere is driven by vertically propagating waves has important implications for data assimilation systems. (1) Tropospheric waves (whether correctly simulated or not) impact zonal mean fields in the stratosphere and mesosphere. This means that apparently random signals (e.g. waves) can produce nonlocal systematic errors (e.g. a zonal mean bias). (2) Since not all waves are correctly simulated, and the large wavenumber part of the spectrum is not resolved, we should expect a model bias (errors in the zonal mean) in the mesosphere and stratosphere. This has implications for observation bias correction schemes that assume the background forecast is unbiased. (3) Mesospheric analyses are sensitive to errors in tropospheric analyses. On the other hand, perhaps we can use this sensitivity to help choose assimilation parameters in the troposphere. (4) Information propagates up (through resolved waves during the forecast step). Some of the large scales in the mesosphere can be improved even with no mesospheric observations if tropospheric wave forcing is captured and the middle atmosphere is well modelled. The assimilation of mesospheric observations will additionally improve mesospheric analyses on large scales thus providing a better upper boundary condition with which to constrain forecasts of the troposphere and lower stratosphere, particularly on longer time scales.

Given the fact that the middle atmosphere is largely driven by vertically propagating waves including gravity waves, and the fact that global climate models often have coarse resolution, it is necessary to parameterize the impact of the dissipation of subgrid scale waves on the zonal mean flow thus introducing a potential source of model error. Data assimilation is useful for estimating the missing drag attributed to such waves as well as for estimating parameters involved in gravity wave drag schemes.

Acknowledgements We are grateful to Ted Shepherd, Josep Aparicio, Martin Charron and Karl Hoppel for providing helpful comments on an earlier version of this article. We are also grateful to the OSIRIS team and the Canadian Space Agency for providing the figure depicting the ozone distribution for March 2004 from OSIRIS. The work involving the CMAM-DAS was supported by the C-SPARC project which was funded by the Canadian Foundation for Climate and Atmospheric Sciences (CFCAS) and the Canadian Space Agency (CSA). MP's work was partially funded by a CONICET Grant PIP 112-20120100414CO.

References

- Alexander MJ, Geller M, McLandress C, Polavarapu S, Preusse P, Sassi F, Sato K, Eckermann S, Ern M, Hertzog A, Kawatani Y, Pulido M, Shaw T, Sigmund M, Vincent R, Watanabe S (2010) Recent developments in gravity wave effects in climate models, and the global

- distribution of gravity wave momentum flux from observations and models. *Q J Roy Meteorol Soc* 136:1103–1124
- Allen DR, Hoppel KW, Kuhl DD (2014) Wind extraction potential from 4D-Var assimilation of stratospheric O₃, N₂O, and H₂O using a global shallow water model. *Atmos Chem Phys* 14:3347–3360. doi:[10.5194/acp-14-3347-2014](https://doi.org/10.5194/acp-14-3347-2014)
- Allen DR, Hoppel KW, Kuhl DD (2015) Wind extraction potential from ensemble Kalman filter assimilation of stratospheric ozone using a global shallow water model. *Atmos Chem Phys* 15:5835–5850. doi:[10.5194/acp-15-5835-2015](https://doi.org/10.5194/acp-15-5835-2015)
- Andrews DG, Holton JR, Leovy CB (1987) Middle atmosphere dynamics. Academic Press, p 489
- Austin J, Shindell D, Beagley SR, Bruhl C, Dameris M, Manzini E, Nagashima T, Newman P, Pawson S, Pitari G, Rosanov E, Schnadt C, Shepherd TG (2003) Uncertainties and assessments of chemistry-climate models of the stratosphere. *Atmos Chem Phys* 3:1–27
- Baldwin MP, Dunkerton TJ (2001) Stratospheric Harbingers of anomalous weather regimes. *Science* 294:581–584. doi:[10.1126/science.1063315](https://doi.org/10.1126/science.1063315)
- Bloom SC, Takaes LL, Da Silva AM, Ledvina D (1996) Data assimilation using incremental analysis updates. *Mon Weather Rev* 124:1256–1271
- Bocquet M, Elbern H, Eskes H, Hirtl M, Žabkar R, Carmichael GR, Flemming J, Inness A, Pagowski M, Pérez Camaño JL, Saide, PE, San Jose R, Sofiev M, Vira J, Baklanov A, Carnevale C, Grell G, Seigneur C (2015) Data assimilation in atmospheric chemistry models: current status and future prospects for coupled chemistry meteorology models. *Atmos Chem Phys* 15:5325–5358. doi:[10.5194/acp-15-5325-2015](https://doi.org/10.5194/acp-15-5325-2015)
- Boer GJ, Hamilton K (2008) QBO influence on extratropical predictive skill. *Clim Dyn* 31: 987–1000
- Buehner M, McTaggart-Cowan R, Beaulne A, Charette C, Garand L, Heilliette S, Lapalme E, Laroche S, Macpherson SR, Morneau J, Zadra A (2015) Implementation of deterministic weather forecasting systems based on ensemble-variational data assimilation at Environment Canada. Part I: the global system. *Mon Weather Rev* 143:2532–2559
- Campbell WF, Bishop CH, Hodyss D (2010) Vertical covariance localization for satellite radiances in Ensemble Kalman Filters. *Mon Weather Rev* 138:282–290. doi:[10.1175/2009MWR3017.1](https://doi.org/10.1175/2009MWR3017.1)
- Cardinali C, Healy S (2014) Impact of GPS radio occultation measurements in the ECMWF system using adjoint-based diagnostics. *Q J R Meteorol Soc* 140:2315–2320. doi:[10.1002/qj.2300](https://doi.org/10.1002/qj.2300)
- Charlton AJ, O'Neill A, Lahoz WA, Massacand AC (2004) Sensitivity of tropospheric forecasts to stratospheric initial conditions. *Q J R Meteorol Soc* 130:1771–1792
- Charlton AJ, O'Neill A, Berrisford P, Lahoz WA (2005a) Can the dynamical impact of the stratosphere on the troposphere be described by large-scale adjustment to the stratospheric PV distribution? *Q J R Meteorol Soc* 131:525–543
- Charlton AJ, O'Neill A, Lahoz WA, Massacand AC, Berrisford P (2005b) The impact of the stratosphere on the troposphere during the southern hemisphere stratospheric sudden warming, September 2002. *Q J R Meteorol Soc* 131:2171–2188
- Charron M, Polavarapu S, Buehner M, Vollandcourt PA, Charrette C, Roch M, Morneau J, Garand L, Aparicio JM, MacPherson S, Pellerin S, St-James J, Heilliette S (2012) The stratospheric extension of the Canadian operational deterministic medium range weather forecasting system and its impact on tropospheric forecasts. *Mon Weather Rev* 140:1924–1944. doi:[10.1175/MWR-D-11-00097.1](https://doi.org/10.1175/MWR-D-11-00097.1)
- Cucurull L, Anthes RA, Tsao L-L (2014) Radio occultation observations as anchor observations in numerical weather prediction models and associated reduction of bias corrections in microwave and infrared satellite observations. *J. Atmos. Oceanic Technol.* 31:20–32. doi:[10.1175/JTECH-D-13-00059.1](https://doi.org/10.1175/JTECH-D-13-00059.1)
- Daley R (1991) Atmospheric data analysis. Cambridge University Press, p 457
- Dee DP et al (2011) The ERA-Interim reanalysis: configuration and performance of the data assimilation system. *Q J R Meteorol Soc* 137:553–597. doi:[10.1002/qj.828](https://doi.org/10.1002/qj.828)

- Dee DP, Uppala S (2009) Variational bias correction of satellite radiance data in the ERA-Interim reanalysis. *Q J R Meteorol Soc* 135:1830–1841
- Derber JC, Wu W-S (1998) The use of TOVS cloud-cleared radiances in the NCEP SSI analysis system. *Mon Weather Rev* 126:2287–2299
- Douville H (2009) Stratospheric polar vortex influence on Northern Hemisphere winter climate variability. *Geophys Res Lett* 36:L18703. doi:[10.1029/2009GL039334](https://doi.org/10.1029/2009GL039334)
- Errera Q, Fujiwara M, Long C, Jackson D (2015) Report from the 10th SPARC data assimilation workshop and the 2014 SPARC reanalysis intercomparison project (S-RIP) workshop in Washington DC, USA. *SPARC Newsletter* 44. <http://www.sparc-climate.org/publications/newsletter/>
- Fox-Rabinovitz MS, Lindzen RS (1993) Numerical experiments on consistent horizontal and vertical resolution for atmospheric models and observing systems. *Mon Weather Rev* 121:264–271
- Gerber EP, Butler A, Calvo N, Charlton-Perez A, Giorgetta M, Manzini E, Perlitz J, Polvani LM, Sassi F, Scaife AA, Shaw TA, Son S-W, Watanabe S (2012) Assessing and Understanding the Impact of Stratospheric Dynamics and Variability on the Earth System. *Bull Am Meteorol Soc* 93:845–859. doi:[10.1175/BAMS-D-11-00145.1](https://doi.org/10.1175/BAMS-D-11-00145.1)
- Healy SB, Thépaut J-N (2006) Assimilation experiments with CHAMP GPS radio occultation measurements. *Q J R Meteorol Soc* 132:605–623. doi:[10.1256/qj.04.182](https://doi.org/10.1256/qj.04.182)
- Hegglin MI, Plummer DA, Shepherd TG, Scinocca JF, Anderson J, Froidevaux L, Funke B, Hurst D, Rozanov A, Urban J, von Claramann T, Walker KA, Wang HJ, Tegtmeier S, Weigel K (2014) Vertical structure of stratospheric water vapour trends derived from merged satellite data. *Nat Geosci* 7:768–776. doi:[10.1038/ngeo2236](https://doi.org/10.1038/ngeo2236)
- Hines CO (1997) Doppler spread parametrization of gravity-wave momentum deposition in the middle atmosphere. Part 1: basic formulation. *J Atmos Sol Terr Phys* 59:371–386
- Hoppel KW, Baker NL, Coy L, Eckermann SD, McCormack JP, Nedoluha GE, Siskind DE (2008) Assimilation of stratospheric and mesospheric temperatures from MLS and SABER into a global NWP model. *Atmos Chem Phys* 8:6103–6116. doi:[10.5194/acp-8-6103-2008](https://doi.org/10.5194/acp-8-6103-2008)
- Hoppel KW, Eckermann SD, Coy L, Nedoluha GE, Allen DR, Swadley SD, Baker NL (2013) Evaluation of SSIMIS upper atmosphere sounding channels for high-altitude data assimilation. *Mon Weather Rev* 141:3314–3330. doi:[10.1175/MWR-D-13-00003.1](https://doi.org/10.1175/MWR-D-13-00003.1)
- Iga S, Tomita H, Satoh M, Goto K (2007) Mountain-wave-like spurious waves associated with simulated cold fronts due to inconsistencies between horizontal and vertical resolutions. *Mon Weather Rev* 135:2629–2641
- Kobayashi S, Ota Y, Harada Y, Ebita A, Moriya M, Onoda H, Onogi K, Kamahori H, Kobayashi C, Endo H, Miyaoka K, Takahashi K (2015) The JRA-55 reanalysis: general specifications and basic characteristics. *J Meteorol Soc Jpn* 93(1):5–48. doi:[10.2151/jmsj.2015-001](https://doi.org/10.2151/jmsj.2015-001)
- Koshyk JN, Boville BA, Hamilton K, Manzini E, Shibata K (1999) Kinetic energy spectrum of horizontal motions in middle-atmosphere models. *J Geophys Res* 104:27177–27190
- Leovy C (1964) Simple models of thermally driven mesospheric circulation. *J Atmos Sci* 21:327–341
- Lei L, Stauffer DR, Deng A (2012) A hybrid nudging-ensemble Kalman filter approach to data assimilation. Part II: Application in a shallow-water model. *Tellus A* 64 (1):18485
- Li F, Austin J, Wilson J (2008) The strength of the Brewer-Dobson circulation in a changing climate: coupled chemistry–climate model simulations. *J Climate* 21:40–57
- Lindzen RS (1981) Turbulence and stress owing to gravity wave and tidal breakdown. *J Geophys Res* 86:9707–9714
- Lorenc AC, Bowler NE, Clayton AM, Pring SR, Fairbairn D (2015) Comparison of hybrid-4DEnvar and hybrid-4DVar data assimilation methods for global NWP. *Mon Weather Rev* 143:212–229. doi:[10.1175/MWR-D-14-00195.1](https://doi.org/10.1175/MWR-D-14-00195.1)
- Lübken F, Rapp M, Hoffmann P (2002) Neutral air turbulence and temperatures in the vicinity of polar mesosphere summer echoes. *J Geophys Res* 107(D15):4273. doi:[10.1029/2001JD000915](https://doi.org/10.1029/2001JD000915)

- Marshall AG, Scaife AA (2009) Impact of the QBO on surface winter climate. *J Geophys Res* 114: D18110. doi:[10.1029/2009JD011737](https://doi.org/10.1029/2009JD011737)
- McLandress C (1998) On the importance of gravity waves in the middle atmosphere and their parameterization in general circulation models. *J Atmos Sol Terr Phys* 60:1357–1383
- McLandress C, Shepherd TG (2009) Simulated anthropogenic changes in the brewer-dobson circulation, including its extension to high latitudes. *J Climate* 22:1516–1540. doi:[10.1175/2008JCLI2679.1](https://doi.org/10.1175/2008JCLI2679.1)
- McLandress C, Shepherd TG, Polavarapu S, Beagley SR (2012) Is missing orographic gravity wave drag near 60 S the cause of the stratospheric zonal wind biases in chemistry-climate models? *J Atmos Sci* 69:802–818
- McLandress C, Scinocca JF, Shepherd TG, Reader MC, Manney GL (2013) Dynamical Control of the mesosphere by orographic and nonorographic gravity wave drag during the extended northern winters of 2006 and 2009. *J Atmos Sci* 70:2152–2169. doi:[10.1175/JAS-D-12-0297.1](https://doi.org/10.1175/JAS-D-12-0297.1)
- McLandress C, Plummer DA, Shepherd TG (2014) Technical Note: A simple procedure for removing temporal discontinuities in ERA-Interim upper stratospheric temperatures for use in nudged chemistry-climate model simulations. *Atmos Chem Phys* 14:1547–1555. doi:[10.5194/acp-14-1547-2014](https://doi.org/10.5194/acp-14-1547-2014)
- McNally T (2004) The assimilation of stratospheric satellite data at ECMWF. In: Proceedings of the ECMWF/SPARC workshop on modelling and assimilation for the stratosphere and tropopause. <http://www.ecmwf.int/publications/library/do/references/list/17123>. Accessed 23–26 June 2003
- Milewski T, Bourqui MS (2011) Assimilation of stratospheric temperature and ozone with an ensemble kalman filter in a chemistry-climate model. *Mon Weather Rev* 139:3389–3404. doi:[10.1175/2011MWR3540.1](https://doi.org/10.1175/2011MWR3540.1)
- Milewski T, Bourqui MS (2013) Potential of an ensemble Kalman smoother for stratospheric chemical-dynamical data assimilation. *Tellus A* 65:18541. doi:[10.3402/tellusa.v65i0.18541](https://doi.org/10.3402/tellusa.v65i0.18541)
- Monge-Sanz BM, Chipperfield MP, Dee DP, Simmons AJ, Uppala SM (2013) Improvements in the stratospheric transport achieved by a chemistry transport model with ECMWF (re)analyses: identifying effects and remaining challenges. *Q J R Meteorol Soc* 139:654–673
- Nezlin Y, Rochon YJ, Polavarapu S (2009) Impact of tropospheric and stratospheric data assimilation on mesospheric prediction. *Tellus* 61A(1):154–159
- Ngan K, Eperon GE (2012) Middle atmosphere predictability in a numerical weather prediction model: Revisiting the inverse error cascade. *Q J R Meteorol Soc* 138:1366–1378. doi:[10.1002/qj.984](https://doi.org/10.1002/qj.984)
- Orr A, Bechtold P, Scinocca J, Ern M, Janiskova M (2010) Improved middle atmosphere climate and forecasts in the ECMWF model through a nonorographic gravity wave drag parameterization. *J Climate* 23:5905–5926. doi:[10.1175/2010JCLI3490.1](https://doi.org/10.1175/2010JCLI3490.1)
- Polavarapu SM, Shepherd TG, Ren S, Rochon YJ (2005) Some challenges of middle atmosphere data assimilation. *Q J R Meteorol Soc* 131:3513–3527
- Pulido M, Thuburn J (2005) Gravity wave drag estimation from global analyses using variational data assimilation principles. I: Theory and implementation. *Q J R Meteorol Soc* 131:1821–1840
- Pulido M, Thuburn J (2006) Gravity wave drag estimation from global analyses using variational data assimilation principles. II: A case study. *Q J R Meteorol Soc* 132:1527–1543
- Pulido M, Thuburn J (2008) The seasonal cycle of gravity wave drag in the middle atmosphere. *J Climate* 21:4664–4679
- Pulido M, Polavarapu S, Shepherd T, Thuburn J (2012) Estimation of optimal gravity wave parameters for climate models using data assimilation. *Q. J. Roy. Meteorol. Soc.* 138:298–309. doi:[10.1002/qj.932](https://doi.org/10.1002/qj.932)
- Pulido M (2014) A simple technique to infer the missing gravity wave drag in the middle atmosphere using a general circulation model: potential vorticity budget. *J Atmos Sci* 71: 683–696

- Ren S, Polavarapu S, Shepherd TG (2008) Vertical propagation of information in a middle atmosphere data assimilation system by gravity-wave drag feedbacks. *Geophys Res Lett* 35(6): L06804. doi:[10.1029/2007GL032699](https://doi.org/10.1029/2007GL032699)
- Ren S, Polavarapu S, Beagley SR, Nezlin Y, Rochon YJ (2011) The impact of gravity wave drag on mesospheric analyses of the 2006 stratospheric major warming. *J Geophys Res* 116: D19116. doi:[10.1029/2011JD015943](https://doi.org/10.1029/2011JD015943)
- Rienecker MM, Suarez MJ, Gelaro R, Todling R, Bacmeister J, Liu E et al (2011) MERRA: NASA's modern-era retrospective analysis for research and applications. *J. Climate* 24:3624–3648
- Ruiz J, Pulido M, Miyoshi T (2013) Estimating parameters with ensemble-based data assimilation. A Review. *J Meteorol Soc Jpn* 91:79–99
- Ruiz J, Pulido M (2015) Parameter Estimation Using Ensemble Based Data Assimilation in the Presence of Model Error. *Mon Weather Rev* 143:1568–1582. doi:[10.1175/MWR-D-14-00017.1](https://doi.org/10.1175/MWR-D-14-00017.1)
- Sandu A, Chai T (2011) Chemical data assimilation—an overview. *Atmosphere* 2:426–463. doi:[10.3390/atmos2030426](https://doi.org/10.3390/atmos2030426)
- Sankey D, Ren S, Polavarapu S, Rochon YJ, Nezlin Y, Beagley S (2007) Impact of data assimilation filtering methods on the mesosphere. *J Geophys Res* D112(24):D24104. doi:[10.1029/2007JD008885](https://doi.org/10.1029/2007JD008885)
- Santitissadeekorn N, Jones C (2015) Two-stage filtering for joint state-parameter estimation. *Mon Weather Rev* 143:2028–2042
- Scinocca JF (2003) An accurate spectral non-orographic gravity wave drag parameterization for general circulation models. *J Atmos Sci* 60:667–682
- Scinocca JF, McFarlane NA, Lazare M, Li J (2008) The CCCma 3rd generation AGCM and its extension into the middle atmosphere. *Atmos Chem Phys* 8:7055–7074
- Semane N, Peuch V-H, Pradier S, Desroziers G, El Amraoui L, Brousseau P, Massart S, Chapnik B, Peuch A (2009) On the extraction of wind information from the assimilation of ozone profiles in Météo-France 4-D-Var operational NWP suite. *Atmos Chem Phys* 9: 4855–4867. doi:[10.5194/acp-9-4855-2009](https://doi.org/10.5194/acp-9-4855-2009)
- Shaw TA, Shepherd TG (2008) Raising the roof. *Nat Geosci* 1:12–13
- Shepherd TG (2000) The middle atmosphere. *J Atmos Sol Terr Phys* 62:1587–1601
- Shepherd TG (2002) Issues in stratosphere-troposphere coupling. *J Meteorol Soc Jpn* 80B:769–792
- Shepherd TG (2007) Transport in the middle atmosphere. *J Meteorol Soc Jpn* 85B:165–191
- Shepherd TG, Semeniuk K, Koshyk JN (1996) Sponge layer feedbacks in middle-atmosphere models. *J Geophys Res* 101:23, 447–23, 464
- Shepherd TG, Plummer DA, Scinocca JF, Hegglin MI, Fioletov VE, Reader MC, Remsberg E, von Clarmann T, Wang HJ (2014) Reconciliation of halogen-induced ozone loss with the total-column ozone record. *Nat Geosci* 7:443–449. doi:[10.1038/NGEO2155](https://doi.org/10.1038/NGEO2155)
- Sigmond M, Scinocca J, Kharin VV, Shepherd TG (2013) Enhanced seasonal forecast skill following stratospheric sudden warmings. *Nat Geosci* 6:98–102. doi:[10.1038/geo1698](https://doi.org/10.1038/geo1698)
- Smith A (2004) Physics and chemistry of the mesopause region. *J Atmos Solar Terr Phys* 66: 839–857
- Stockdale TN, Molteni F, Ferranti L (2015) Atmospheric initial conditions and the predictability of the Arctic Oscillation. *Geophys Res Lett* 42:1173–1179. doi:[10.1002/2014GL062681](https://doi.org/10.1002/2014GL062681)
- Tandee P, Pulido M, Lott F (2015) Offline parameter estimation using EnKF and maximum likelihood error covariance estimates: Application to a subgrid-scale orography parametrization. *Q J R Meteorol Soc* 141:383–395. doi:[10.1002/qj.2357](https://doi.org/10.1002/qj.2357)
- Thompson DWJ, Wallace JM (2001) Regional climate impacts of the Northern Hemisphere Annual Model. *Science* 293:85. doi:[10.1126/science.1058958](https://doi.org/10.1126/science.1058958)
- Tripathi OP, Baldwin M, Charlton-Perez A, Charron M, Eckermann SD, Gerber E, Harrison RG, Jackson DR, Kim B-M, Kuroda Y, Lang A, Mahmood S, Mizuta R, Roff G, Sigmond M, Son S-W (2014) The predictability of the extratropical stratosphere on monthly time-scales and its impact on the skill of tropospheric forecasts. *R Meteorol Soc, Q.J.* doi:[10.1002/qj.2432](https://doi.org/10.1002/qj.2432)
- Vallis GK (2006) Atmospheric and oceanic fluid dynamics. Cambridge University Press, p 745

- Wang H, Fuller-Rowell TJ, Akmaev RA, Hu M, Kleist DT, Iredell MD (2011) First simulations with a whole atmosphere data assimilation and forecast system: the January 2009 major sudden stratospheric warming. *J Geophys Res* 116:A12321. doi:[10.1029/2011JA017081](https://doi.org/10.1029/2011JA017081)
- Watanabe S (2008) Constraints on a non-orographic gravity wave drag parameterization using a gravity wave resolving general circulation model. *SOLA* 4:61–64
- Watanabe S, Kawatani Y, Tomikawa Y, Miyazaki K, Takahashi M, Sato K (2008) General aspects of a T213L256 middle atmosphere general circulation model. *J Geophys Res* 113:D12110. doi:[10.1029/2008JD010026](https://doi.org/10.1029/2008JD010026)
- Xu X, Manson AH, Meek CE, Jacobi C, Hall CM, Drummond JR (2011a) Mesospheric wind semidiurnal tides within the canadian middle atmosphere model data assimilation system. *J Geophys Res* 116:D17102. doi:[10.1029/2011JD015966](https://doi.org/10.1029/2011JD015966)
- Xu X, Manson AH, Meek CE, Jacobi C, Hall CM, Drummond JR (2011b) Verification of the mesospheric winds within the canadian middle atmosphere model data assimilation system using radar measurements. *J Geophys Res* 116:D16108. doi:[10.1029/2011JD015589](https://doi.org/10.1029/2011JD015589)

A Coupled Atmosphere-Chemistry Data Assimilation: Impact of Ozone Observation on Structure of a Tropical Cyclone

Seon Ki Park, Sujeong Lim and Milija Županski

Abstract Ozone (O_3) generally shows lower concentration inside the eyewall and higher concentration around the eye in tropical cyclones (TCs). In this study, we identify the impact of O_3 observations on TC structure through a coupled atmosphere-chemistry data assimilation (DA) system. We applied the Weather Research and Forecasting model coupled with Chemistry (WRF-Chem) and an ensemble-based DA algorithm—the maximum likelihood ensemble filter (MLEF) to a case TC over East Asia, Typhoon Nabi (2005). The ensemble forecast with 32 ensembles shows larger background state uncertainty over the TC. The assimilation of O_3 observations, with a 6 h assimilation window, impacts both O_3 itself and wind field in the vicinity of TC. Several measures for verification, including the cost function, root mean square (RMS) error with respect to observations and degrees of freedom for signal (DFS), indicate improvement of the analysis fields through the O_3 DA. The cost function and RMS error have decreased by 17 and 9 %, respectively. The DFS shows large reduction in uncertainty, indicating a strong positive impact of observations in the TC area.

S.K. Park (✉)

Department of Environmental Science and Engineering,
Ewha Womans University, Seoul 03760, Republic of Korea
e-mail: spark@ewha.ac.kr

S. Lim

Department of Atmospheric Science and Engineering,
Ewha Womans University, Seoul 03760, Republic of Korea
e-mail: sjlim1202@gmail.com

S. Lim

Present Address:
Korea Institute of Atmospheric Prediction System, Seoul, Republic of Korea

M. Županski

Cooperative Institute for Research in the Atmosphere, Colorado State University,
Fort Collins, CO 80523-1375, USA
e-mail: Milija.Zupanski@colostate.edu

1 Introduction

Data assimilation (DA) is a process to incorporate observational data into a numerical model to obtain an optimal estimate of model states (Županski 1993; Park and Županski 2003; Reichle 2008; Navon 2009). Since the whole Earth system is composed of several subsystems, e.g., atmosphere, hydrosphere, cryosphere, biosphere and lithosphere, some weather/climate/environment prediction models combine two or more subsystem to better represent the phenomena of interest in the coupled model system (e.g., Johnson et al. 2001; Grell et al. 2005; Betts 2009; Kerkweg and Jockel 2012; Macias et al. 2014; Zhang et al. 2014). DA in a such coupled system is quite challenged due to differences in time scale and availability of observations among different subsystems (Zhang et al. 2007; Sugiura et al. 2008; Zhang et al. 2010; Tardif et al. 2014; Laloyaux et al. 2016).

In coupled DA, the cross-variable components of the forecast error covariances play a major role in transferring information among different variables in a coupled system (Han et al. 2013; Park et al. 2015; Županski 2016). Borovikov et al. (2005) investigated the impact of assimilation of ocean temperature profiles on various model variables such as temperature, salinity, and zonal and meridional velocities. They found that the multivariate error covariance enabled both the temperature and entire ocean-state fields to be updated during an assimilation cycle. They also showed that univariate assimilation scheme made the temperature field fit to observations, yet the structure of the unobserved salinity and current fields deteriorated quickly, which was caused by neglect of the correlation between temperature and salinity when assimilating temperature alone. Park et al. (2015) examined the structure of forecast error covariance in a coupled system and reported that the cross-variable components of the coupled error covariance allowed a physically meaningful adjustment of all control variables and a much wider impact of observations (e.g., atmospheric observation on chemistry analysis, and vice versa). Laloyaux et al. (2016) showed that a coupled model allowed increments to be propagated in the other component, implying that observations from one component have the potential to affect the analysis in the other component within the same assimilation window. Some important issues in coupled data assimilation, especially in terms of error covariance, are discussed in Županski (2016).

For the air quality forecast, associated with emissions, transport and predominant meteorological conditions, the coupled atmosphere-chemistry model is essential (e.g., Carmichael et al. 2008). Ozone (O_3) has a relatively long photochemical lifetime and is a passive tracer at synoptic scale or mesoscale; thus variations of total column O_3 in space and time are strongly linked to the atmospheric flow and many meteorological variables, especially in the upper troposphere (Wu and Zou 2008). In tropical cyclones (TCs), O_3 shows a lower concentration just inside the eyewall and higher concentration around the eye (e.g., Carsey and Willoughby 2005; Zou and Wu 2005; Wu and Zou 2008), due to updraft in the eyewall and downdraft in the eye (Zou and Wu 2005). The daily total column O_3 from Total Ozone Mapping Spectrometer (TOMS) has linear relationship with mean vertically-integrated potential

vorticity (MPV), and has been used to improve hurricane or winter storm prediction (e.g., Jang et al. 2003; Zou and Wu 2005; Wu and Zou 2008).

In this study, we investigate the impact of O_3 observations on TC structure by directly assimilating the total column O_3 from the Ozone Monitoring Instrument (OMI), using a coupled atmosphere-chemistry model and an ensemble-based DA system. Here the cross-correlations between meteorological and chemical variables are obtained directly from ensemble forecasts (e.g., Park et al. 2015).

2 Description on Model and Observation

2.1 Model

In this study, we employ the Weather Research and Forecasting (WRF) model coupled with Chemistry (WRF-Chem; Grell et al. 2005), which can simulate the emission, transport, mixing and chemical transformation of aerosols and atmospheric chemical constituents concurrently with meteorology. Physical options for various atmospheric processes from WRF include the WRF Single-Moment 6-class (WSM6) scheme (Hong and Lim 2006) for the microphysics, the Community Atmospheric Model (CAM) scheme (Collins et al. 2006) for the radiation physics, the Monin-Obukhov scheme (Monin and Obukhov 1953) for the surface layer, the Noah land surface model (Chen and Dudhia 2001) for the land surface, the Yonsei University (YSU) scheme (Hong et al. 2006) for the planetary boundary layer, and the Kain-Fritsch scheme (Kain 2004) for the cumulus convection. For advection, the monotonic transport scheme (Skamarock 2006) is applied to turbulent kinetic energy and scalars such as mixing ratios of water vapor, cloud water, rain, snow and ice and chemical species (see also Wang et al. 2009; Freitas et al. 2011). For gas-phase chemistry, the Carbon Bond Mechanism version Z (CBMZ) without Dimethylsulfide scheme is used.

2.2 Data Assimilation Scheme

For the DA system, we use an ensemble-based method called the maximum likelihood ensemble filter (MLEF; Źupanski 2005; Źupanski et al. 2008). The MLEF generates the analysis solution which maximizes the likelihood of the posterior probability distribution, obtained by minimization of a cost function. It is a hybrid between variational and ensemble DA methods; thus employing a cost function based on a Gaussian probability density function and producing both the analysis and the background error covariance (Źupanski 2005). The MLEF is well suited for nonlinear observation operators, with minimization of cost function using the Hessian preconditioning (Źupanski 2005; Źupanski et al. 2007, 2008), and has been employed in

studies of uncertainty analysis and data assimilation (e.g., Županski and Županski 2006; Kim et al. 2010; Apodaca et al. 2014; Tran et al. 2014).

An interface module has been developed to couple MLEF and WRF-Chem, which transforms the control variables of MLEF into WRF-Chem, and vice versa. For implementing the interface, WRF-Chem is not modified at all.

2.3 *Observational Data*

Satellite data often provide estimates of chemical concentration as a total vertical column through retrieval techniques. They usually cover a wide geographical range compared to other measurements. For this study, we use the total column O_3 from OMI as an observation. The OMI is a nadir-viewing near-UV/Visible CCD spectrometer on board NASA's Aura satellite (OMI Team 2012). The total column O_3 is Level 2 data (OMTO3) based on the Total Ozone Mapping Spectrometer (TOMS) v8.5 algorithm, which is obtained from an orbital swath with a resolution of 13 km \times 24 km at nadir (OMI Team 2012). It achieves global coverage in one day. In this experiment, we employ the OMI data without quality flags because the first appearance of the row anomaly that affects particular viewing directions did not occur in 2005, when Typhoon Nabi occurred.

2.4 *Observation Operator and Bias Correction*

In our DA problem, it is necessary to develop an observation operator transforming the O_3 forecast of WRF-Chem to the total column O_3 observation. The operator contains the calculation of total column O_3 , unit conversion from ppmv (parts per million by volume) to Dobson Units (DU) and the bilinear interpolation to the observation location. The most demanding part of the observation operator is bias correction of total column O_3 observation. Although we use the reference pressure at the model top as 10 hPa, there are still considerable amounts of O_3 in the stratosphere that could not be included in the calculation of the model guess. Since this creates a negative bias in the mean observation error, we introduce a multiplicative bias correction ε to preserve positive-definiteness of the bias-corrected guess (Apodaca et al. 2014) as

$$h_B(\mathbf{x}) = \varepsilon \cdot h(\mathbf{x}). \quad (1)$$

With the multiplicative bias correction in Eq. (1), we can make a new cost function in unbiased form as

$$J(\mathbf{x}) = \frac{1}{2}(\mathbf{x} - \mathbf{x}_b)^T \mathbf{P}_f^{-1} (\mathbf{x} - \mathbf{x}_b) + \frac{1}{2} [\mathbf{y} - h_B(\mathbf{x})]^T \mathbf{R}^{-1} [\mathbf{y} - h_B(\mathbf{x})] \quad (2)$$

where \mathbf{x} is the model state vector, \mathbf{x}_b is the prior (background) state, \mathbf{y} is the observation vector. Here, h is the nonlinear observation operator, \mathbf{P}_f is the forecast (background) error covariance matrix in the ensemble subspace, and \mathbf{R} is the observation error covariance matrix. Equation (2) is the cost function used in DA, provided ε can be estimated. The optimal value of parameter ε is obtained by implicitly assuming lognormal probability density function errors for a multiplicative bias correction in Eq. (1), similarly to Apodaca et al. (2014)—see Eq. (5) therein. The parameter ε is calculated once in every DA cycle.

3 Case Description and Experimental Design

Typhoon Nabi (2005), the case TC in this study, lasted several days from 29 August 2005 to 8 September 2005. Figure 1a shows the best track of Nabi. It moved westward after its formation and passed near Saipan on 31 August as an intensifying TC, transformed to a super typhoon on 1 September, and reached its peak with winds of 175 km h^{-1} (10-min average) on 2 September. It became weak while turning to the north and striking Kyushu on 6 September. Nabi turned to the northeast after passing by the Korean Peninsula, and transformed to an extratropical cyclone passing over Hokkaido on 8 September. Figure 1b shows the total column O_3 from OMI at 0405 UTC 3 September 2005. It shows a lower concentration just inside the eyewall and higher concentration around the eye. This feature is well described when the TC has the strongest intensity in the intensifying stages (e.g., Carsey and Willoughby 2005). Although Typhoon Nabi (2005) reached the maximum intensity on 2 September, OMI entered in the zoom-in mode (i.e., no ozone data) in our domain on that date; thus we have alternatively chosen 3 September for the analysis of O_3 properties during the maximum development of the case TC.

We focus on a single DA cycle from 0000 to 0600 UTC 3 September 2005, which is one of the strongest periods of its lifetime. We conduct the experiment with 32 ensembles and 6 h assimilation window. Note that the OMI observations have an approximate frequency of once per day over the typhoon and the surrounding geographical area. Therefore, adding more DA cycles would not be beneficial since no additional data are available.

The initial and lateral boundary conditions for atmospheric states are provided by the National Centers for Environmental Prediction (NCEP) Global Forecasting System (GFS), while those for chemical variables are obtained from the Model for Ozone and Related chemical Tracers (MOZART; Emmons et al. 2010) chemistry global model. The WRF-Chem is set up with a horizontal resolution of 30 km and 51 vertical levels with the bottom at the ground and the top at 10 hPa using a terrain-following hydrostatic pressure coordinate (Skamarock et al. 2008).

The model domain is centered over the Korean Peninsula, covering an area of approximately $3900 \text{ km} \times 4400 \text{ km}$ with 132×147 horizontal grid points. The control variables defined in the coupled atmosphere-chemistry DA are the WRF-Chem prognostic variables that contain dynamical variables such as winds, perturbation

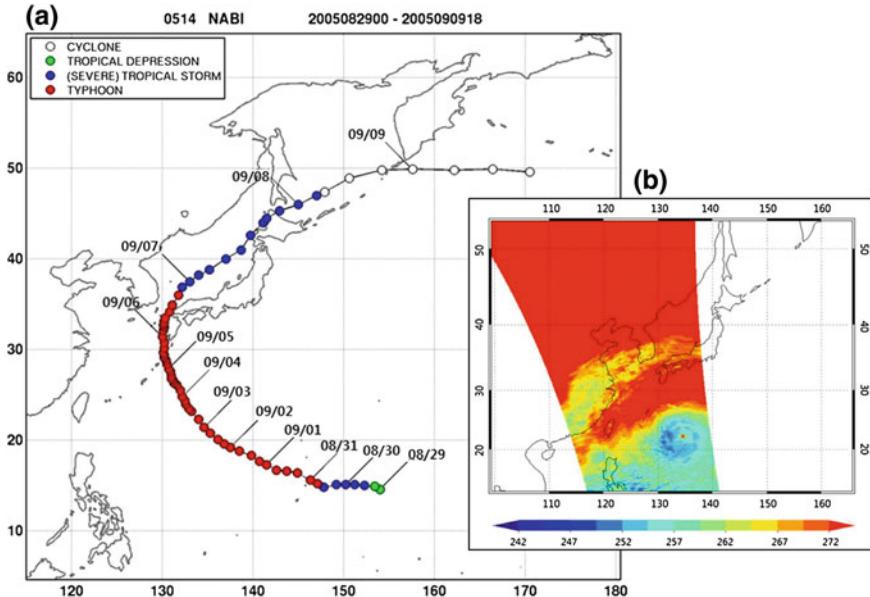


Fig. 1 **a** The best track of Typhoon Nabi (2005) from 29 August to 9 September 2005, and **b** total column O₃ (in DU) from OMI at 0405 UTC, 3 September 2005. The best track analysis is provided by Typhoon Research Center (<http://www.typhoon.or.kr/>) and modified. The OMI image is modified from Lim et al. (2015)

potential temperature, perturbation geopotential, water vapor mixing ratio and perturbation dry air mass in a column, and the chemical variables such as ozone (O₃), nitrates (NO, NO₂, NO₃), and sulfur dioxide (SO₂). The experiments consist of (i) the forecast (without DA) which is useful to understand the synoptic situation and background error covariance, and (ii) the analysis (with DA) which is useful to understand the assimilation impacts.

4 Results

4.1 Background State Uncertainty

The background state uncertainty can be represented by the background error covariance (e.g., Kim et al. 2010). These are estimated by the difference between each of the 32 ensemble members and the control forecast in the ensemble system (Zhang et al. 2013). The initial ensemble perturbations are generated by using the lagged forecast outputs in this experiment, as in Zhang et al. (2013).

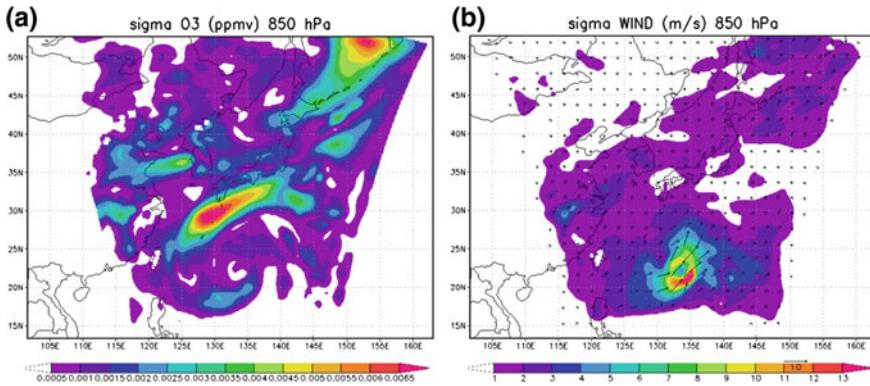


Fig. 2 Standard deviation of background error covariance for **a** O_3 (in ppmv) and **b** wind (in $m s^{-1}$) valid on 0600 UTC 3 September 2005 at 850 hPa. Modified from Lim et al. (2015)

The flow-dependent background error covariance, calculated from the WRF-Chem ensemble forecast and defined for atmospheric and chemical variables, allows chemistry observations to impact atmospheric variables in DA. Our results identify the larger background state uncertainty near the TC, similar to Kim et al. (2010) and Zhang et al. (2013). Figure 2 shows the standard deviation of background error covariance for O_3 and wind at 850 hPa. The O_3 fields depict a large background state uncertainty near the TC (Fig. 2a). The wind fields also show a large background state uncertainty near the TC, especially in the eye region (Fig. 2b). Provided that total column O_3 observations are available, a larger background state uncertainty potentially implies a greater analysis correction.

4.2 Impact of O_3 Data Assimilation

We assess the impact of the assimilated O_3 observations using analysis increments ($\mathbf{x}_a - \mathbf{x}_b$), which show the correction of the background state by the observations. Figure 3 shows the analysis increments ($\mathbf{x}_a - \mathbf{x}_b$) of O_3 and wind at 850 hPa, obtained by assimilating O_3 observations. The O_3 analysis increments are in agreement with background state uncertainties. The O_3 analysis increment has an increase near the TC, but a decrease over China (Fig. 3a). The analysis increments of wind by O_3 assimilation are shown in Fig. 3b. Corresponding to background state uncertainties, the analysis increments of wind show notable impact at the low level. Positive O_3 increments correspond to positive wind increments at 850 hPa, especially in the eye region. Our results illustrate that chemical observations can impact not only the chemical variables but also the atmospheric variables, due to using the ensemble-based coupled atmosphere-chemistry background error covariance.

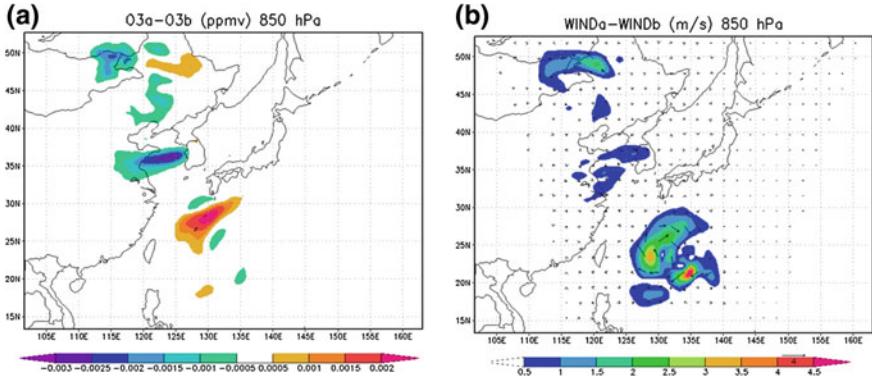


Fig. 3 Same as in Fig. 2 except for analysis increment ($x_a - x_b$) of **a** O_3 (in ppmv) and **b** wind (in $m s^{-1}$), in response to total column O_3 . Modified from Lim et al. (2015)

4.3 Analysis Improvement and Uncertainty Reduction

We verify the impact of O_3 assimilation in terms of the cost function and on the root mean square (RMS) error with respect to O_3 observations—the same data used in the analysis. The cost function of O_3 (see Eq. (2)) has decreased from 0.36924×10^4 (background) to 0.30689×10^4 (analysis), i.e., it is reduced by $\sim 17\%$. The RMS error has also decreased from 0.16684×10^2 DU (background) to 0.15204×10^2 DU (analysis), i.e., by $\sim 9\%$. Our results suggest that O_3 assimilation has produced a notable improvement in the initial conditions.

In addition, the impact of total column O_3 observations is also quantified in terms of the uncertainty reduction. With the Gaussian probability assumption, the information content of observations can be represented as the degrees of freedom for signal (DFS; d_s) (e.g., Rodgers 2000) as

$$d_s = \sum_i \frac{\lambda_i^2}{1 + \lambda_i^2} \quad (3)$$

where λ_i are the eigenvalues of the observation information matrix (e.g., Županski et al. 2007). Note that d_s are strictly a non-negative measure: positive values indicate a reduction of uncertainty due to assimilation, while zero values indicate no impact of observations. The DFS generally coincided with the satellite path and the maximum impact area occurred near the TC location (not shown). This implies that the total column O_3 observation had the strongest impact in the TC area.

5 Conclusions

In this study, the impact of ozone (O_3) data assimilation (DA) on the tropical cyclone (TC) structure has been investigated for Typhoon Nabi (2005). The total column O_3 from the Ozone Monitoring Instrument (OMI) is directly assimilated into a coupled atmosphere-chemistry modelling system—the Weather Research and Forecasting (WRF) model coupled with Chemistry (WRF-Chem). For the DA method, the maximum likelihood ensemble filter (MLEF) is employed and interfaced with the WRF-Chem. Because the OMI observations cover the model domain only once per day, and no other observations were available at the time, only a single DA cycle is performed. The single DA cycle may limit the robustness of the DA system; however, it does not impact the performance of the coupled atmosphere-chemistry DA system.

It turns out that the O_3 DA has a significant improvement on the analyses of O_3 itself and wind field, especially near the case TC. Lim et al. (2015) showed that the O_3 DA also improves the analysis accuracy of other atmospheric variables such as temperature and specific humidity, which are closely related to the TC structure and other properties. Other dynamical and thermodynamical variables of the TC can be affected by the O_3 observations. For example, the TC development is related to temperature, intensity to wind, and rainfall to specific humidity. Thus, the corrections of such variables in the initial conditions of atmospheric environment near the TC areas have a potential to modify the TC structures and eventually improve the the forecast accuracy of TCs.

In our experiments, several measures related to verification of O_3 DA reveal improvement of the analysis fields compared to the first guess. For instance, the ensemble forecast error, represented by the background error standard deviation, had larger uncertainty over the TC area; however, both the cost function and root mean square error reduced notably after the O_3 DA. Such reduction indicates an improvement of the analysis state. The degrees of freedom for signal showed strong positive values near the TC area, indicating a reduction of the uncertainty of analysis. Based on our results, with only one available observation product per day for O_3 , a positive impact of assimilation can be expected for the TC forecasts

Acknowledgements This work is supported by the Korea Environmental Industry & Technology Institute through the Eco Innovation Program (ARQ201204015), and partly by the National Research Foundation of Korea grant (No. 2009-0083527) funded by the Korean government (MSIP). The third author acknowledges a partial support from the National Science Foundation Collaboration in Mathematical Geosciences Grant 0930265 and the NASA Modeling, Analysis and Prediction (MAP) Program Grant NNX13AO10G.

References

- Apodaca K, Županski M, DeMaria M, Knaff JA, Grasso LD (2014) Development of a hybrid variational-ensemble data assimilation technique for observed lightning tested in a mesoscale model. *Nonlinear Process Geophys* 21:1027–1041
- Betts AK (2009) Land-surface-atmosphere coupling in observations and models. *J Adv Model Earth Syst* 1:4. doi:[10.3894/JAMES.2009.1.4](https://doi.org/10.3894/JAMES.2009.1.4)
- Borovikov A, Rienecker MM, Keppenne CL, Johnson GC (2005) Multivariate error covariance estimates by Monte Carlo simulation for assimilation studies in the Pacific Ocean. *Mon Weather Rev* 133:2310–2334
- Carmichael GR, Sandu A, Chai T, Daescu DN, Constantinescu EM, Tang Y (2008) Predicting air quality: improvements through advanced methods to integrate models and measurements. *J Comput Phys* 227:3540–3571
- Carsey TP, Willoughby HE (2005) Ozone measurements from eyewall transects of two Atlantic tropical cyclones. *Mon Weather Rev* 133:166–174
- Chen F, Dudhia J (2001) Coupling an advanced land surface-hydrology model with the Penn State-NCAR MM5 modeling system. Part I: Model implementation and sensitivity. *Mon Weather Rev* 129:569–585
- Collins WD, Rasch PJ, Boville BA, Hack JJ, McCaa JR, Williamson DL, Briegleb BP, Bitz CM, Lin SJ, Zhang M (2006) The formulation and atmospheric simulation of the Community Atmosphere Model Version 3 (CAM3). *J Clim* 19:2144–2161
- Emmons LK, Walters S, Hess PG, Lamarque J-F, Pfister GG, Fillmore D, Granier C, Guenther A, Kinnison D, Laepple T, Orlando J, Tie X, Tyndall G, Wiedinmyer C, Baughcum SL, Kloster S (2010) Description and evaluation of the Model for Ozone and Related chemical Tracers, version 4 (MOZART-4). *Geosci Model Dev* 3:43–67
- Freitas SR, Rodrigues LF, Longo KM, Panetta J (2011) Impact of a monotonic advection scheme with low numerical diffusion on transport modeling of emissions from biomass burning. *J Adv Model Earth Syst* 3:M01001. doi:[10.1029/2011MS000084](https://doi.org/10.1029/2011MS000084)
- Grell G, Peckham S, Schmitz R, McKeen S, Frost G, Skamarock W, Eder B (2005) Fully coupled “online” chemistry within the WRF model. *Atmos Environ* 39:6957–6975
- Han G, Wu X, Zhang S, Liu Z, Li W (2013) Error covariance estimation for coupled data assimilation using a Lorenz atmosphere and a simple pycnocline ocean model. *J Climate* 26:10218–10231
- Hong S-Y, Lim J-OJ (2006) The WRF single-moment 6-class microphysics scheme (WSM6). *J Korean Meteor Soc* 42:129–151
- Hong S-Y, Noh Y, Dudhia J (2006) A new vertical diffusion package with an explicit treatment of entrainment processes. *Mon Weather Rev* 134:2318–2341
- Jang KI, Zou X, De Pondeca MSFV, Shapiro M, Davis C, Krueger A (2003) Incorporating TOMS ozone measurements into the prediction of the Washington, DC, winter storm during 24–25 January 2000. *J Appl Meteor* 42:797–812
- Johnson CE, Stevenson DS, Collins WJ, Derwent RG (2001) Role of climate feedback on methane and ozone studied with a coupled ocean-atmosphere-chemistry model. *Geophys Res Lett* 28:1723–1726
- Kain JS (2004) The Kain-Fritsch convective parameterization: An update. *J Appl Meteor* 43:170–181
- Kerkweg A, Jockel P (2012) The 1-way on-line coupled atmospheric chemistry model system MECO(n)—Part 1: Description of the limited-area atmospheric chemistry model COSMO/MESSy. *Geosci Model Dev* 5:87–110
- Kim HH, Park SK, Županski D, Županski M (2010) Uncertainty analysis using the maximum likelihood ensemble filter and WRF and comparison with dropwindsonde observations in Typhoon Sinlaku (2008). *Asia-Pac J Atmos Sci* 46:317–325
- Laloyaux P, Balmaseda M, Dee D, Mogensen K, Janssen P (2016) A coupled data assimilation system for climate reanalysis. *Q J R Meteor Soc* 142:65–78

- Lim S, Park SK, Županski M (2015) Ensemble data assimilation of total column ozone using a coupled meteorology-chemistry model and its impact on the structure of Typhoon Nabi (2005). *Atmos Chem Phys* 15:10019–10031
- Macias DM, Guerreiro CT, Prieto L, Peliz A, Ruiz J (2014) A high-resolution hydrodynamic-biogeochemical coupled model of the Gulf of Cadiz—Alboran Sea region. *Medit Mar Sci* 15:739–752
- Monin AS, Obukhov AM (1953) Dimensionless characteristics of turbulence in the atmospheric surface layer. *Dok AN SSSR* 93:223–226
- Navon IM (2009) Data assimilation for numerical weather prediction: a review. In: Park SK, Xu L (eds) *Data assimilation for atmospheric oceanic and hydrologic applications*. Springer, Berlin, pp 21–65
- OMI Team (2012) Ozone Monitoring Instrument (OMI) Data User's Guide. NASA, Greenbelt, 62 pp
- Park SK, Županski D (2003) Four-dimensional variational data assimilation for mesoscale and storm-scale applications. *Meteor Atmos Phys* 82:173–208
- Park SK, Lim S, Županski M (2015) Structure of forecast error covariance in coupled atmosphere-chemistry data assimilation. *Geosci Model Dev* 8:1315–1320
- Reichle RH (2008) Data assimilation methods in the Earth sciences. *Adv Water Res* 31:1411–1418
- Rodgers CD (2000) *Inverse methods for atmospheric sounding: theory and practice*. World Scientific, Singapore 256 pp
- Skamarock WC (2006) Positive-definite and monotonic limiters for unrestricted-time-step transport schemes. *Mon Weather Rev* 134:2241–2250
- Skamarock WC, Klemp JB, Dudhia J, Gill DO, Barker DM, Duda MG, Huang X-Y, Wang W, Powers JG (2008) A description of the advanced research WRF version 3. NCAR/TN-475+STR, National Center For Atmospheric Research, Boulder, CO, 113 pp
- Sugiura N, Awaji T, Masuda S, Mochizuki T, Toyoda T, Miyama T, Igarashi H, Ishikawa Y (2008) Development of a four-dimensional variational coupled data assimilation system for enhanced analysis and prediction of seasonal to interannual climate variations. *J Geophys Res* 113:C10017. doi:[10.1029/2008JC004741](https://doi.org/10.1029/2008JC004741)
- Tardif R, Hakim GJ, Snyder C (2014) Coupled atmosphere-ocean data assimilation experiments with a low-order climate model. *Clim Dyn* 43:1631–1643
- Tran AP, Vanclooster M, Županski M, Lambot S (2014) Joint estimation of soil moisture profile and hydraulic parameters by ground-penetrating radar data assimilation with maximum likelihood ensemble filter. *Water Resour Res* 50:3131–3146. doi:[10.1002/2013WR014583](https://doi.org/10.1002/2013WR014583)
- Wang H, Skamarock WC, Feingold G (2009) Evaluation of scalar advection schemes in the Advanced Research WRF model using large-eddy simulations of aerosol-cloud interactions. *Mon Weather Rev* 137:2547–2558
- Wu Y, Zou X (2008) Numerical test of a simple approach for using TOMS total ozone data in hurricane environment. *Q J R Meteor Soc* 134:1397–1408
- Zhang S, Harrison MJ, Rosati A, Wittenberg A (2007) System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon Weather Rev* 135:3541–3564
- Zhang S, Rosati A, Delworth T (2010) The adequacy of observing systems in monitoring the Atlantic meridional overturning circulation and North Atlantic climate. *J Clim* 23:5311–5324
- Zhang S, Zhao M, Lin S-J, Yang X, Anderson W (2014) Retrieval of tropical cyclone statistics with a high-resolution coupled model and data. *Geophys Res Lett* 41:652–660
- Zhang SQ, Županski M, Hou AY, Lin X, Cheung SH (2013) Assimilation of precipitation-affected radiances in a cloud-resolving WRF ensemble data assimilation system. *Mon Weather Rev* 141:754–772
- Zou X, Wu Y (2005) On the relationship between Total Ozone Mapping Spectrometer (TOMS) ozone and hurricanes. *J Geophys Res* 110:D06109. doi:[10.1029/2004JD005019](https://doi.org/10.1029/2004JD005019)
- Županski D, Županski M (2006) Model error estimation employing an ensemble data assimilation approach. *Mon Weather Rev* 134:1337–1354

- Županski D, Hou AY, Zhang SQ, Županski M, Kummerow CD, Cheung SH (2007) Applications of information theory in ensemble data assimilation. *Q J R Meteor Soc* 133:1533–1545
- Županski M (1993) Regional four-dimensional variational data assimilation in a quasi-operational forecasting environment. *Mon Weather Rev* 121:2396–2408
- Županski M (2005) Maximum likelihood ensemble filter: theoretical aspects. *Mon Weather Rev* 133:1710–1726
- Županski M (2016) Data assimilation for coupled modeling systems. In: Park SK, Xu L (eds) *Data assimilation for atmospheric oceanic and hydrologic applications*, vol. III. Springer, Berlin, pp xx–xx
- Županski M, Navon IM, Županski D (2008) The maximum likelihood ensemble filter as a non-differentiable minimization algorithm. *Q J R Meteor Soc* 134:1039–1050

Adjoint Sensitivity with a Nested Limited Area Atmospheric Model

Clark Amerault

Abstract A one-way interactive nesting capability was added to the tangent linear and adjoint models of a limited area atmospheric forecasting system. In the tangent linear model, perturbation information is passed through the lateral boundaries to higher resolution nests. Gradient information moves through the lateral boundaries in the direction of the lower resolution parent domains in the adjoint model. The system was demonstrated by forcing the adjoint model over a small volume resolved on the finest nest and constructing optimal perturbations from the gradient information on the parent domain. Small perturbations ($<1 \text{ m s}^{-1}$ in wind speed) on the coarsest domain resulted in relatively large changes in the boundary layer flow over the localized area on the finest scale nest where the adjoint model was originally forced.

1 Introduction

A primitive equation numerical weather prediction (NWP) model provides an estimate of the future state of the atmosphere based on the initial conditions input to the model. The atmospheric model's state is discretized in both space and time. Smaller grid spacings reduce model error, but also increase the computational expense. Incorporating a nest with relatively small grid spacing over an area of interest inside of a larger domain is a way to balance computing cost with the desire to resolve finer scale features.

Early nested models were utilized in simulating the mesoscale features of tropical cyclones Birchfield (1960). Ley and Elsberry (1976) provide a summary of the pioneering accomplishments in nested modeling. Because of the spectral nature of global NWP models, nesting is a capability reserved for grid point limited area models. Current operational forecast quality limited area NWP models incorporate various nesting capabilities Hodur (1997), Pielke et al. (1992), Dudhia (1993), Skamarock et al. (2001), Michalakes et al. (2001). Although it is not the focus

C. Amerault (✉)
Naval Research Laboratory, Monterey, CA, USA
e-mail: clark.amerault@nrlmry.navy.mil

of this chapter, the next generation of NWP models will rely on adaptive mesh techniques Skamarock (1989), meaning the model will be run on a single domain (encompassing either the entire globe or a limited area) with varying horizontal grid spacing.

Adjoint models of NWP systems provide gradients of scalar forecast aspects with respect to an earlier model state. They have been widely employed for data assimilation and sensitivity studies. On larger scales, global NWP adjoint models are utilized in creating four dimensional variational (4D-Var) analyses (see Rabier (2005) for a review) and for targeting observations (see Langland (2005) for a review). Limited area adjoint models Sun and Crook (1997), Zou et al. (1997), Županski et al. (2005) have been utilized in experiments to produce analyses of meso and cloud scale structures like tropical cyclones Zou and Xiao (2000).

As with their corresponding NWP systems, adjoint models require greater computing resources as the horizontal grid spacing decreases. Because of the increased computing and developmental expenses associated with nesting, adjoint models have only been developed for single domain configurations. In this study, a nesting capability was added to the tangent linear and adjoint models of a limited area NWP system. With this capability, the evolution of perturbations can be tracked in the tangent linear model across nests down to the smallest grid spacings. In the nested adjoint model, forcings related to a fine scale feature are integrated backward in time to reveal sensitivities to larger scales.

This chapter includes an outline of the steps taken to add a nesting capability to these models and examples of the fields produced by the enhanced system. The limited area atmospheric model utilized for this work is presented in the next section along with a discussion on the process of adding nests to the tangent linear and adjoint models. The new capabilities are tested for coastal atmospheric flow in Sect. 3 and a brief summary is provided in Sect. 4.

2 COAMPS

2.1 *COAMPS Atmospheric Model*

The atmospheric portion of the Naval Research Laboratory's (NRL) Coupled Ocean Atmosphere Mesoscale Prediction System (COAMPS®¹) is the nested model Hodur (1997) used in this study. The COAMPS atmospheric model is a limited area, relocatable, grid point model. The model is non-hydrostatic and contains predictive equations for zonal wind u , meridional wind v , vertical velocity w , the dimensionless Exner pressure function π , the potential temperature θ , water vapor q_v , and turbulent kinetic energy e . The bulk cloud microphysics scheme calculates the source and sink terms in the prognostic equations for cloud droplets q_c , cloud ice q_i , rain

¹COAMPS® registered trademark of the Naval Research Laboratory.

water q_r , snow q_s , and graupel q_g . The other major parameterizations in the model for subgrid-scale processes include turbulent mixing, surface fluxes, cumulus convection, and radiation. The vertical coordinate of the model is a terrain following σ_z defined as

$$\sigma_z = \frac{z_t(z - z_s)}{z_t - z_s}, \quad (1)$$

where the constant z_t is the depth of the model domain and z_s is the terrain height.

Lateral boundary conditions for the coarsest parent domain are provided by forecast fields from NRL's global atmospheric model. Analysis fields are created with the three dimensional NRL Atmospheric Variational Data Assimilation System Daley and Barker (2001).

2.2 COAMPS Tangent Linear and Adjoint Atmospheric Models

Adjoint models provide the gradient of some scalar function J of the state vector of a model \mathbf{x}_t at time t with respect to the initial state vector of the model \mathbf{x}_0 . The state vector depends on the initial conditions of the model, so

$$J(\mathbf{x}_t) = J(M(\mathbf{x}_0)), \quad (2)$$

where M is the nonlinear model. In sensitivity studies J is referred to as a response function. The gradient of J with respect to the initial model state is

$$\frac{\partial J}{\partial \mathbf{x}_0} = \mathbf{M}^T \frac{\partial J}{\partial \mathbf{x}_t}, \quad (3)$$

where \mathbf{M} is the tangent linear model and the superscript T denotes the transpose operation. In real number space, the adjoint model \mathbf{M}^T is formulated by realizing the transpose of the tangent linear model. The tangent linear model is needed to construct the adjoint model and can also be used to integrate perturbations fields. The forcing for the adjoint model, $\frac{\partial J}{\partial \mathbf{x}_t}$ is calculated simultaneously with J by differentiating J with respect to model state at t . The nonlinear and tangent linear models are integrated with time increasing and will also be referred to as forward models in this chapter. Conversely, the adjoint model is integrated with time decreasing and will also be referred to as the backward model.

In addition to the dynamical core, the COAMPS tangent linear and adjoint models include the respective components of the nonlinear model turbulent mixing, surface flux, moist physics, and cumulus schemes. More information on the adjoint COAMPS atmospheric model can be found in Amerault et al. (2008). In this study the moist physics and cumulus schemes are excluded from the tangent linear and adjoint model integrations.

Table 1 Weightings for perturbation calculation

Variable	Weight
u	$4.0 \text{ m}^2 \text{ s}^{-2}$
v	$4.0 \text{ m}^2 \text{ s}^{-2}$
w	$0.04 \text{ m}^2 \text{ s}^{-2}$
π	0.0001 (dimensionless)
θ	4.0 K^2
q_v	$4.0 \text{ g}^2 \text{ kg}^{-2}$

To showcase the nesting capability, the adjoint model will be used to create optimal perturbations Oortwijn (1995), Rabier et al. (1996), Errico and Raeder (1999). The j th element of the perturbation vector $\delta\mathbf{x}$ is calculated using,

$$\delta x_j = s w_j \frac{\partial J}{\partial x_0^j}. \quad (4)$$

The weightings w_j only vary by model variable and the values are listed in Table 1 for the variables included in the perturbation vector (the hydrometeors and turbulent kinetic energy are not perturbed). A constraint on the size of the initial perturbation is imposed by the choice of the scaling parameter s . The units of s are the inverse of the units of J ensuring that the elements of $\delta\mathbf{x}$ have the proper units. Here, s is chosen so that the maximum magnitude of the u perturbation is less than or equal to 1.0 m s^{-1} everywhere in the domain. In practice, the adjoint model is integrated backward in time. The gradient values are multiplied by the proper values of s and w_j to form the initial perturbation. The initial perturbation fields retain the same spatial structures they had in the final gradient fields. The perturbation is then input to the tangent linear model to track its evolution.

2.3 Nesting

The COAMPS atmospheric model contains many nesting options. These include one and two-way interactions between nests, delayed start times and early ending times for child nests, and the ability to move nests during the model's integration. In a one-way nesting configuration, information is only communicated through the lateral boundaries from the parent domain to its child. For two-way interactions, there is a feedback of information over the entire area covered by the child domain onto the parent domain in addition to the lateral boundary exchange. This feedback is performed at a specified time interval, usually the length of a time step on the parent domain.

In this study, one-way interactions for telescoping static nests were added to the COAMPS tangent linear and adjoint models. Information is only communicated through the lateral boundaries for this type of nest interaction. Figure 1 shows the

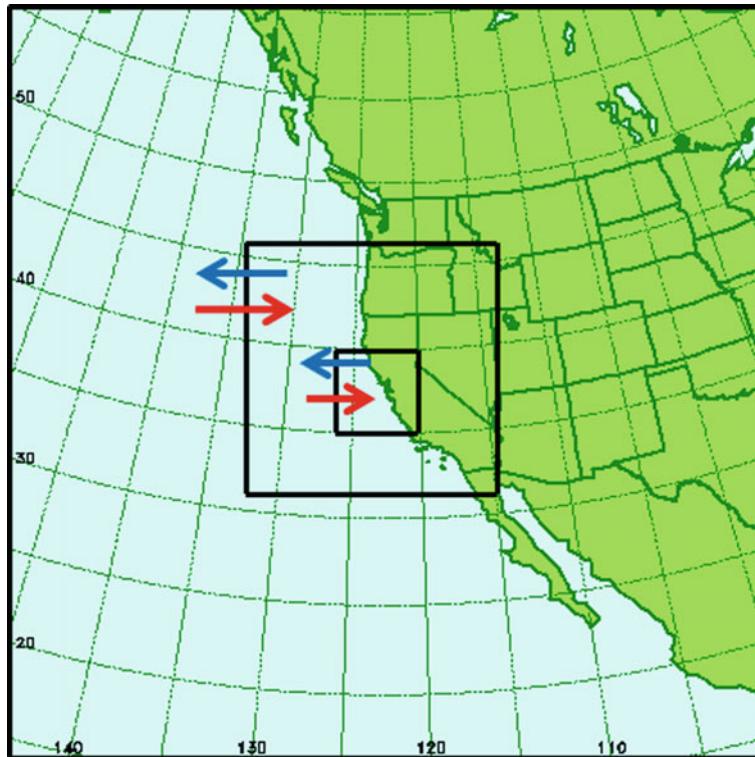
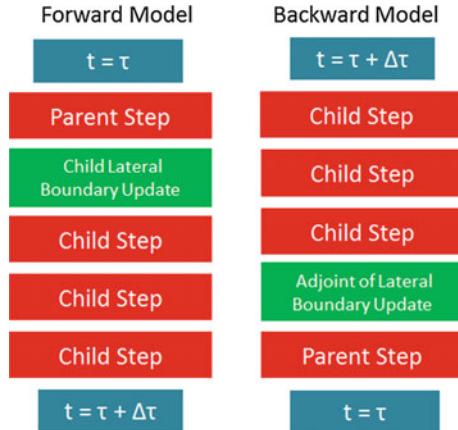


Fig. 1 Area covered by the three domains used for the COAMPS atmospheric model integrations in this study. The *red arrows* indicate the direction of information flow from the coarse parent domain to the finest scale nest across the lateral boundaries in the nonlinear and tangent linear models. The flow is reversed (*blue arrows*) in the adjoint model

area covered by the 3 domains used in this study. COAMPS utilizes a 3 to 1 horizontal grid spacing ratio, so the outermost domain is nine times as coarse as the innermost nest in this three nest example. The vertical grid spacing is identical on every domain. In the nonlinear and tangent linear models, information flows from the coarse mesh down to the finest scale nest as indicated by the red arrows in Fig. 1. In the adjoint model, the flow of information is reversed, going from the finest scale nest to the outermost domain (blue arrows).

Figure 2 outlines the execution sequence in one coarse model time step for the case of only 2 domains (parent and child) with one-way interactions. Since the spatial and temporal flow of information is in the same direction in the nonlinear and tangent linear models and the process of passing the information through the nest lateral boundaries is linear, most of the code needed to add the nesting capability to the tangent linear model was already available. Code was added to pass the perturbation fields to finer scale nests using the routines already developed for the full nonlinear fields. However, in the adjoint model the interaction is reversed, so an adjoint routine

Fig. 2 Schematic of the forward (nonlinear and tangent linear models) and backward (adjoint model) integrations for one coarse model time step



that feeds the information on the child nest through the lateral boundaries to the parent domain was constructed.

In the forward models, the execution sequence begins with the parent domain advancing forward one time step. The lateral boundaries on the child domain are updated based on the new parent domain values. The child nest is then integrated forward on three smaller time steps. The ratio of the length of the time step on the parent and child domains is equivalent to the horizontal grid spacing ratio, 3 to 1.

In the adjoint model, the child nest is first integrated backward in time on the three small time steps. The adjoint of the lateral boundary update is then performed to pass information from the child to the parent nest. Finally, an adjoint model time step is performed for the parent domain.

A nonlinear model trajectory is required for the tangent linear and adjoint models. The trajectory is saved for each domain on every time step of the respective domain. For the 3 domain setup used in this study, the trajectory on the innermost nest is saved nine times as frequently as the outermost nest. This requires ample storage. The storage cost can be reduced with less frequent saves and temporal interpolation of trajectory values. However, this would require more code construction and is not the focus of this study.

This demonstration of the nested tangent linear and adjoint models is concerned with creating the optimal perturbations discussed in Sect. 2.2. To initiate the tangent linear model, the perturbation field is only input to the coarsest domain. This perturbation field is interpolated to the child nests before the model integration begins. In the adjoint model, forcings are only input to the finest scale child nest. At the end of the adjoint integration, the gradient information on each nest is placed on the coarsest parent domain using the adjoint version of the code used to interpolate the perturbations to finer scales in the tangent linear models. Just as with the choice of trajectory, this setup was chosen for simplicity. In the future, options could be added to force the adjoint model on multiple nests and create perturbations on each domain, not just the coarsest.

In addition to the optimal perturbations presented in this chapter, the nested tangent linear and adjoint models were run through standard tests to ensure the added code was done correctly. The asymptotic behavior of the following was checked

$$\lim_{\lambda \rightarrow 0} \frac{M(\mathbf{x} + \lambda \delta \mathbf{x}) - M(\mathbf{x})}{\mathbf{M} \lambda \delta \mathbf{x}} = 1 \quad (5)$$

and displayed the same behavior with and without the nesting capability. The following identity was used to verify the correctness of the adjoint model against the tangent linear model,

$$(\mathbf{M} \delta \mathbf{x})^T \mathbf{M} \delta \mathbf{x} = \delta \mathbf{x}^T \mathbf{M}^T \mathbf{M} \delta \mathbf{x}. \quad (6)$$

This identity is checked by running the tangent linear model with a perturbation and calculating the dot product of the result (left hand side of Eq. 6). The evolved perturbation is then input to the adjoint model and dot product is computed between the resulting field and the initial perturbation (right hand side of Eq. 6). The identity was verified to machine precision with the nesting capability active.

3 Application to Coastal Flow

Optimal perturbations using the nested tangent linear and adjoint models were computed for coastal California atmospheric flow. During the summer months, northerly flow in the marine atmospheric boundary layer is capped by a strong temperature inversion and channeled along the California coastline by the accompanying terrain. The flow becomes supercritical when the Froude number (dimensionless ratio of the fluid speed to the phase speed of internal gravity waves) is greater than unity. Capes and points modify the flow so that localized regions of both super and subcritical flow exist. Downwind of the coastal feature, air accelerates and the boundary layer compresses, while on the upwind side of a cape, the flow decelerates and the boundary layer deepens. Beardsley et al. (1987) were the first to investigate these flow features. Others Samelson (1992), Haack et al. (2001) have researched various aspects of this flow usually focusing on the geographic features that influence it. This study will both demonstrate the nested capabilities of the COAMPS adjoint and tangent linear models and also highlight how the flow can be sensitive to small perturbations in the atmospheric state away from the coastal boundary layer.

The configuration for this study uses three telescoping domains covering the area shown in Fig. 1. The horizontal grid spacings for the outermost (nest 1), middle (nest 2), and innermost (nest 3) domains are 45, 15, and 5 km, respectively. Each domain has 45 unequally spaced vertical levels with smaller spacings in the lower levels of the atmosphere. The time step for nest 1 is 120 s. The experiment was initialized at 1200 UTC Jun 3 2015. The 12 h forecasts of wind speed and stream function at the lowest and 30th (corresponding to a pressure level of roughly 350 mb over the ocean) levels are shown on nest 3 in Fig. 3. These forecast are valid at 1700 local time, corresponding to the late afternoon. Variations in the flow are resolved by the

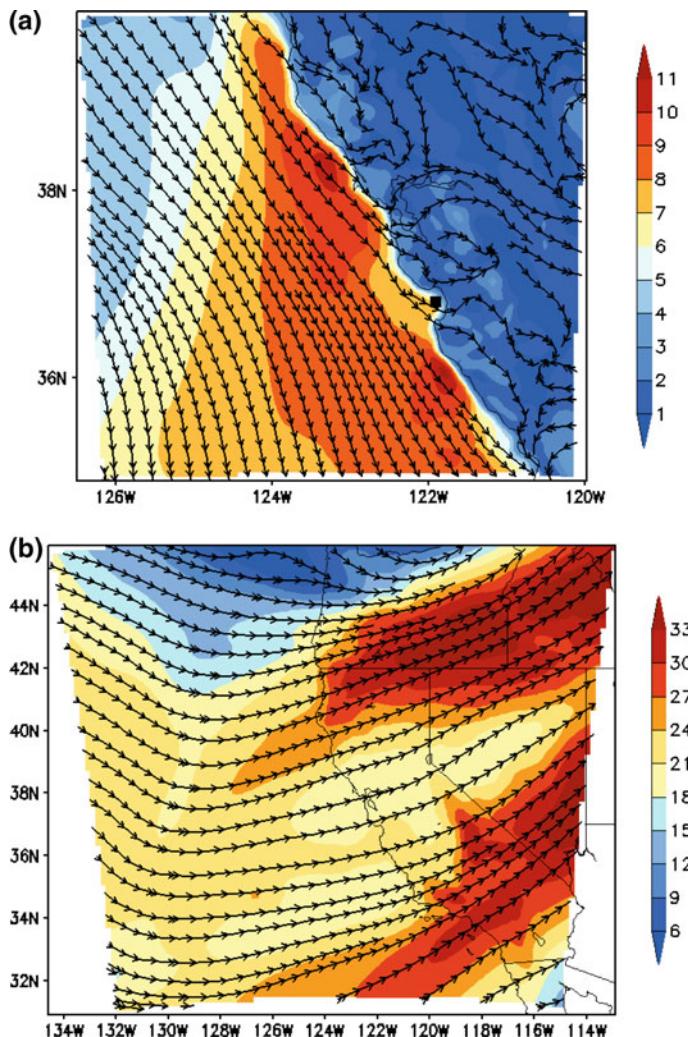


Fig. 3 12 h forecast of wind speeds (m s^{-1}) and streamlines on nest 3 valid at 0000 UTC Jun 4 2015 at model level **a** 1 and **b** 30

innermost nest and two maxima in excess of 10 m s^{-1} are present in the flow. The northern maximum is located north of Point Reyes while the southern maximum is south of Point Sur. Primarily zonal flow is dominant in the model levels above the boundary layer.

This study is focused on the strength of the flow over Monterey Bay. The response function for these experiments is the integrated horizontal kinetic energy $u^2 + v^2$ calculated over a small volume using the 12 h forecast fields. The volume comprises 3 grid points on each side (15 km^2) in the horizontal and the lowest 7 model layers (roughly 300 m) in the vertical. The horizontal area covered by the volume is indi-

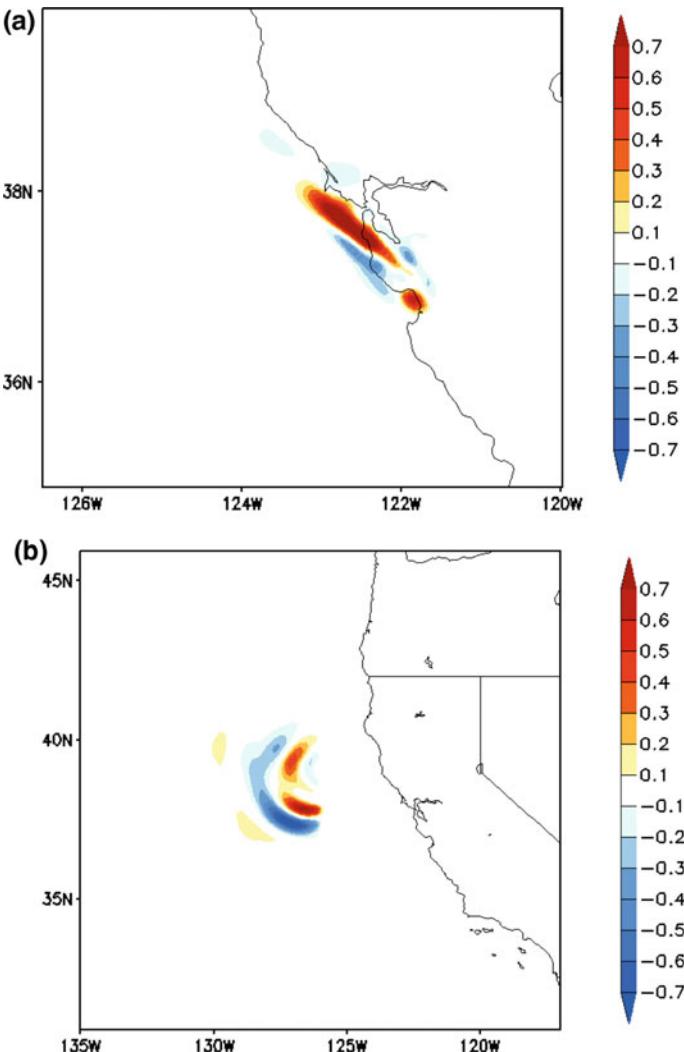


Fig. 4 Sensitivities of response function with respect to u at **a** 2000 UTC June 3 2015 on model level 8 and nest 3 and **b** 1200 UTC June 3 2015 on model level 21 and nest 2. The units are $10^4 \text{ m s}^{-1} \text{ km}^{-3}$

cated by the black box in Fig. 1a. The largest sensitivities move in opposition to the primary flow during the adjoint model integration, as can be seen in Fig. 4. As indicated by Lewis et al. (2001), adjoint sensitivities need to be properly scaled to account for varying resolution. The sensitivities are divided by the grid cell volume as was done in Ancell and Mass (2006). In the boundary layer, the sensitivities travel northward parallel to the coast (Fig. 4a). Above the boundary layer, sensitivities head in a more westward direction due to the zonal flow (Fig. 4b).

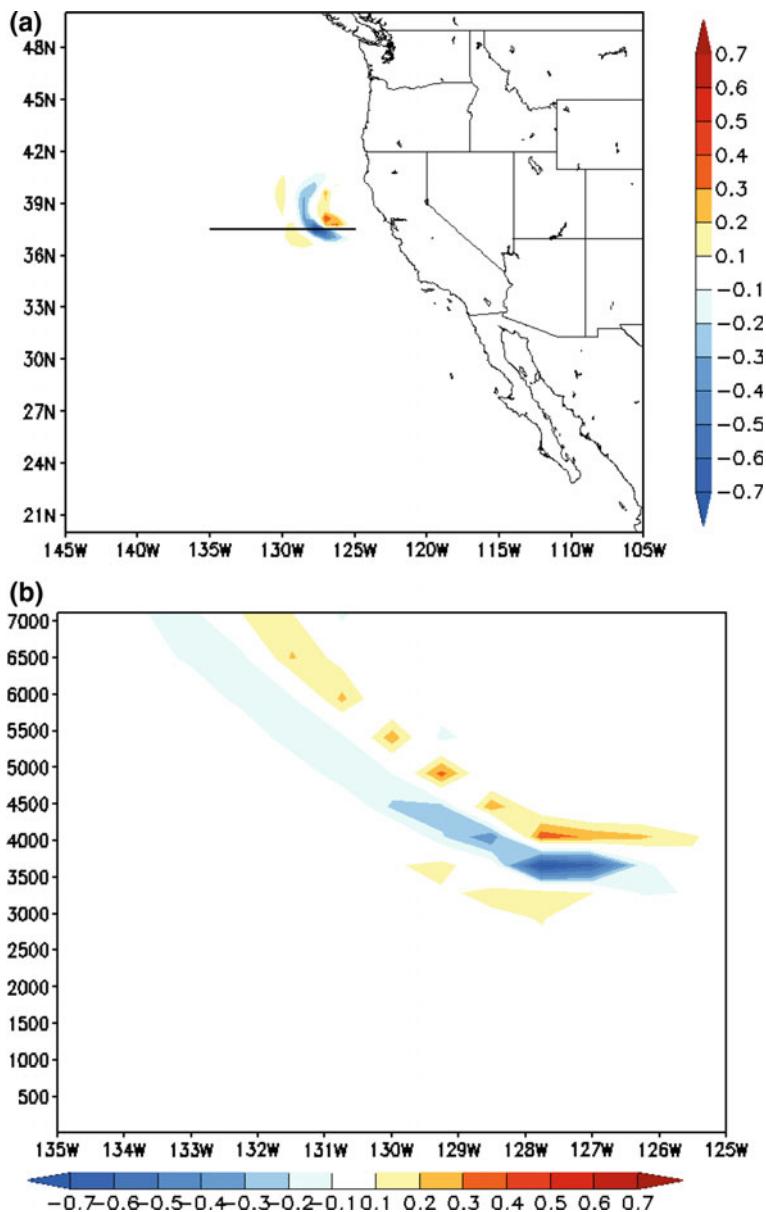
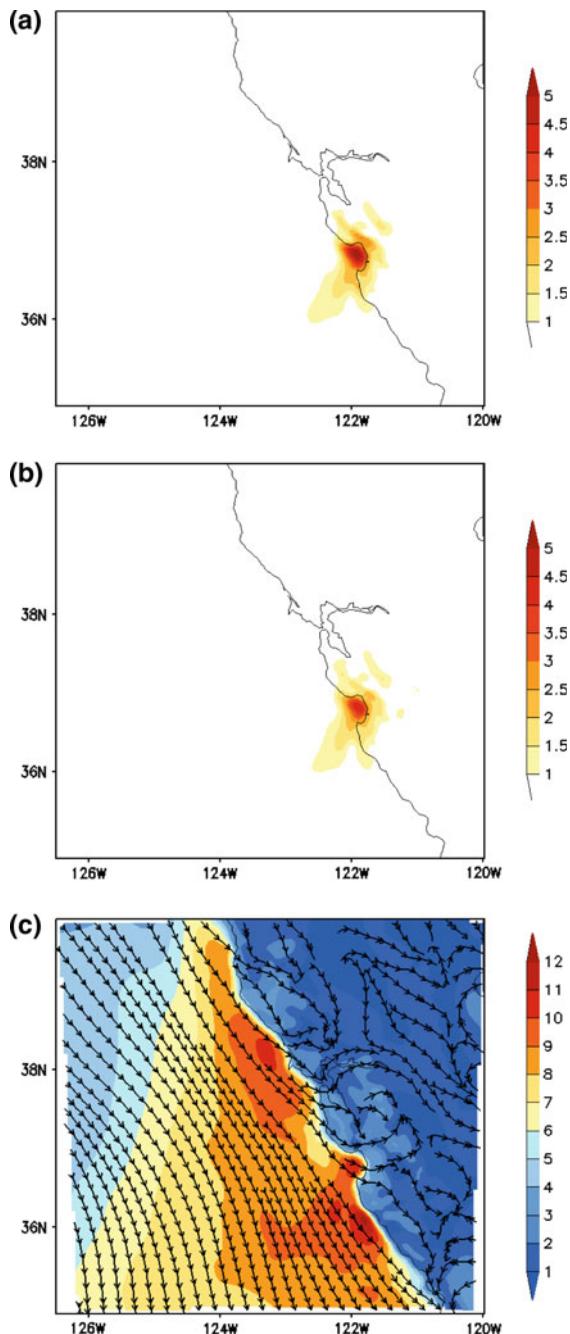


Fig. 5 Initial u perturbation **a** on nest 1 at model level 21 and **b** on a cross section indicated by the line in (a). The units are m s^{-1} and the perturbations are valid at 1200 UTC June 3 2015

After the 12 h backward integration, the largest sensitivities are located at roughly 650 mb and 300 Km from the location of the response function. The initial perturbation (Fig. 5) has the same structure as the sensitivity field and spans a larger volume

Fig. 6 Magnitude of the perturbation wind **a** forecast by the tangent linear model and **b** calculated from the difference in nonlinear forecasts run with and without the optimal perturbation. The full wind speed for the optimally perturbed forecast is shown in **c**. The units are m s^{-1} and all fields are at the lowest model level on nest 3 and valid at 0000 UTC June 4 2015



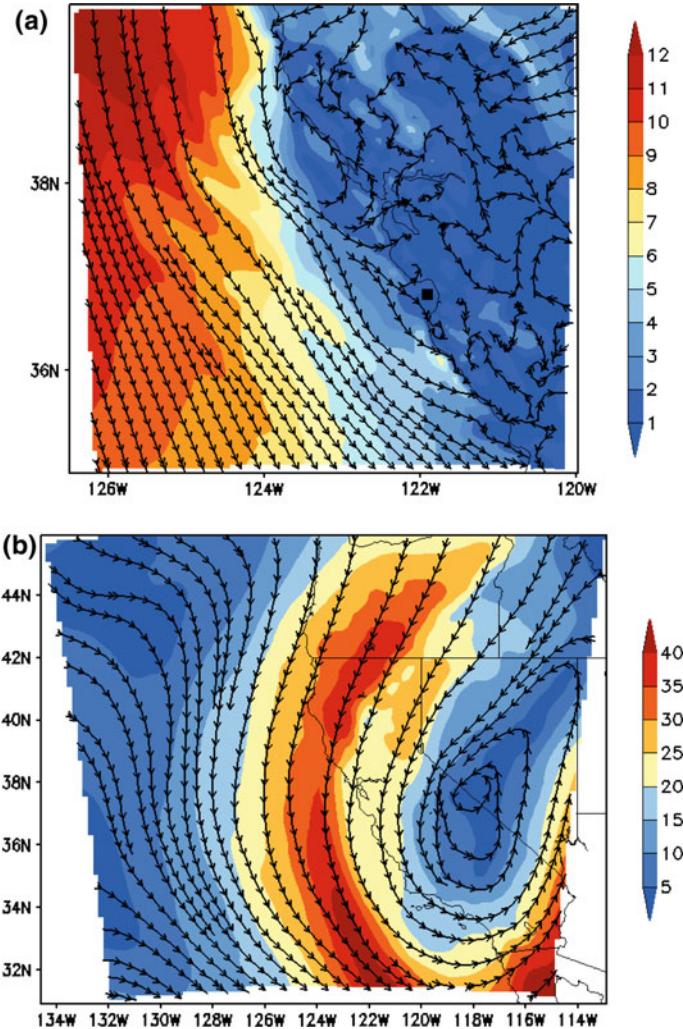


Fig. 7 Same as Fig. 3, but valid at 0000 UTC June 6 2015

than the original forcing input to the adjoint model. Using the process to choose s in Eq. 4 and set the maximum u perturbation magnitude to 1 m s^{-1} results in an initial v perturbation of roughly the same size and θ perturbations that are less the 0.5 K . The perturbation values below 700 mb are almost non existent away from the coast.

The magnitude of the tangent linear model forecast perturbation wind field at 12 h in the lowest model layer on nest 3 is shown in Fig. 6a. The structure of the perturbation is almost identical to the difference in forecast fields for nonlinear model runs with and without the optimal perturbation added to the initial state (Fig. 6b). The tangent linear forecast is about 1 m s^{-1} stronger over the response area compared to

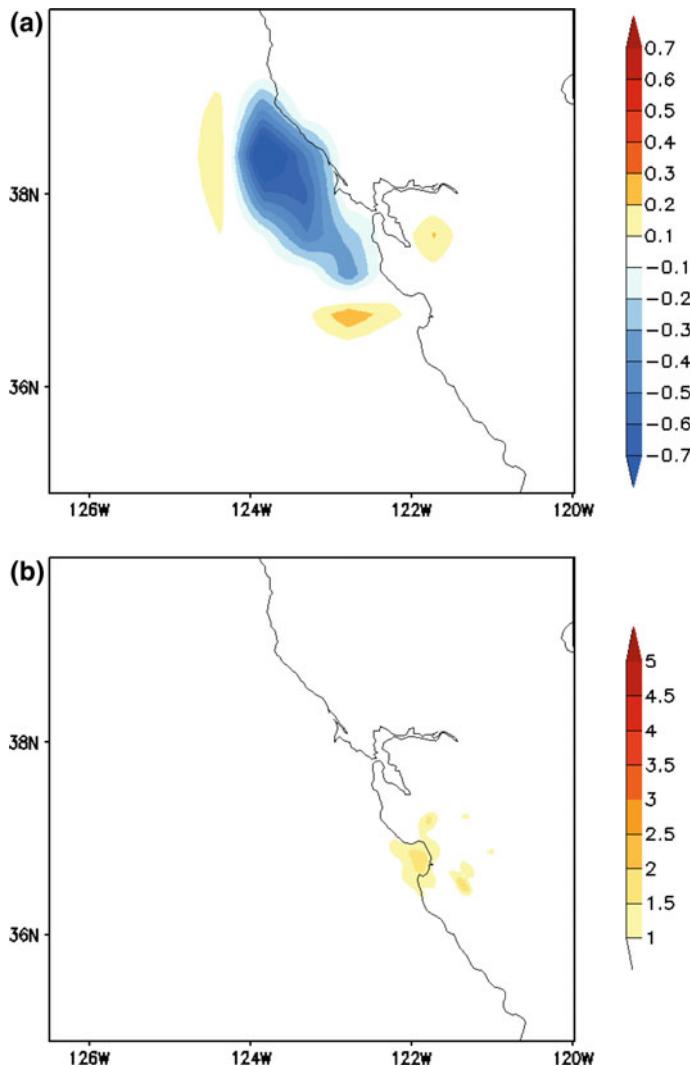


Fig. 8 Perturbation of u **a** at the initial time 1200 UTC June 5 2015 on model level 9 and **b** forecast by the tangent linear model at 0000 UTC June 6 2015 on the lowest model level. The units are ms^{-1}

the nonlinear model difference. The largest response is localized over Monterey Bay (the location of the response function). Comparing Figs. 3a and 6c, the addition of the optimal perturbation increases the lowest model level wind speed over Monterey Bay to be comparable to the maxima north of Point Reyes and south of Point Sur ($>10 \text{ m s}^{-1}$). It is also worth noting that during the first 10 h of the tangent linear forecast, there is no evidence of perturbation growth and the largest values remain above the boundary layer. Only in the last 2 h of the forecast period do the wind speed

perturbation values grow from less than 1 m s^{-1} to over 5 m s^{-1} and propagate into the boundary layer (not shown).

An additional experiment was conducted for a weaker surface flow regime. The model was initialized at 1200 UTC Jun 5 2015. In Fig. 7, the 12 h forecasts of wind speed and stream function at the lowest and 30th levels are again shown. A trough is present in the upper levels over southern California, this leads to a weaker temperature inversion at the top of the marine layer (not shown). Therefore, the surface winds close to the central California coast are not as strong as the previous experiment. The same response function was used and the adjoint model was again forced at the 12 h forecast time. In this case, the largest sensitivities remain confined to the boundary layer. Since the advective flow in the trajectory field is not as strong, the location of the largest initial optimal perturbations are not far removed from the response function area (Fig. 8a). Also, the response is not nearly as dramatic, the largest perturbation magnitudes at 12 h are $< 2 \text{ m s}^{-1}$ (Fig. 8b).

4 Summary

Nesting capabilities were added to the COAMPS tangent linear and adjoint models. The system was demonstrated by constructing optimal perturbations for coastal atmospheric flow. Small perturbations applied over relatively large areas in the upper levels of the atmosphere increased the surface wind speeds over Monterey Bay by 5 m s^{-1} 12 h later. This study primarily serves as a demonstration of the nested system. There are countless atmospheric phenomena where the relatively high resolution achievable with this system would be beneficial for both data assimilation and sensitivity studies. However, nonlinearity and discontinuity increases with resolution, especially as more physical processes are included in the model integrations. Therefore, the included processes and the integration times of the tangent linear and adjoint models must be considered. Nevertheless, the nesting capabilities added to this system are important tools for elucidating predictability aspects of meso and finer scale atmospheric flows.

Acknowledgements This research is supported by the Chief of Naval Research through the NRL Base Program, PE 0602435N. Computational resources from the Department of Defense's High Performance Computing Modernization Program were vital to this work.

References

- Amerault C, Zou X, Doyle J (2008) Tests of an adjoint mesoscale model with explicit moist physics on the cloud scale. *Mon Weather Rev* 136:2120–2132. doi:[10.1175/2007MWR2259.1](https://doi.org/10.1175/2007MWR2259.1)
- Ancell B, Mass C (2006) Structure, growth rates, and tangent linear accuracy of adjoint sensitivities with respect to horizontal and vertical resolution. *Mon Weather Rev* 134:2971–2988. doi:[10.1175/MWR3227.1](https://doi.org/10.1175/MWR3227.1)
- Beardsley R, Dorman C, Friehe C, Rosenfeld L, Winant C (1987) Local atmospheric forcing during CODE. Part 1: A description of the marine boundary layer and atmospheric condi-

- tions over a northern California upwelling region. *J Geophys Res* 92:1467–1488. doi:[10.1029/JC092iC02p01467](https://doi.org/10.1029/JC092iC02p01467)
- Birchfield GE (1960) Numerical prediction of hurricane movement with the use of a fine grid. *J Meteor* 17:404–414
- Daley R, Barker E (2001) NAVDAS: formulation and diagnostics. *Mon Weather Rev* 129:869–883. doi:[10.1175/1520-0493\(2001\)129<0869:NFAD>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0869:NFAD>2.0.CO;2)
- Dudhia J (1993) A nonhydrostatic version of the Penn State-NCAR mesoscale model: Validation tests and simulation of an Atlantic cyclone and cold front. *Mon Weather Rev* 121:1493–1513
- Errico R, Raeder K (1999) An examination of the accuracy of the linearization of a mesoscale model with moist physics. *Q J R Meteorol Soc* 125:169–195. doi:[10.1002/qj.49712555310](https://doi.org/10.1002/qj.49712555310)
- Haack T, Burk S, Dorman C, Rodgers D (2001) Supercritical flow interaction with the Cape Blanco–Cape Mendocino orographic complex. *Mon Weather Rev* 129:688–708. doi:[10.1175/1520-0493\(2001\)129<0688:SFIWTC>2.0.CO;2](https://doi.org/10.1175/1520-0493(2001)129<0688:SFIWTC>2.0.CO;2)
- Hodur R (1997) The Naval Research Laboratory's coupled ocean/atmosphere mesoscale prediction system (COAMPS). *Mon Weather Rev* 125:1414–1430. doi:[10.1175/1520-0493\(1997\)125<1414:TNRLSC>2.0.CO;2](https://doi.org/10.1175/1520-0493(1997)125<1414:TNRLSC>2.0.CO;2)
- Langland R (2005) Issues in targeted observing. *Q J R Meteorol Soc* 131:3409–3425
- Lewis J, Raeder K, Errico R (2001) Vapor flux associated with return flow over the Gulf of Mexico: a sensitivity study using adjoint modeling. *Tellus* 53A:74–93. doi:[10.3402/tellusa.v53i1.12177](https://doi.org/10.3402/tellusa.v53i1.12177)
- Ley GW, Elsberry RL (1976) Forecasts of Typhoon Irma using a nested-grid model. *Mon Weather Rev* 104:1154–1160
- Michalakes J, Chen S, Dudhia J, Hart L, Klemp J, Middlecoff J, Skamarock W (2001) Development of a next generation regional weather research model. In: Zwiefelhofer W, Krietz N (eds) *Developments in TeraComputing: proceedings of the ninth ECMWF workshop on the use of high performance computing in meteorology*. World Scientific
- Oortwijn J (1995) Perturbations that optimally trigger weather regimes. *J Atmos Sci* 52:3932–3944. doi:[10.1175/1520-0469\(1995\)052<3932:PTOTWR>2.0.CO;2](https://doi.org/10.1175/1520-0469(1995)052<3932:PTOTWR>2.0.CO;2)
- Pielke R, Cotton W, Walko R, Tremback C, Lyons W, Grasso L, Nicholis M, Moran M, Wesley D, Lee T, Copeland J (1992) A comprehensive meteorological modeling system—RAMS. *Meteorol Atmos Phys* 49:69–91
- Rabier F (2005) Overview of global data assimilation developments in numerical weather-prediction centres. *Q J R Meteorol Soc* 131:3215–3233
- Rabier F, Klinker E, Courtier P, Hollingsworth A (1996) Sensitivity of forecast errors to initial conditions. *Q J R Meteorol Soc* 122:121–150. doi:[10.1002/qj.49712252906](https://doi.org/10.1002/qj.49712252906)
- Samelson R (1992) Supercritical marine-layer flow along a smoothly varying coastline. *J Atmos Sci* 49:1571–1584. doi:[10.1175/1520-0469\(1992\)049<1571:SMLFAA>2.0.CO;2](https://doi.org/10.1175/1520-0469(1992)049<1571:SMLFAA>2.0.CO;2)
- Skamarock W (1989) Adaptive grid refinement for numerical weather prediction. *J Comput Phys* 80:27–60. doi:[10.1016/0021-9991\(89\)90089-2](https://doi.org/10.1016/0021-9991(89)90089-2)
- Skamarock W, Klemp J, Dudhia J (2001) Prototypes for the WRF (Weather Research and Forecasting) model. In: Ninth conference on mesoscale processes. American Meteorological Society, Fort Lauderdale, FL, USA
- Sun J, Crook N (1997) Dynamical and microphysical retrieval from doppler radar observations using a cloud model and its adjoint. Part I: Model development and simulated data experiments. *J Atmos Sci* 54:1642–1661. doi:[10.1175/1520-0469\(1997\)054<1642:DAMRFD>2.0.CO;2](https://doi.org/10.1175/1520-0469(1997)054<1642:DAMRFD>2.0.CO;2)
- Zou X, Vandenberghe F, Pondeca MD, Kuo YH (1997) Introduction to adjoint techniques and the MM5 adjoint modeling system. Technical Report NCAR/TN-435-STR, NCAR. doi:[10.1175/1520-0469\(1965\)022<0040:OFAIOT>2.0.CO;2](https://doi.org/10.1175/1520-0469(1965)022<0040:OFAIOT>2.0.CO;2)
- Zou X, Xiao Q (2000) Studies on the initialization and simulation of a mature hurricane using a variational bogus data assimilation scheme. *J Atmos Sci* 57:836–860
- Županski M, Županski D, Vukicevic T, Eis K, Haar TV (2005) CIRA/CSU four-dimensional variational data assimilation system. *Mon Weather Rev* 133:829–843. doi:[10.1175/MWR2891.1](https://doi.org/10.1175/MWR2891.1)

On the Impact of the Diabatic Component in the Forecast Sensitivity Observation Impact Diagnostics

Marta Janisková and Carla Cardinali

Abstract Over the years, a comprehensive set of the linearized physical parameterization schemes has been developed at ECMWF. These linearized schemes, operationally used in data assimilation, parametrize both the dry physical processes (vertical diffusion, gravity wave drag, shortwave and longwave radiation) and the moist processes (convection, large-scale condensation and clouds) consistently with the physical parametrization of the nonlinear model (though some simplifications are applied). In this work, the representation of the moist physical processes in the adjoint assimilation model is compared with the representation of humidity in the energy norm used to compute the forecast sensitivity to observations in the short-range forecasts. Forecast Sensitivity Observation Impact using the adjoint model with only dry processes (dry adjoint) but moist energy norm in the sensitivity gradient calculation is examined in contrast with the observation impact obtained when moist processes (moist adjoint) and dry energy norm are used. The performed study indicates that the use of the humidity term in the norm produces unrealistic humidity and temperature sensitivity gradients, which largely affect the observation forecast impact results.

1 Introduction

Nowadays sophisticated data assimilation schemes are used for exploiting information from irregularly distributed observations in order to provide initial conditions for a numerical weather prediction (NWP) model. One of them is the four-dimensional variational (4D-Var) data assimilation, which is the operational system at the European Centre for Medium-Range Weather Forecast (ECMWF) since November 1997 (Rabier et al. 2000). 4D-Var minimizes the distance between the model trajectory and the observations over a given time interval, using the adjoint equations of the

M. Janisková (✉) · C. Cardinali
European Centre for Medium-Range Weather Forecasts,
Shinfield Park, Reading RG2 9AX, UK
e-mail: marta.janiskova@ecmwf.int

model to compute the gradient of the cost function with respect to the model state at the beginning of the assimilation period. The mismatch between model solution and observations can remain large if the adiabatic adjoint model would only be used in the minimization. In addition, many satellite observations, such as radiances, rainfall and cloud measurements, cannot be directly assimilated with such overly simple adjoint model. Therefore representation of physical processes in the assimilating models is crucial. Initially, adjoint models used only very simple parametrization schemes, such as Buizza (1994), which aimed at removing very strong increments produced by the adiabatic adjoint models. Gradually, more complex, but still incomplete schemes were developed by Zou et al. (1993), Županski and Mesinger (1995), Janisková et al. (1999), Mahfouf (1999), Laroche et al. (2002), Mahfouf (2005). More comprehensive schemes, which can describe the whole set of physical processes and interactions between them, almost as in the non-linear model, with just a few simplifications and/or regularizations compared to the reference non-linear model, were implemented more recently (e.g. Janisková et al. 2002; Tompkins and Janisková 2004; Lopez and Moreau 2005; Janisková and Lopez 2013).

Using sophisticated data assimilation schemes, such as 4D-Var, also requires effective performance monitoring of such a complex system. A traditional tool for estimating data impact in a forecasting system is provided by Observing System Experiments (OSEs). These are usually performed by removing subsets of observations from the assimilating system and the resulting forecasts are compared against a control experiment that includes all observations (e.g. Bouttier and Kelly 2001; English et al. 2004; Kelly and Thépaut 2007; Bauer et al. 2014). Recently, new diagnostics in data assimilation and numerical weather prediction provides an assessment of each observation contribution to the analysis. For example, techniques have been derived to indicate which level of influence is given to observations and which one to the background during the assimilation procedure (Purser and Huang 1993; Cardinali et al. 2004; Chapnik et al. 2004; Cardinali 2015), thus allowing some tuning of the weights assigned in the assimilation system. To measure the observation contribution to the forecast quality, the adjoint methodology can also be used where the observation impact is evaluated with respect to a scalar function representing the short-range forecast error, see for example Baker and Daley (2000), Cardinali and Buizza (2004), Langland and Baker (2004), Xu et al. (2006), Zhu and Gelaro (2008), Cardinali (2009, 2015) or Lorenc and Marriot (2014).

An advantage of the adjoint-based observation sensitivity compared to OSE is that it measures the impact of observations when the entire observation dataset is present in the assimilation system. It provides the response of a single forecast metric to all perturbations of the observing system. However, this technique is influenced by simplified adjoint model used to carry the forecast error information backwards and therefore limited by the validity of the tangent-linear assumptions in a different way from OSE. Generally, experiments performed with adjoint technique for estimating the forecast sensitivity to observations, use a different level of complexity for simplified adjoint model. Some of them contain only a basic description of physical processes, mainly dry processes (Zhu and Gelaro 2008; Gelaro and Zhu 2009), while others use a comprehensive set of physical parametrizations describing both

moist and dry processes (Cardinali 2009, 2015). Another factor which can have a significant impact in adjoint-based method is the selection of the total energy (TE) norm used for the sensitivity gradient computations. Dry energy norm is used by Cardinali (2009, 2015), Daescu and Todling (2010), Zhu and Gelaro (2008), Gelaro and Zhu (2009), while Langland and Baker (2004), Lorenc and Marriot (2014) apply the moist TE norm for their gradient computations. Which norm is the most appropriate for these computations is a matter of permanent discussion.

At ECMWF, over the years an extensive set of linearized physical parametrizations (Janisková and Lopez 2013) has been developed for the global data assimilation system and sensitivity studies. It comprises dry parametrization schemes (radiation, vertical diffusion, orographic gravity wave drag and non-orographic gravity wave activity) and moist parametrizations (moist convection, large-scale condensation/precipitation). The current linearized physics package is therefore quite sophisticated and is believed to be the most comprehensive one currently used in operational global data assimilation. As mentioned above, the dry TE norm is used in the adjoint-based observation sensitivity studies at ECMWF (Cardinali 2009, 2015). Being in position of having a comprehensive description of physical processes in the adjoint model, different experiments have been performed to compare observation impacts obtained by using different sets of the physical processes in the adjoint assimilation model and different representation of energy norms. In this paper, the Forecast Sensitivity Observation Impact (FSOI) using the adjoint model with only dry processes (dry adjoint) but moist energy norm in the sensitivity gradient calculation is examined in contrast with FSOI obtained with moist processes (moist adjoint) and dry energy norm. This type of study provides information on the use of moist processes in the adjoint models and some information on the role of the moist component in the TE norm. In Sect. 2, the methodology of adjoint-based observation sensitivity and its relation with the energy norm and the simplified adjoint model is described. Details of the experimental framework and the results are provided in Sect. 3. Finally, conclusions are given in Sect. 4.

2 Forecast Sensitivity Impact

2.1 Method

The aim of 4D-Var assimilation is to find the optimal initial atmospheric state (the analysis, \mathbf{x}^a) for numerical weather forecast. Information on short-range model forecast (background, \mathbf{x}^b) and observations, \mathbf{y} (over a given time interval), are combined accordingly to their weights. Once the cost (objective) function, J , measuring the weighted misfit between the model trajectory and the observations has been defined, the gradient of the cost function with respect to the model state at the beginning of

assimilation period can be computed using the adjoint equations of the model. The analysis \mathbf{x}^a can be obtained by providing this gradient to an iterative minimization algorithm (Courtier et al 1998). The analyses weights are obtained by the Kalman gain matrix \mathbf{K} .

The forecast sensitivity equation with respect to the observations in the context of variational data assimilation has been derived by Baker and Daley (2000). The sensitivity of the objective function, J_E , with respect to the observations can be written using a simple derivative chain as:

$$\frac{\partial J_E}{\partial \mathbf{y}} = \frac{\partial J_E}{\partial \mathbf{x}^{a,b}} \frac{\partial \mathbf{x}^a}{\partial \mathbf{y}} \quad (1)$$

where $\partial J_E / \partial \mathbf{x}^{a,b}$ is the mean sensitivity of the forecast error with respect to the analysis and the background (second order gradients, see for example Errico 2007). As explained for instance by Cardinali et al. (2004) or Cardinali (2009), $\partial \mathbf{x}^a / \partial \mathbf{y}$ is the sensitivity of the analysis system with respect to observations, that is \mathbf{K}^T .

Once the forecast sensitivity is computed, FSOI, i.e. the variation δJ_E of the forecast error due to the assimilated observations can be found by applying the adjoint property for a linear operator as:

$$\begin{aligned} \delta J_E &= \left\langle \frac{\partial J_E}{\partial \mathbf{x}^{a,b}}, \delta \mathbf{x}^a \right\rangle = \left\langle \frac{\partial J_E}{\partial \mathbf{x}^{a,b}}, \mathbf{K}(\mathbf{y} - H[\mathbf{x}^b]) \right\rangle = \left\langle \mathbf{K}^T \frac{\partial J_E}{\partial \mathbf{x}^{a,b}}, \mathbf{y} - H[\mathbf{x}^b] \right\rangle = \left\langle \mathbf{K}^T \frac{\partial J_E}{\partial \mathbf{x}^{a,b}}, \delta \mathbf{y} \right\rangle \\ &= \left\langle \frac{\partial J_E}{\partial \mathbf{y}}, \delta \mathbf{y} \right\rangle \end{aligned} \quad (2)$$

where $\delta \mathbf{x}^a = \mathbf{x}^a - \mathbf{x}^b$ are the analysis increments, $\delta \mathbf{y} = \mathbf{y} - H[\mathbf{x}^b]$ is the innovation vector, H is the nonlinear observation operator (moving the background value to the observation location) and \mathbf{K} and \mathbf{K}^T are the gain matrix and its adjoint, respectively. According to Eq. 2, FSOI is then a function of the sensitivity gradient $\partial J_E / \partial \mathbf{x}^{a,b}$, the adjoint of the gain matrix, \mathbf{K}^T , and the innovation vector, i.e.

$$\delta J_E = f\left(\frac{\partial J_E}{\partial \mathbf{x}^{a,b}}, \mathbf{K}^T, \mathbf{y} - H[\mathbf{x}^b]\right) \quad (3)$$

For instance, at ECMWF, δJ_E is computed for a 12-h window. The second order sensitivity gradient $\partial J_E / \partial \mathbf{x}^{a,b}$ is valid at the starting time of the 4D-Var window, typically 09 and 21 UTC (Cardinali 2009). The variation of the forecast error due to a specific measurement can be summed up over time and space in different subsets to compute the total contribution of the different components of the observing system towards reduction of the forecast error.

The role of the energy norm objective function and the simplified adjoint model in FSOI computation will be assessed in this study. The description of the energy norm is provided in the following subsection.

2.2 Energy Norm

As explained in Sect. 1 (Introduction), the total energy (TE) norm in sensitivity studies is used either in the dry form (Cardinali 2009; Daescu and Todling 2010) or with the moist contribution (Langland and Baker 2004). The only difference between the dry and moist norms in these studies is in the additional term which explicitly measures specific humidity q . The TE norm in the moist form can be expressed in a continuous formulation as follows:

$$J_E = \frac{1}{2} \int_0^1 \int_{\Sigma} \left(\nabla \Delta^{-1} \xi_x \cdot \nabla \Delta^{-1} \xi_x + \nabla \Delta^{-1} D_x \cdot \nabla \Delta^{-1} D_x + \frac{c_p}{T_r} T_x T_x + w_q \frac{L_c^2}{c_p T_r} q_x q_x \right) d\Sigma \left(\frac{\partial p}{\partial \eta} \right) d\eta + \frac{1}{2} \int_{\Sigma} R_d T_r P_r \ln \pi_x \cdot \ln \pi_x d\Sigma \quad (4)$$

where c_p is the specific heat of dry air at constant pressure, R_d is the dry constant of dry air, L_c is the latent heat of condensation, T_r is the reference temperature and P_r is the reference pressure (e.g. $T_r = 350$ K and $P_r = 1000$ hPa at ECMWF). The TE norm (Eq. 4) has contributions from vorticity ξ_x , divergence D_x , temperature T_x , specific humidity q_x with certain weight w_q and logarithm of surface pressure $\ln \pi_x$ of the model state \mathbf{x} . In the case of the dry TE norm, the term with specific humidity is missing, i.e. $w_q = 0$.

Questions were raised on what is more appropriate—to use dry or moist energy norm. A lot of studies using the moist norm refer to the paper of Ehrendorfer et al. (1999) about singular-vector perturbation growth in a primitive equation model with moist physics. They made experiments with both dry and moist norm based on the fact that a strict theoretical basis for the humidity component does not exist. Unlike for dry TE norm and dry model, it is not evident that the moist TE norm will be conserved if other physical processes than condensation occur.

Using the moist TE norm requires a definition of the weight w_q . Several studies (such as Buizza et al. 1996; Mahfouf et al. 1996; Ehrendorfer et al. 1999; Barkmeijer et al. 2001) used the moist TE norm with $w_q = 1$ based on condensation physics and with the aim to ensure significant contribution to the norm in the singular vector computation by all components of the state vector both at initial and final time. Barkmeijer et al. (2001) also experimented with q weight derived from background-error statistics and Ehrendorfer et al. (1999) made some investigations by applying $w_q = 0.1$ or $w_q = 10$. The different weights would lead to different contributions of moisture and therefore qualitatively different results. Because of that, for instance, Ehrendorfer et al. (1999) concluded that further studies are required in order to understand how to specify the appropriate weight and the role of moist physics for a proper accounting for moist processes in the growth of perturbations (e.g. fast growing components of the analysis error).

2.3 Simplified Adjoint Model

In the adjoint-based technique, a simplified adjoint model is usually used to carry the forecast error information backwards. This model should have a certain level of reality, i.e. to be comprehensive enough to ensure that the observations are given a dynamically realistic, as well as statistically likely response in the analysis. However, it is important to achieve some trade-off between reality and linearity of the model since any adjoint-based technique is restricted by the tangent-linear (TL) assumption and its validity. Therefore one must be very careful with the non-linear nature of physical processes, especially in the presence of thresholds, which can affect the range of validity of the TL approximation.

The better the tangent-linear approximation, the more realistic and useful the sensitivity patterns. Results obtained through the adjoint integration when using a too simplified adjoint model with large inaccuracies or adjoint models without a proper treatment of nonlinearities and discontinuities can be incorrect. Most of the adjoint sensitivity experiments (e.g. Baker and Daley 2000; Langland and Baker 2004; Xu et al. 2006; Zhu and Gelaro 2008) are performed with simplified adjoint models that only describe dry processes with different levels of complexity. Adjoint models accounting for both dry and moist processes were only used by Cardinali (2009, 2015) and to some extent by Lorenc and Marriot (2014).

At ECMWF, the current set of physical parametrizations used in the linearized model describes both dry and moist processes: vertical diffusion, subgrid-scale orographic effects, radiation (shortwave and longwave), non-orographic gravity wave activity (not yet used in this study), clouds with large-scale condensation and convection as described by Janisková and Lopez (2013). Therefore, in the context of operational global data assimilation, this linearized physics package is quite sophisticated and comprehensive.

Validation studies of the ECMWF linearized model clearly demonstrate the impact and the importance of including physical processes for the validity of the tangent-linear approximation. For this validation, the accuracy of the linearization is studied with respect to pairs of non-linear simulations. The difference between two non-linear integrations (one starting from a background field and the other one starting from analysis) of the full nonlinear model is used as the standard reference to which the TL integrations from the analysis increments are compared. For a quantitative evaluation of the impact of linearized schemes, their relative importance is determined by using mean absolute errors between tangent-linear and non-linear integrations. The absolute mean error of the TL model without physics is usually taken as a reference for the comparisons. Errors and improvements relative to the reference can then be computed. Validity tests of the TL approximations are usually performed over the time period and at the resolution at which adjoint models will be applied in practice: resolution and time length of 4D-Var inner-loop integration (e.g. 12 h, T255¹ and 91 vertical levels at ECMWF) or longer time periods for singular vectors and sensitivity applications (e.g. 24 h at ECMWF).

¹T255 corresponding approximately to 80 km.

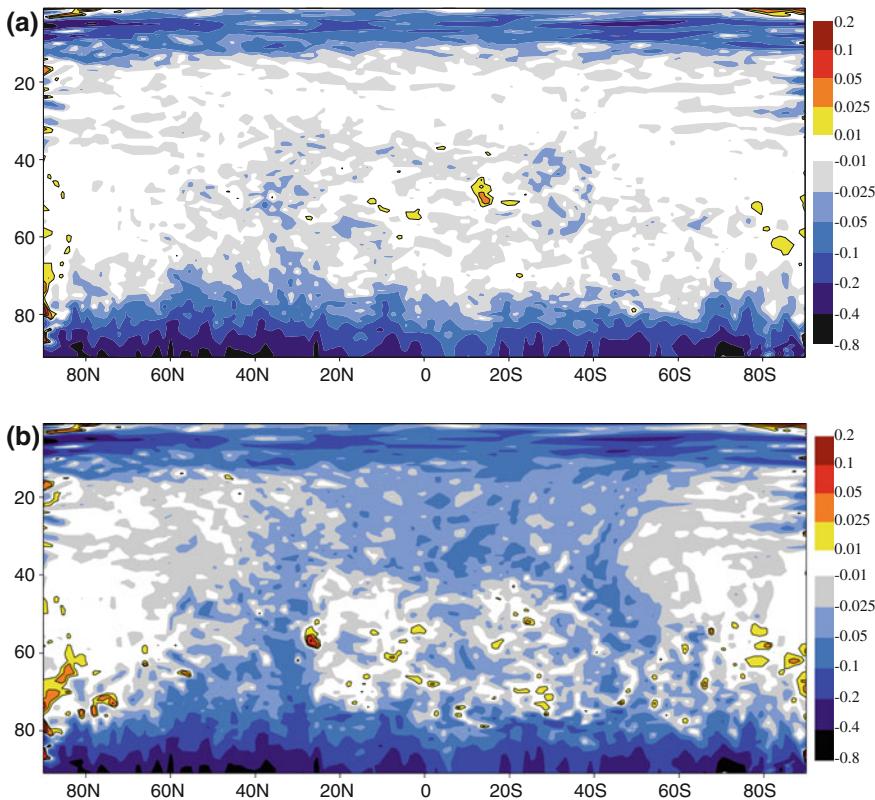


Fig. 1 Zonal mean impact of the different ECMWF linearized parametrization schemes on the evaluation of temperature increments. Results are presented as the error differences (in terms of fit to the non-linear model with full physics) between the TL model with physical parametrization schemes (including **a** dry processes alone—vertical diffusion, gravity wave drag and radiation or **b** in combination with moist processes—convection and cloud with large-scale condensation) and the purely adiabatic TL model

Examples of results from the TL approximation assessment using the ECMWF linearized physics are shown in Fig. 1. The impact of the different physical processes on the TL evolution of temperature increments is presented in Fig. 1 as zonal mean cross-sections of the error difference (in terms of fit to the nonlinear model with full physics) between the TL model including different parametrization schemes and the adiabatic TL model. Negative values are associated with an improvement of the model using the parametrization schemes with respect to the adiabatic TL model, since they correspond to a reduction of the errors. The improvement is observed over most of the atmosphere, and is maximum in the lower troposphere for two sets of parametrization schemes, one describing dry processes only (vertical diffusion, gravity wave drag and radiation; Fig. 1a) and the other one with moist processes also included (clouds with large-scale condensation and convection; Fig. 1b). Results also clearly show that taking into account moist processes lead to additional significant

improvement which is, however, not only coming from these schemes, but also from cloud-radiation interactions. The global relative improvement of the TL approximation for temperature coming from including physical parametrization schemes into the linearized model compared to the purely adiabatic TL model is $\sim 12\%$ for dry parametrization schemes alone and close to 18 % when combined with the moist schemes (Janisková and Lopez 2013). For specific humidity, the relative improvement becomes even larger when including moist processes (improvement of $\sim 20\%$) on top of the dry ones (improvement of $\sim 10\%$). Thus the TL model with all physical processes included performs remarkably better than its dry version. The representation of moist processes in the adjoint model not only provides a better description of the time evolution of the model state during the assimilation procedure and sensitivity calculations, but also allows the assimilation of observations sensitive to precipitation or clouds.

3 Experiments

Several experiments with the ECMWF assimilation system have been performed in order to study the impact of using different representations of the objective function and different representations of physical processes in the adjoint model in the FSOI computation. Which norm is the most appropriate for this type of computations is a matter of permanent discussion in the scientific community dealing with sensitivity studies. Thanks to the comprehensive description of physical processes in the adjoint model of ECMWF (Janisková and Lopez 2013), experiments with different combinations of norm definition and moist physics directly included in adjoint model could be performed.

3.1 Experimental Setup

The experiments have been run for the period of 25 August–10 September 2010 using the ECMWF 4D-Var system at that time operational resolution, i.e. 91 levels in vertical (L91) combined with the horizontal resolution of T1279² for the standard forecast model run and a much lower resolution of T159/T255³ for the minimization in assimilation computation. Although moist processes in the adjoint model allows the assimilation of observations related to clouds and precipitation, these observations have not been assimilated in order to unify observation usage among all performed experiments.

Adjoint sensitivity computations have then been performed at the resolution T255L91. Sensitivity calculations have been done using:

²T1279 corresponding approximately to 16 km.

³T159/T255 corresponding approximately to 130/80 km, respectively.

- (1) dry parametrization schemes alone: vertical diffusion, gravity wave drag and radiation;
- (2) moist parametrization schemes: convection and cloud with large-scale condensation in combination with dry processes.

These two adjoint model versions have been combined with either the dry or the moist TE norm as described by Eq. 4. In the case of the moist norm, different weights for the moisture contribution (w_q), such as 1 (results not presented), 0.5 or 0.1, have been used. Most results presented here have been obtained with $w_q = 0.5$.

The results from experiments using the following combinations of adjoint model versions and norm specifications will be shown:

- **dryAD_dryN**: dry processes in adjoint (AD) model and dry TE norm;
- **dryAD_moistN_0.5**: dry processes and moist TE norm with the weight for moist contribution equal to 0.5;
- **moistAD_dryN**: moist processes in combination with the dry ones in AD model and dry TE norm;
- **moistAD_moistN_0.5**: moist and dry processes in AD model combined with the moist norm using $w_q = 0.5$;
- **moistAD_moistN_0.1**: moist and dry processes in AD model combined with the moist norm using $w_q = 0.1$.

3.2 Sensitivity Gradient

Several results from the combination of either dry or moist adjoint model with either dry or moist energy norm in the sensitivity gradient calculations are presented here.

Figures 2, 4, 5, and 7 display the global horizontal distribution of the second order sensitivity gradient (SG) for the situation on 28 August 2010 at 21:00 UTC only. Similar sensitivity structures have been observed for all the other days (not shown). Zonal mean and vertical profiles of sensitivity gradient averaged over the whole test period (i.e. 25 August 2010–10 September 2010) are shown in Figs. 6, 8 and 9, respectively.

Figure 2 displays the specific humidity sensitivity gradient at the lowest model level for four different combinations of the physical processes in adjoint model and TE norms: **dryAD_dryN** (Fig. 2a), **dryAD_moistN_0.5** (Fig. 2b), **moistAD_dryN** (Fig. 2c) and **moistAD_moistN_0.5** (Fig. 2d). When using dry physical processes and dry TE norm, the sensitivity to specific humidity is quite small and mainly localized in areas of intense dynamical activity. Adding moist processes in the adjoint model (Fig. 2c) brings additional structures in the sensitivity, especially in areas of condensation and convective development. When using the moist norm in combination with the dry adjoint (Fig. 2b), no new structures appear, and on the contrary, a lot of structures are masked by large-scale positive sensitivities. Using all processes and moist norms (Fig. 2d) leads to overall enhanced sensitivities, with a clear preva-

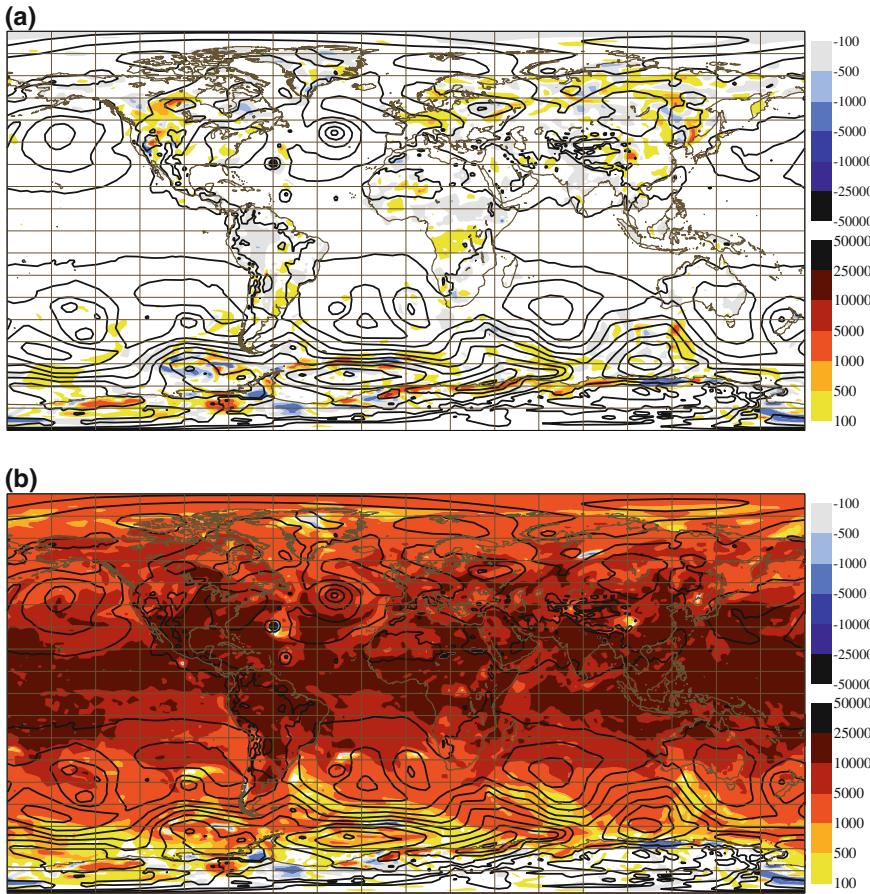


Fig. 2 Specific humidity sensitivity gradient ($\text{J kg}^{-1}/(\text{g kg}^{-1})$) at the lowest model level for the situation on 28 August 2010 at 21:00 UTC. The results are presented for experiments with dry parametrization schemes (i.e. vertical diffusion, gravity wave drag and radiation) included in the adjoint model using **a** dry or **b** moist norm, and for experiments with moist processes also added using **c** dry or **d** moist norm. Sensitivities are shown with colour shading. *Black isolines* represent mean-sea-level pressure (hPa)

lence of positive values over the globe. However, there are no additional sensitivity patterns when compared with **moistAD_dryN** (Fig. 2c).

The described behaviour of the forecast sensitivity with respect to specific humidity is more obvious from the zoom over the tropical cyclone in the Atlantic Ocean (Fig. 3). Missing moist physical processes clearly lead to underdetermined structures around the cyclone. Using the moist TE norm does not enhance the humidity structure (Fig. 3b). When moist processes are added in the adjoint model (Fig. 3c), a lot of sensitivity structures related to condensation and convection in the area of cyclone development are observed, as expected. When adding the moist norm, the humidity

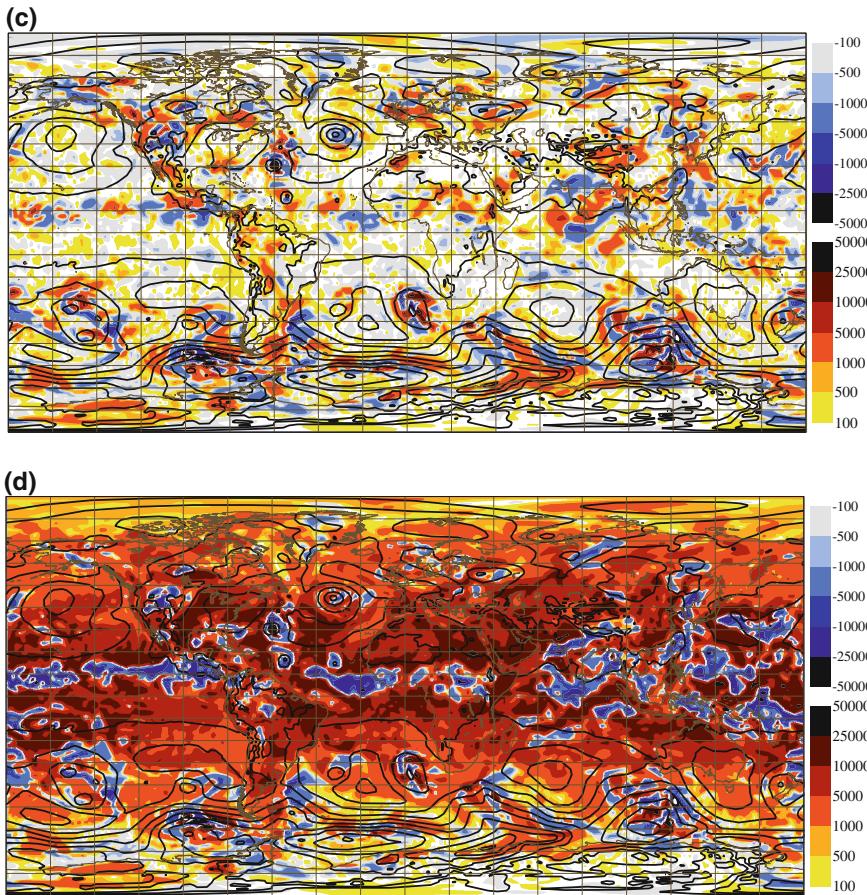


Fig. 2 (continued)

sensitivity pattern around the tropical cyclone remains very similar, but it is somehow embedded in a constant and larger humidity sensitivity background. This comparison indicates that only when moist processes are represented in the adjoint model the sensitivity gradients realistically depict the expected physical structure pattern, for example around tropical cyclones, whilst the dry adjoint model is not sufficient to correctly describe it. Moreover, adding the moisture term in the TE norm, either in combination of the dry or the moist adjoint model, provides a larger but somehow physically meaningless humidity pattern spread everywhere.

The temperature sensitivity gradient displayed in Fig. 4 suggests that the impact of moist processes used in the adjoint model on the sensitivity structures is less dramatic (Fig. 4c) than for humidity. In fact, larger sensitivity patterns already appear in the SG with the dry adjoint model and the dry TE norm (Fig. 4a). On top of the enhanced patterns around tropical cyclones, few additional ones associated to con-

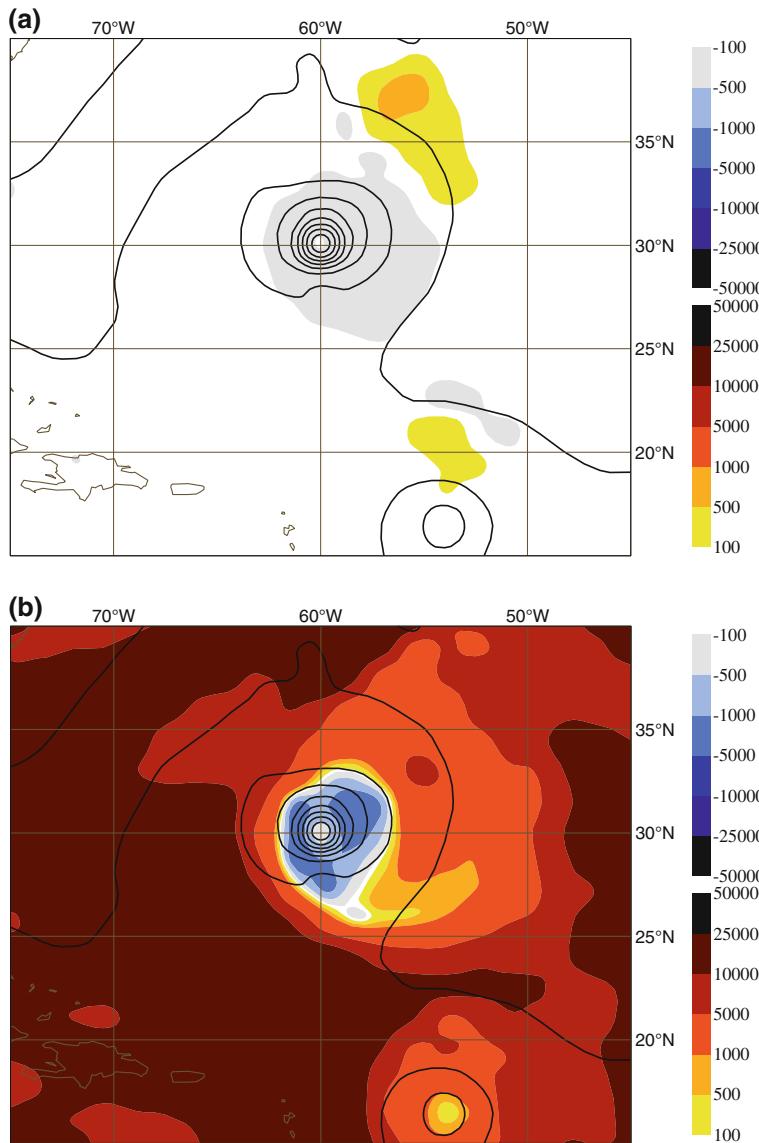


Fig. 3 Same as Fig. 2, but for the areas around tropical cyclones over the Atlantic ocean

vective and condensation activities emerge. In general, in **moistAD_dryN**, both the specific humidity and the temperature structures are quite consistent. Using the moist norm has a similar impact on temperature as observed for specific humidity (i.e. positive sensitivities prevailing), though a lot of structures, which already appeared when using the dry TE norm, are preserved in contrast with the specific humidity sensitiv-

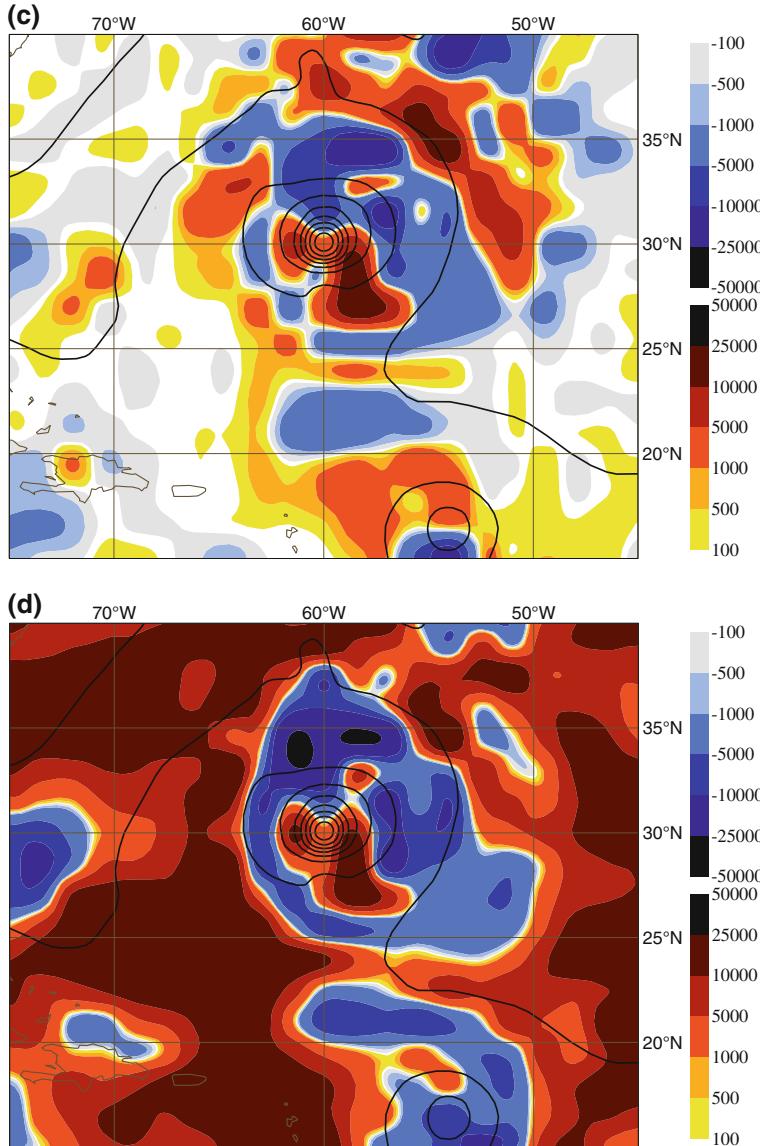


Fig. 3 (continued)

ity gradient (Fig. 4b). In the areas with larger specific humidity when the moist TE norm is used, the sensitivities are enhanced significantly more in the case of moist adjoint model (Fig. 4d).

In the case of sensitivity to surface pressure, when using moist TE norms (Fig. 5b, d), significantly enhanced sensitivities are found in the tropics, especially

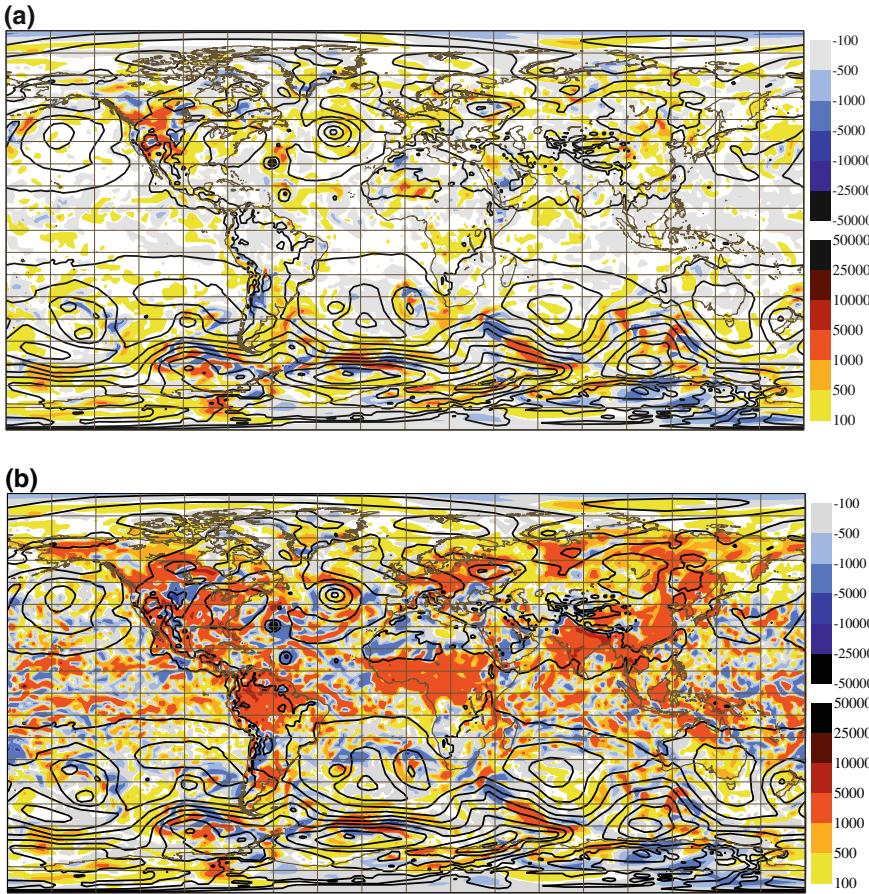


Fig. 4 Same as Fig. 2, but for temperature sensitivity gradient ($\text{J kg}^{-1}/\text{K}$) at the lowest model level

around the whole Inter Tropical Convergence Zone (ITCZ). As indicated by the isobars (black lines), the surface pressure there is not changing significantly and therefore one would assume a small overall forecast error sensitivities with some slightly more pronounced variations in convective regions (as seen in Fig. 5c for **moistAD_dryN**). However, when using the moist norm, wherever the humidity is large the sensitivity to surface pressure is large as well.

Overall, sensitivities with respect to different variables suggest two main pattern differences between the experiments. When the moist physics is used in the adjoint model, new structures emerge on top of those which already appeared when the dry adjoint model is used. Using the moist norm rather than the dry one leads to globally larger sensitivity patterns, but inconsistent with the physical processes observed.

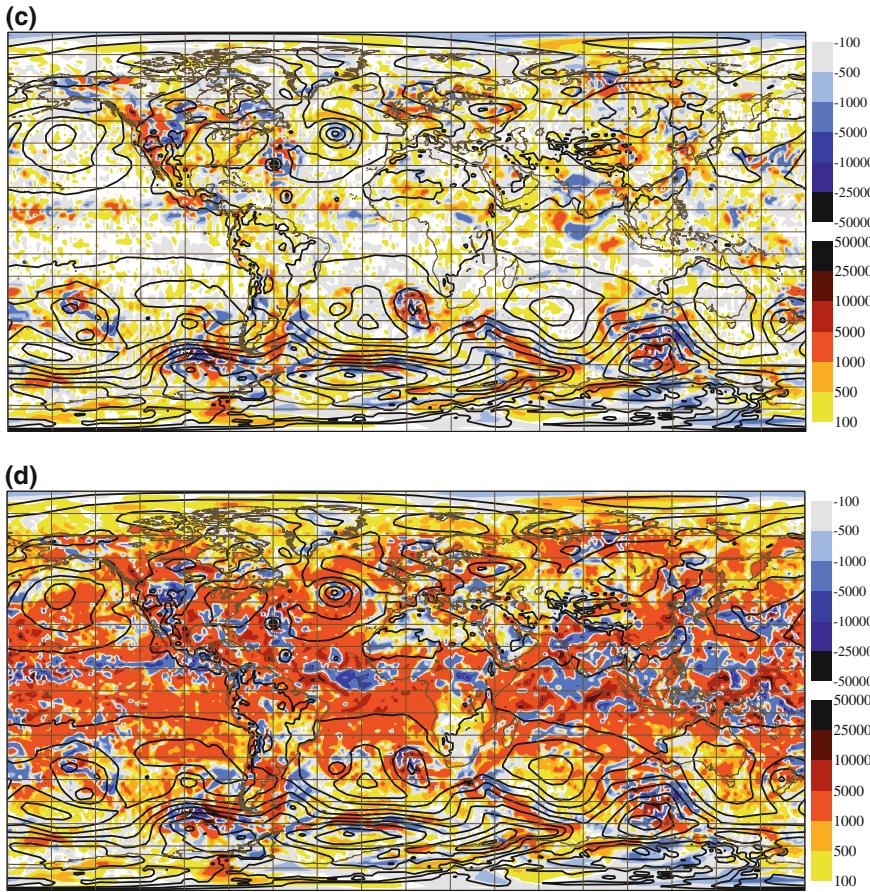


Fig. 4 (continued)

Many of the results described for the horizontal structures are confirmed when assessing zonal means (Fig. 6) and vertical profiles (Figs. 8 and 9) of the set averaged over the whole test period.

The zonal mean SG with respect to specific humidity confirms that the sensitivity is very small and located in the mid-low troposphere when using only dry adjoint model and dry norm (Fig. 6a). Including moist processes in the adjoint model leads to the appearance of much more structures in the boundary layer (Fig. 6e) and also to substantial changes in the sensitive regions of Fig. 6a. When the moist norm is used, the zonal mean of sensitivity reminds more of the structure of the humidity background error (as will be illustrated later on the vertical profiles of sensitivity), just modulated by the amount of the available moisture, i.e. decreasing from tropics towards the poles and from the low to the middle troposphere (Fig. 6c). The moist TE norm also leads to extensive, mainly positive temperature SGs in the zonal mean

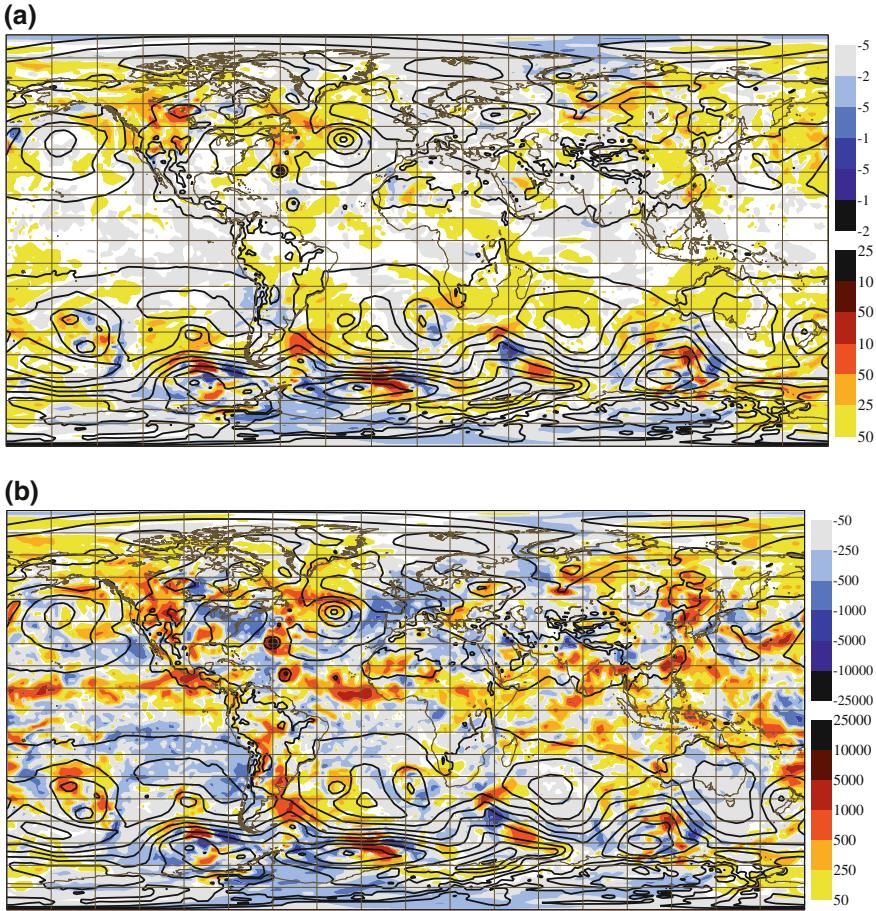


Fig. 5 Same as Fig. 2, but for the logarithm of surface pressure, p_s , sensitivity gradient ($\text{J kg}^{-1}/\text{Pa}$.)

(Fig. 6b, d, f). This feature is more pronounced when the moist TE norm is combined with the moist adjoint model (not shown).

Generally, experiments with the moist TE norm indicate that the weight $w_q = 0.5$ of the moisture term is too big. When the experiments were performed with $w_q = 1$ (not shown here) as often used in different studies, the described predominance of positive sensitivities is even more striking. Decreasing the weight for moisture term will lead to generally lower sensitivities. This is illustrated in Fig. 7, which compares specific humidity (Fig. 7a, c) and temperature (Fig. 7b, d) sensitivity gradient using $w_q = 0.5$ (Fig. 7a, b) and $w_q = 0.1$ (Fig. 7c, d). By reducing the weight not only sensitivities are reduced, but also some negative sensitivities emerge.

For a quantitative comparison of the different experiments performed in this study, vertical profiles of mean SGs with respect to specific humidity, q , and tem-

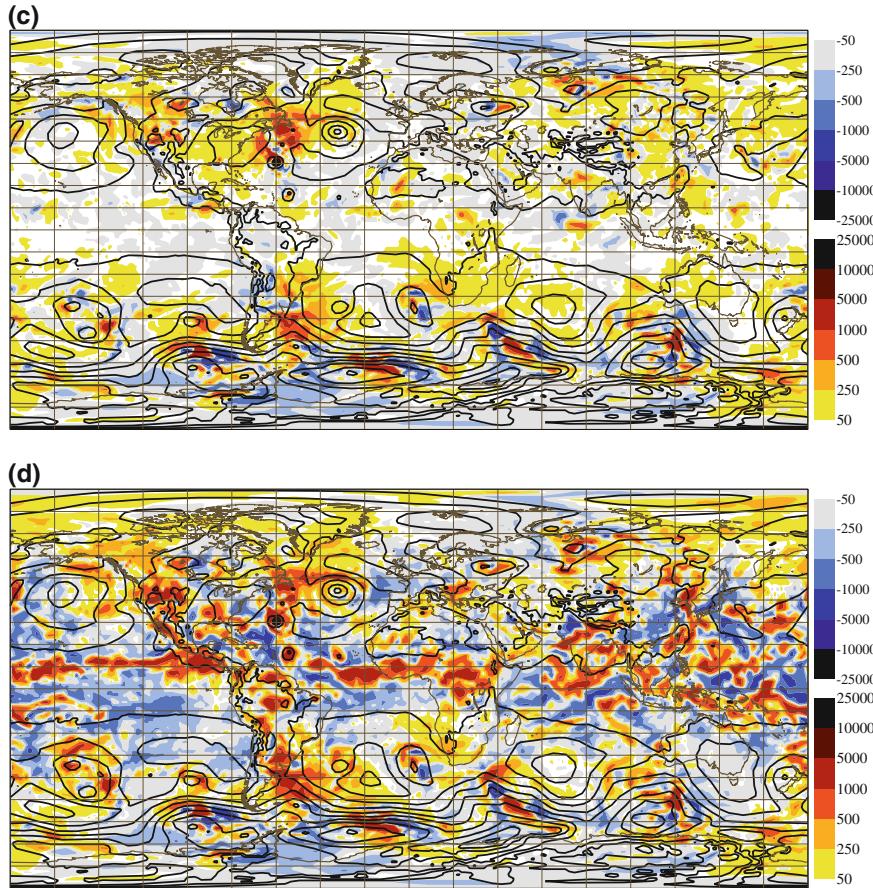


Fig. 5 (continued)

perature, T , averaged over the whole period together with their standard deviations are presented in Figs. 8 and 9, respectively. In principle, one would expect that SG averaged over a long period of time and over the whole globe should be unbiased, i.e. positive sensitivities should not prevail over negative ones or vice versa. For specific humidity (Fig. 8) using the dry TE norm, combined with either dry or moist processes in the adjoint model, leads to very small mean sensitivity values (Fig. 8a). The standard deviations (Fig. 8b) are also quite small (values are even smaller when using only dry processes given the generally very small sensitivities to specific humidity). On the contrary, when the moist TE norm is used, SGs are significantly positively biased and the standard deviations are also large. Definitely, the standard deviation shape is similar to the typical standard deviation of the ECMWF background errors for specific humidity (see Fig. 8c). Decreasing the weighting factor for moisture contribution in the moist TE norm from 0.5 to 0.1 leads to smaller bias

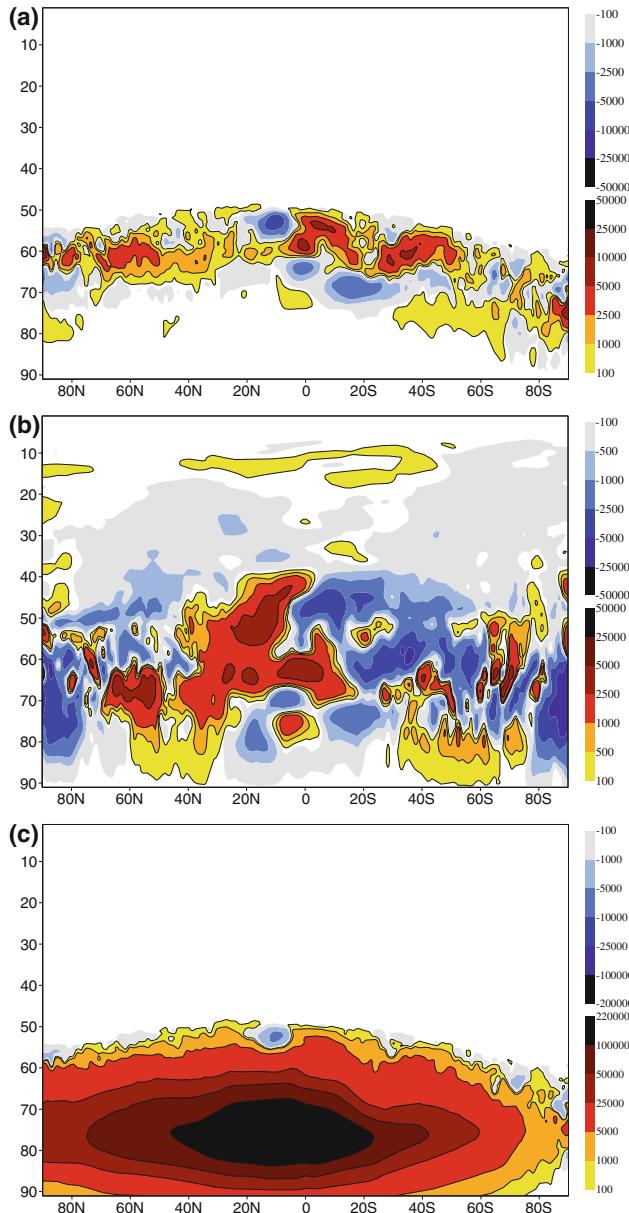


Fig. 6 Zonal mean of (a, c, e) specific humidity sensitivity gradient ($\text{J kg}^{-1}/(\text{g kg}^{-1})$) and (b, d, f) temperature sensitivity gradient ($\text{J kg}^{-1}/\text{K}$) for the whole test period (25 August 2010–10 September 2010). The results are presented for experiments with dry parametrization schemes (i.e. vertical diffusion, gravity wave drag and radiation) included in the adjoint model using **a**, **b** dry or **c**, **d** moist norm, and (e, f) for experiments with moist processes also added and using dry norm. Sensitivities are shown with colour shading

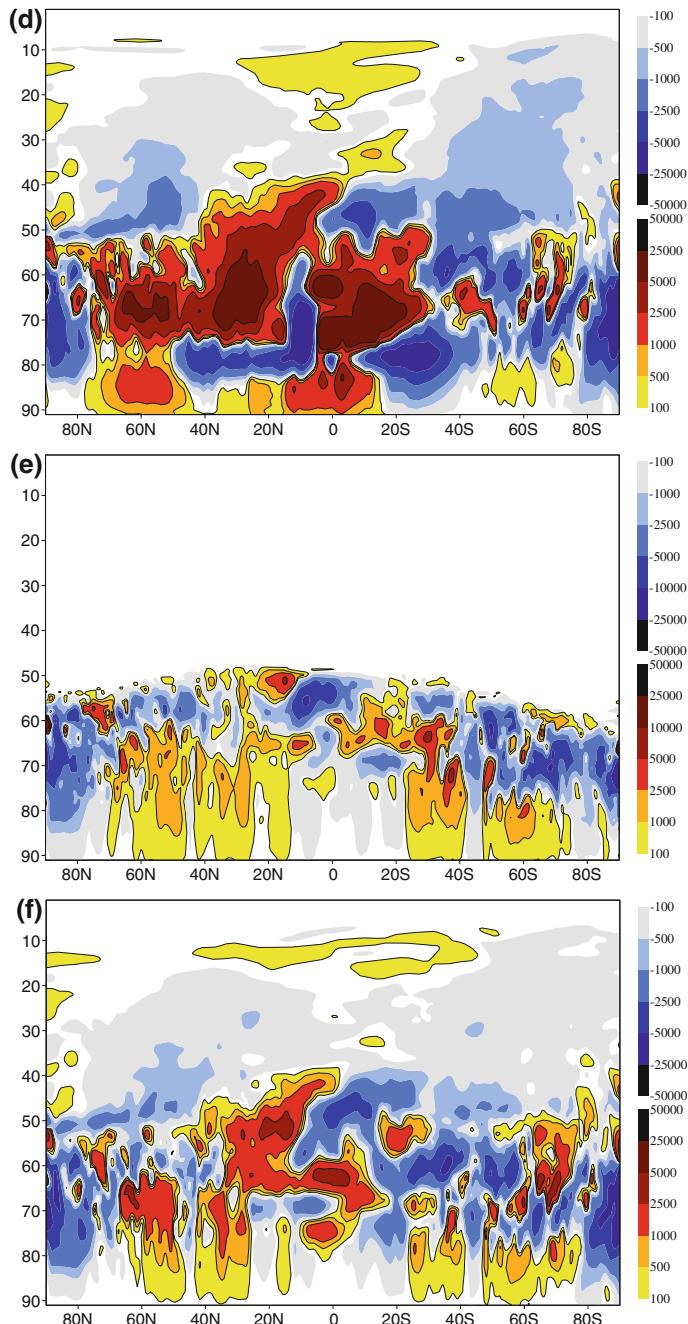


Fig. 6 (continued)

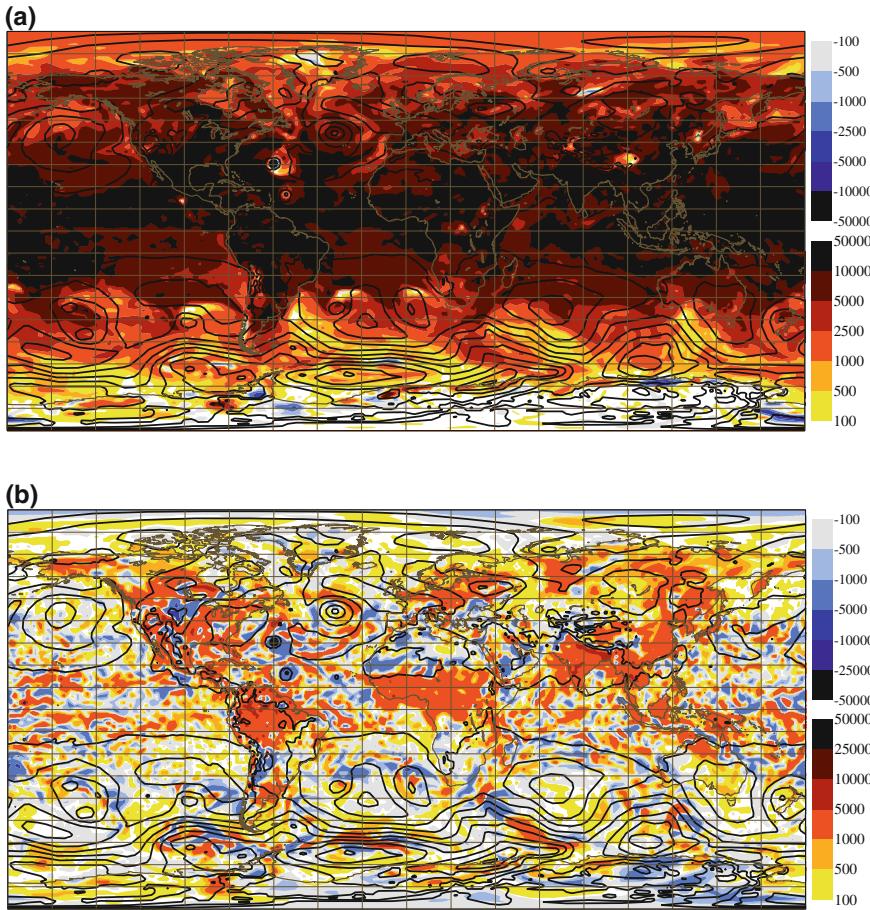


Fig. 7 **a, c** specific humidity ($\text{J kg}^{-1}/(\text{g kg}^{-1})$) and **b, d** temperature ($\text{J kg}^{-1}/\text{K}$) sensitivity gradients at the lowest model level for the situation on 28 August 2010 at 21:00 UTC. The results are presented for experiments with dry adjoint model and using moist norm with weighting factor of **a, b** 0.5 or **c, d** 0.1. Black isolines represent mean-sea-level pressure (hPa)

and standard deviation (Fig. 8a, b dotted line). However, even when the moisture contribution in the TE norm is significantly decreased, the sensitivity with respect to specific humidity remains biased. Similarly for the sensitivity to temperature (Fig. 9), mean values and standard deviations are close to zero when using the dry TE norm. The least biased SGs are determined by the moist adjoint and the dry TE norm. Interestingly when the moist TE norm is combined with the moist adjoint, the mean SGs with respect to temperature are the most biased. To demonstrate that sensitivities are indeed biased when the moist TE norm is used, an experiment has been done in which the mean humidity contribution was extracted from specific humidity SG

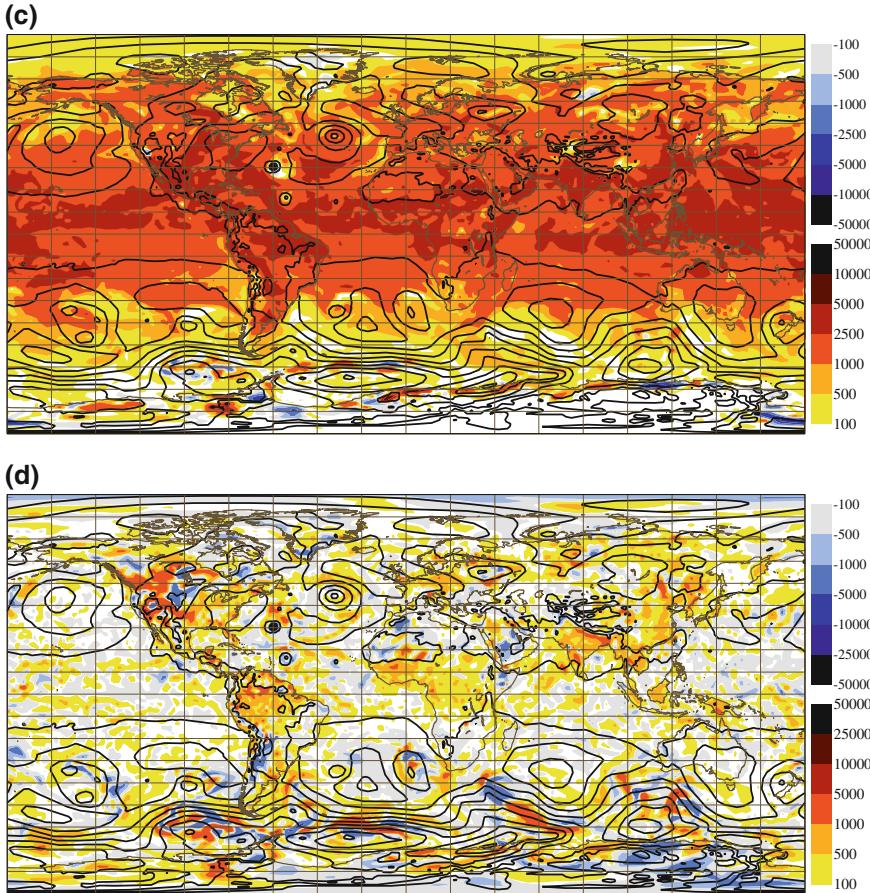


Fig. 7 (continued)

(not shown). Doing that, the overrepresentation of the positive structures disappear, though a small positive bias remains.

In conclusions, results indicate that the inclusion of moist processes through the adjoint model compared to accounting for the moisture through the energy norm leads to substantial differences in the intensity and the structure of the perturbations. The largest differences are obviously observed for the SGs with respect to specific humidity. While including moisture information in the adjoint model generate new structures on top of the already generated by the dry adjoint model, the usage of the moist TE norm usually globally enhances the sensitivity. Decreasing (respectively increasing) the weight of moist contribution in the norm (i.e. the weighting factor w_q in Eq. 4) will also decrease (respectively increase) the intensity of the sensitivity. Thus the different weights lead to qualitatively different results.

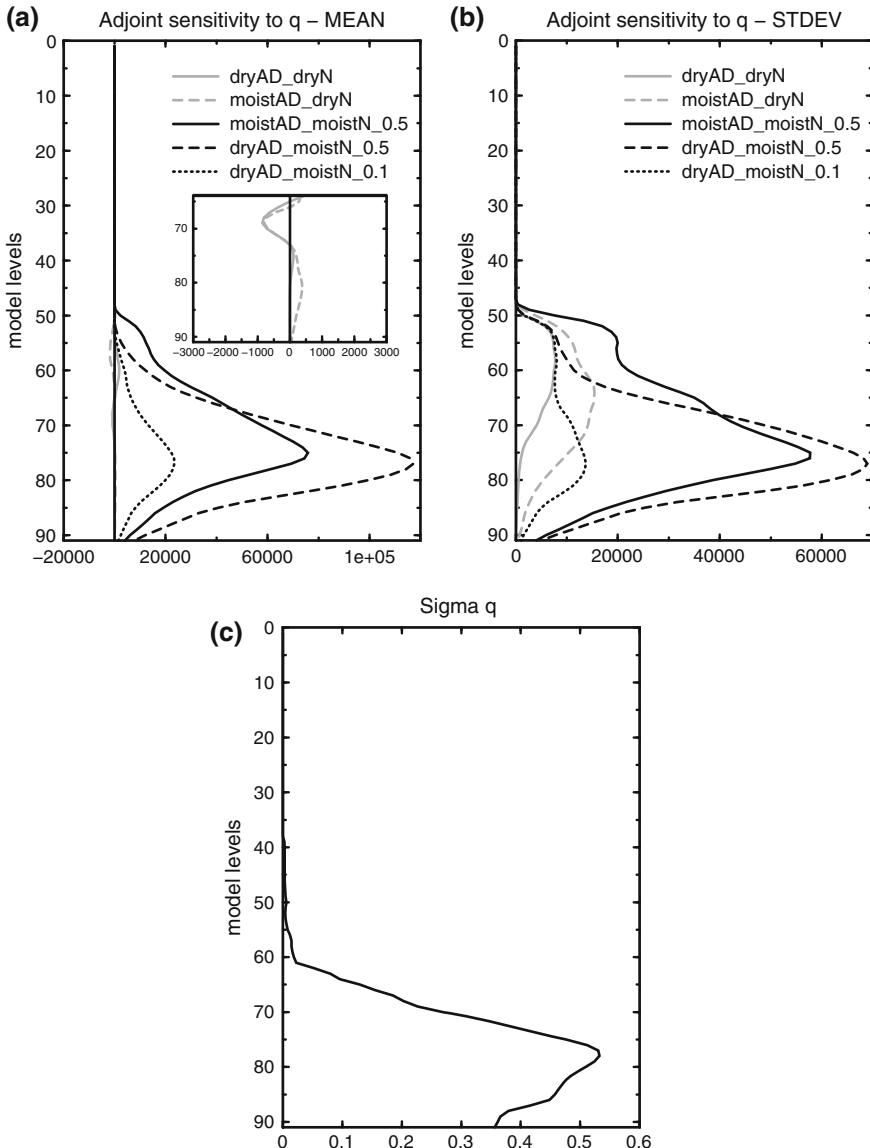


Fig. 8 Vertical profile of **a** mean specific humidity SG ($\text{J kg}^{-1}/(\text{g kg}^{-1})$) and **b** its standard deviation obtained by averaging over the whole test period (25 August 2010–10 September 2010). The results are presented for experiments with dry parametrization schemes (i.e. vertical diffusion, gravity wave drag and radiation) included in the adjoint model using dry (grey solid line) or moist norm (grey dashed line), and for experiments with moist processes also added using dry (black solid line) or moist norm with weighting factor of 0.5 (black dashed line) and 0.1 (black dotted line). **Panel c** displays vertical profile of typical values of the standard deviation of the ECMWF background errors for specific humidity (in g kg^{-1})

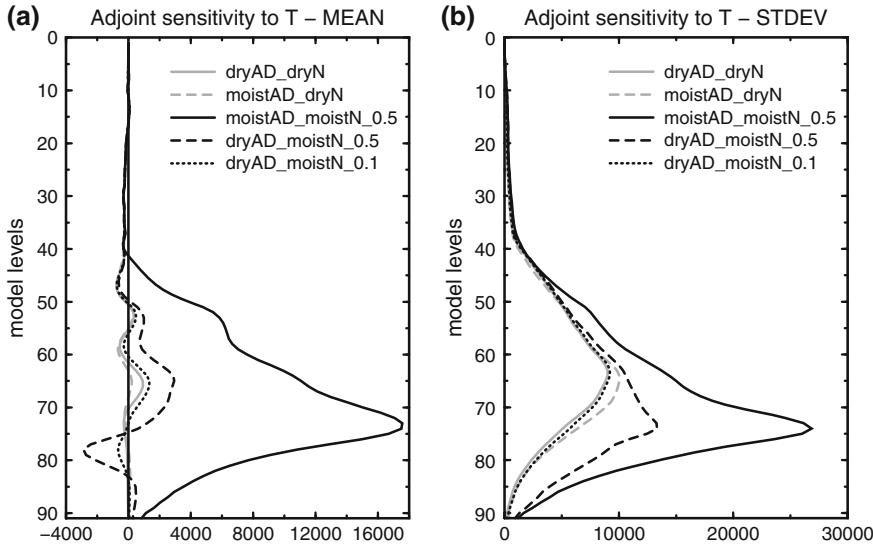


Fig. 9 Same as Fig. 8, but for temperature SG ($\text{J kg}^{-1}/\text{K}$) and over the whole test period

The impact on the FSOI diagnostics coming from different representation of physical processes in the adjoint model and different norms (dry or moist) used in SG will be illustrated in the following subsection.

3.3 FSOI

In this section, Forecast sensitivity Observation Impacts, FSOIs, computed using different sensitivity gradient configurations are compared. To obtain statistically significant results, the observation impacts have been averaged over 31 cases. Positive (negative) values mean forecast error increase (decrease) due to the assimilation of the observations. In this section, FSOI is shown in percentage, therefore positive values indicate forecast improvement. The observation types analysed are summarized in Table 1. Figure 10 compares FSOI computed with dry and moist adjoint model when the dry norm is used. FSOI differences are in general very small (less than 1 %) and for the majority of the observation types they are within the FSOI error calculation as shown by the error bars. These small differences reflect the small SG differences between **dryAD_dryN** and **moistAD_dryN** experiments (Figs. 2a and 4a versus Figs. 2c and 4c). The results as shown in Fig. 10 have been validated (by different FSOI developers and for different cases) by comparison with OSE observation impact diagnostics and an agreement within 10 % error was found (see for instance Gelaro and Zhu 2009). Completely different results are obtained when a moist norm is used to compensate for the lack of moist processes in the model adjoint.

Table 1 Observation types assimilated in the experiments. Number of observations per assimilated cycle is $\sim 10^7$

Data name	Data kind	Information
OZONE (O3)	Backscattered solar UV radiation, retrievals	Ozone, stratosphere
GEO-Rad	US/Japanese/EUMETSAT geostationary satellite infrared sounder radiances	Moisture, mid/upper troposphere
AMSU-B	Microwave sounder radiances	Moisture, troposphere
MHS	Microwave sounder radiances	Moisture, troposphere
MERIS	Differential reflected solar radiation, retrievals	Total column water vapour
GPS-RO	GPS radio occultation bending angles	Temperature, surface pressure
IASI	Infrared sounder radiances	Temperature, moisture, ozone
AIRS	Infrared sounder radiances	Temperature, moisture, ozone
AMSU-A	Microwave sounder radiances	Temperature
HIRS	Infrared sounder radiances	Temperature, moisture, ozone
SCAT	Microwave scatterometer backscatter coefficients	Surface wind
AMV-WV	Atmospheric Motion Vectors, retrievals from Water Vapour	Wind, troposphere
AMV-VIS	Atmospheric Motion Vectors, retrievals from Visible	Wind, troposphere
AMV-IR	Atmospheric Motion Vectors, retrievals from Infrared	Wind, troposphere
PROFILER	American, European and Japanese Wind profiles	Wind, troposphere
DROP	Dropsondes from aircrafts	Wind, temperature, moisture, pressure, troposphere
TEMP	Radiosondes from land and ships	Wind, temperature, moisture, pressure, troposphere
DRIBU	Drifting and Stationary buoys	Surface pressure, temperature, moisture, wind
Aircraft	Aircraft measurements	Wind, temperature, troposphere
SYNOP	Surface Observations at land stations and on ships	Surface pressure, temperature, moisture, wind

Figure 11 shows the observation impact computed by using either the dry adjoint model and the moist norm or the moist adjoint model and the dry norm. GEO-Rad, AIRS, HIRS and GPS-RO show a $\sim 5\%$ larger impact in **dryAD_moistN_0.5** than in **moistAD_dryN** whilst SCAT and SYNOP exhibit a detrimental impact of -2% and -1% , respectively. The differences between two configurations are actually -5% for SCAT and -7% for SYNOP. On the contrary, MHS and TEMP have $\sim 2\%$ and

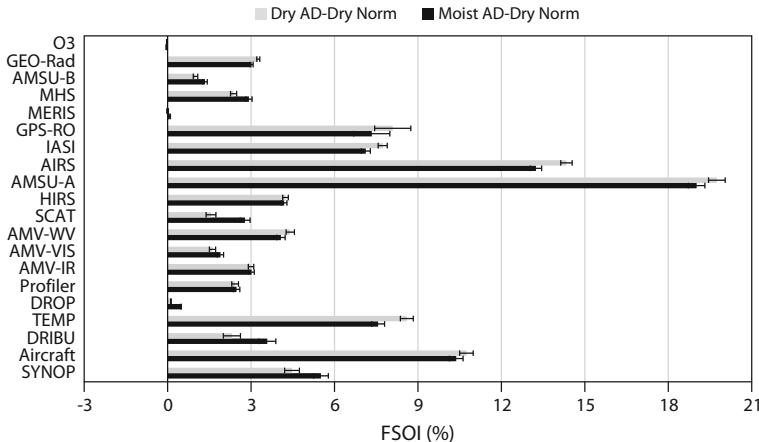


Fig. 10 Comparison of FSOI computed with the dry and the moist adjoint models (grey and black bars, respectively) using the dry norm. Results were averaged over 31 cases

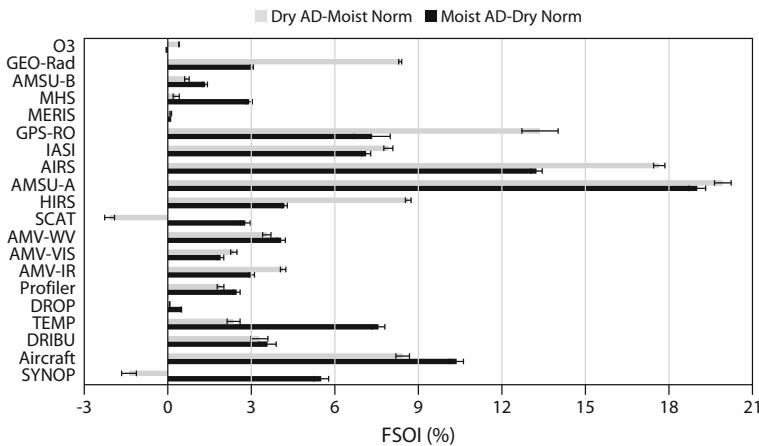


Fig. 11 Same as Fig. 10, but using either the dry adjoint model and the moist norm (grey bars) or the moist adjoint model and the dry norm (black bars)

~5 % smaller impact, respectively. The observation impact differences are due to the large mismatch between the forecast error sensitivity pattern with respect to both the humidity and temperature fields in the two experiments compared (Figs. 2c and 4c versus Figs. 2b and 4b).

Finally, the impact of using two different humidity weight factors, $w_q = 0.5$ and $w_q = 0.1$ in the forecast error energy norm is illustrated in Fig. 12. With $w_q = 0.1$, the negative impact of SCAT and SYNOP is replaced by a small positive one, which is

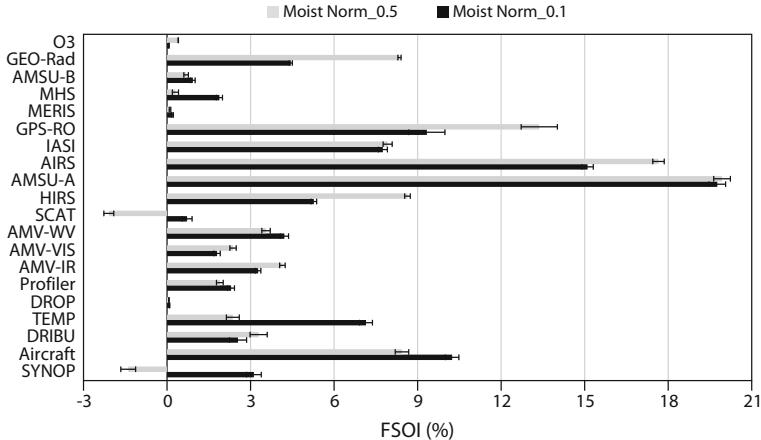


Fig. 12 Same as Fig. 11, but using the dry adjoint model and the moist norm with humidity weight factors $w_q = 0.5$ (grey bars) or $w_q = 0.1$ (black bars)

still 2 % smaller than in **moistAD_dryN** (Fig. 11). Despite the larger impact reduction (~ 3 %) when $w_q = 0.1$ is used instead of $w_q = 0.5$, GEO-Rad, GPS-RO, AIRS and HIRS still show at least 2 % more forecast impact than in **moistAD_dryN**. In conclusion, by decreasing the weight factors the negative impact of some observations vanishes and more consistency with the moist adjoint and dry norm experiment is found. However, the factor cannot be exactly determined but can only be experimentally tuned and the moist norm mainly introduces large-scale positive values in the sensitivity gradients on which FSOI strongly depends.

4 Conclusions

Forecast sensitivity observation impact experiments have been performed by using either dry or moist total energy norm combined with either dry or moist (i.e. combining dry and moist physical processes) adjoint model to evaluate the relevance of the humidity term in the norm.

It has been clearly shown that the use of the humidity term in the norm produces unrealistic humidity and temperature sensitivity gradients, which largely affect the observation forecast impact results. When the moist norm is combined with either dry or moist adjoint model the computed sensitivity gradient shows large scale positive sensitivity patterns, which are more representative of biases of the model than realistic forecast error sensitivity with respect to the initial and background conditions.

The forecast sensitivity observation impact tends to be very different when the moist norm is used as a quite larger impact is found for few observation types that

measure atmospheric humidity (together with temperature), like the infrared sounder radiances. Unexpectedly, both surface observations and surface winds retrieved from satellite over the ocean show degradation. These differences are due to the unrealistic sensitivity patterns mainly with respect to humidity which are generated by the norm. In fact, with the moist norm the humidity forecast error sensitivity patterns are artificially inflated, instead of showing the true effect of moist processes.

An appropriate tuning of the humidity weight factor in the norm is necessary to reduce the large unphysical patterns. Nevertheless, decreasing the humidity contribution up to 90 % still does not solve the problem. It is believed that the extra term in the norm is only an unnecessary artefact and that the humidity contribution in the norm is implicitly expressed through the temperature component, which varies to properly account for condensation and evaporation processes.

A full comprehensive description of the moist physical processes in the adjoint of the linearized model is necessary to properly propagate backward in time and space the humidity component of the forecast error. For systems that miss the representation of such processes the use of the dry adjoint model is recommended together with the dry norm; this combination provides, in percentage, a similar observation contribution to forecast error reduction. Moreover, to account for moist processes in the absence of moist physical parametrization in the adjoint model, the possibility of using hydrometeors in the energy norm could be explored in the future. This could provide more localized sensitivity gradients related to physical processes.

Acknowledgements The authors would like to thank Philippe Lopez and Anton Beljaars for their constructive comments to this article. The anonymous reviewer is also acknowledged for reviewing the manuscript.

References

- Baker NL, Daley R (2000) Observation and background adjoint sensitivity in the adaptive observation targeting problem. *Q J R Meteorol Soc* 126:1431–1454
- Barkmeijer J, Buizza R, Palmer TN, Puri K, Mahfouf J-F (2001) Tropical singular vectors computed with linearized diabatic physics. *Q J R Meteorol Soc* 201:685–708
- Bauer P, Radnoti G, Healey S, Cardinali C (2014) GNSS Radio occultation observing system experiments. *Mon Weather Rev* 142:555–572
- Bouttier F, Kelly G (2001) Observing-system experiments in the ECMWF4D-Var data assimilation system. *Q J R Meteorol Soc* 127:1469–1488
- Buizza R (1994) Impact of simple vertical diffusion and of the optimisation time on optimal unstable structures. ECMWF Technical Memorandum 192, ECMWF, Reading, UK, 25 pp
- Buizza R, Palmer TN, Barkmeijer J, Gelaro R, Mahfouf JF (1996) Singular vectors, norms and large-scale condensation. In: Proceedings of conference on numerical weather prediction, Norfolk, Virginia, American Meteorological Society, pp 50–52
- Cardinali C, Buizza R (2004) Observation sensitivity to the analysis and the forecast: A case study during ATreC targeting campaign. In: Proceedings of the First THORPEX International Science Symposium, 6–10 Dec 2004, Montreal, Canadian WMO TD 1237 WWRP/THORPEX N. 6
- Cardinali C, Pezzulli S, Andersson E (2004) Influence-matrix diagnostic of a data assimilation system. *Q J R Meteorol Soc* 130:2767–2786

- Cardinali C (2009) Monitoring the observation impact on the short-range forecast. *Q J R Meteorol Soc* 135:239–250
- Cardinali C (2015) Advanced data assimilation for geosciences. *École de Physique des Houches, Les Houches 2012. Chaps. 5 and 6.* Oxford University Press
- Chapnik B, Desroziers G, Rabier F, Talagrand O (2004) Properties and first applications of an error-statistics tuning method in variational assimilation. *Q J R Meteorol Soc* 132:543–565
- Courtier P, Andersson E, Heckley W, Pailleux J, Vasiljevic D, Hamrud DM, Hollingsworth A, Rabier F, Fisher M (1998) The ECMWF implementation of three-dimensional variational assimilation (3D-Var). Part I: Formulation. *Q J R Meteorol Soc* 124:1783–1807
- Daescu DN, Todling R (2010) Adjoint sensitivity of the model forecast to data assimilation system error covariance parameters. *Q J R Meteorol Soc* 136:2000–2012
- Ehrendorfer M, Errico RM, Raeder D (1999) Singular-vector perturbation growth in a primitive equation model with moist physics. *J Atmos Sci* 56:1627–1648
- English S, Saunders R, Candy B, Forsythe M, Collard A (2004) Met. Office satellite data OSEs. In: *Proceedings of Third WMO workshop on the impact of various observing systems on numerical weather prediction*, Alpbach, Austria, WMO/TD-1228 Geneva, pp 146–156
- Errico RM (2007) Interpretation of an adjoint-derived observational impact measure. *Tellus* 59A:273–276
- Gelaro R, Zhu Y (2009) Examination of observation impacts derived from observing system experiments (OSEs) and adjoint models. *Tellus* 61A:179–193
- Janisková M, Thépaut J-N, Geleyn J-F (1999) Simplified and regular physical parameterizations for incremental four-dimensional variational assimilation. *Mon Weather Rev* 127:26–45
- Janisková M, Mahfouf J-F, Morcrette J-J, Chevallier F (2002) Linearized radiation and cloud schemes in the ECMWF model: development and evaluation. *Q J R Meteorol Soc* 128:1505–1527
- Janisková M, Lopez P (2013) Linearized physics for data assimilation at ECMWF. In: Park SK, Xu L (eds.) *Data assimilation for atmospheric, oceanic and hydrological applications*, vol II. Springer, Berlin Heidelberg, pp 251–286. doi:[10.1007/978-3-642-35088-7_1](https://doi.org/10.1007/978-3-642-35088-7_1)
- Kelly G, Thépaut JN (2007) Evaluation of the impact of the space component of the global observation system through Observing System Experiments. *ECMWF Newsletter* 113, Autumn, 16–28
- Langland R, Baker NL (2004) Estimation of observation impact using the NRL atmospheric variational data assimilation adjoint system. *Tellus* 56A:189–201
- Laroche S, Tanguay M, Delage Y (2002) Linearization of a simplified planetary boundary layer parametrization. *Mon Weather Rev* 130:2074–2087
- Lopez P, Moreau E (2005) A convection scheme for data assimilation: Description and initial tests. *Q J R Meteorol Soc* 131:409–436
- Lorenc AC, Marriot RT (2014) Forecast sensitivity to observations in the Met Office global numerical weather prediction system. *Q J R Meteorol Soc* 140:209–224. doi:[10.1002/qj.2122](https://doi.org/10.1002/qj.2122)
- Mahfouf JF, Buizza R, Errico RM (1996) Strategy for including physical processes in the ECMWF variational data assimilation system. In: *Proceedings of ECMWF Workshop on non-linear aspects of data assimilation*, 9–11 Sept 1996, Reading, U.K, pp 595–632
- Mahfouf J-F (1999) Influence of physical processes on the tangent-linear approximation. *Tellus* 51:147–166
- Mahfouf J-F (2005) Linearization of a simple moist convection for large-scale NWP models. *Mon Weather Rev* 133:1655–1670
- Pariser RJ, Huang H-L (1993) Estimating effective data density in a satellite retrieval or an objective analysis. *J Appl Meteorol* 32:1092–1107
- Rabier F, Järvinen H, Klinker E, Mahfouf J-F, Simmons A (2000) The ECMWF operational implementation of four-dimensional variational assimilation. Part I: Experimental results with simplified physics. *Q J R Meteorol Soc* 126:1143–1170
- Tompkins AM, Janisková M (2004) A cloud scheme for data assimilation: Description and initial tests. *Q J R Meteorol Soc* 130:2495–2517

- Xu L., Langland R, Baker N, Rosmond T (2006) Development and testing of the adjoint of NAVDAS-AR. In: Proceedings of Seventh international workshop on adjoint applications in dynamic meteorology, 8–13 Oct 2006, Obergurgl, Austria
- Županski D, Mesinger F (1995) Four-dimensional variational assimilation of precipitation data. *Mon Weather Rev* 123:1112–1127
- Zhu Y, Gelaro R (2008) Observation sensitivity calculations using the adjoint of the Gridpoint Statistical Interpolation (GSI) analysis system. *Mon Weather Rev* 136:335–351
- Zou X, Navon IM, Sela JG (1993) Variational data assimilation with moist threshold processes using the NMC spectral model. *Tellus* 45A:370–387

Application of Conditional Nonlinear Optimal Perturbation to Target Observations for High-Impact Ocean-Atmospheric Environmental Events

Qiang Wang and Mu Mu

Abstract This paper reviews the progresses in the target observation studies for high-impact ocean-atmospheric environmental events based on the conditional nonlinear optimal perturbation (CNOP) approach. It is stressed that initial errors have important effects on the predictions of high-impact events, such as tropical cyclone (TC), Kuroshio large meander (KLM) and El Niño-Southern Oscillation (ENSO). To improve the initial conditions, the targeted observation studies of the above events have been performed. These studies used the spatial structure of the CNOP perturbation to determine the sensitive areas of the targeted observations. The validations of the sensitive areas were also investigated through a series of numerical experiments. The results showed that, for the all three kinds of events, if the targeted observations are deployed over the sensitive areas identified through the spatial pattern of the CNOP perturbation, the forecast results will be greatly improved. This implies that the CNOP method is useful for identifying the sensitive areas of the targeted observations of high-impact ocean-atmospheric environmental events.

1 Introduction

High-impact ocean-atmospheric environmental events refer to the oceanic, weather or climate events that cause huge economic and societal losses. Generally, dynamic and thermodynamic processes of atmosphere and ocean and the air-sea interactions all play important roles in the processes of occurrence, development and decay of such events. The tropical cyclone (TC), Kuroshio large meander (KLM) and El

Q. Wang · M. Mu (✉)

Key Laboratory of Ocean Circulation and Waves, Institute of Oceanology,
Chinese Academy of Sciences, Qingdao 266071, China
e-mail: mumu@qdio.ac.cn

Q. Wang · M. Mu

Function Laboratory for Ocean Dynamics and Climate, Qingdao National Laboratory
for Marine Science and Technology, Qingdao 266071, China

Niño-Southern Oscillation (ENSO) are the typical examples of such events. Because of the importantly socioeconomic impacts of these events, it is of significance to predict these events. However, although some progresses have been made, there are still considerable uncertainties in the predictions of these events.

Of the events mentioned above, TC, a rapidly rotating atmospheric storm system, is the cause of severe natural disaster. It can cause huge losses in human and economic each year. Hence, accurate TC forecasts in areas threatened by such storms is of great importance. Researchers have considered many ways to obtain a more accurate forecast of both TC track and strength. Due to the application of advanced numerical models and observations during the past decade, TC track forecasts have improved significantly, and errors have been reduced by nearly 50 % over the period 1980–2008 for forecasts in the Atlantic and eastern North Pacific (Franklin 2009). Notably, some studies have found that the TC track forecast error can be significantly reduced due to the improvement of the initial condition of the atmosphere in numerical model (Wu et al. 2007; Chou et al. 2011). This implies that the initialization of numerical model is very important for the TC track forecasts.

The KLM is an oceanic phenomenon characterized by large fluctuations of the Kuroshio path south of Japan. The occurrence of the KLM has important impact on local weather and climate (Xu et al. 2010; Nakamura et al. 2012) and consequently on fisheries and ship navigation. Although there are a few studies on the mechanism and predictability of KLM and a forecast system has also been developed in Japan Agency for Marine-Earth Science and Technology (JAMSTEC; Kagimoto et al. 2008), but the significant uncertainties still exist in the prediction of KLM. The predictability investigations have indicated that the initial error is an important error source causing the prediction uncertainties of KLM (Ishikawa et al. 2004; Fujii et al. 2008; Wang et al. 2012). Hence, to enhance the forecast skill of the KLM, the initial condition needs to be improved.

ENSO is one of the strongest modes of the climate variability on interannual time scales. Although it originates from the tropical Pacific, it affects climate, societies in many regions of the world. Hence, its prediction has attracted the attention of many scientists and policy makers in recent decades. But our current forecast skill for ENSO is still far from satisfactory. Similar to the TC and KLM, many researchers have also found that the initial errors have important effects on the ENSO prediction. For example, Moore and Kleeman (1996) pointed out that ENSO prediction is sensitive to the initial condition. Furthermore, Duan and Zhang (2010) and Yu et al. (2012a) compared the relative importance of the initial and model parameter errors based on a simple theoretical model and the Zebiak-Cane model, respectively and concluded that initial error is an important source causing the uncertainties in the ENSO prediction. Chen et al. (2004) suggested that improving the initialization for the Zebiak-Cane model (Zebiak and Cane 1987) can greatly overcome the spring predictability barrier (SPB) of ENSO and improve its forecast skill.

The above discussions indicate that the initial error has significant effects on the prediction of high-impact ocean-atmospheric environmental events, such as TC, KLM and ENSO. In fact, during the early development of numerical weather prediction, meteorologists have noticed that the forecast skill in a focused area is limited

by the initial conditions in a local region (Riehl et al. 1956). That is to say, the improvement in forecast skill from adding observations in a local area cannot be less than that induced by the observations in a wide area. This inspires some researchers to perform additional observations in a limited area rather than wide area, which will greatly save the cost of field observations. A question naturally arises: how to find the limited area and design an efficient observation strategy? Fortunately, the idea of targeted observation (also called adaptive observation) that has been developed since 1990s can be used to address this question. The idea is as follows: to obtain better prediction results at a future verification time (denoted as t_v), in a focused verification area, additional observations are deployed at a future targeted time (t_a), where $t_a < t_v$, in a (local) sensitive area where the observations are expected to have a great contribution on the improvement of forecast skills in the verification area (e.g. see Mu 2013; Mu et al. 2015). The additional observations are assimilated by data-assimilation system to generate a more accurate initial condition for the model.

A key of targeted observations is the determination of sensitive areas. To identify sensitive area, several methods have been developed. Examples include the quasi-inverse linear method (Pu et al. 1997), linear singular vector (LSV; Palmer et al. 1998), and the adjoint sensitivity method (Langland et al. 1999). Subsequently, methods based on “ensemble” were also proposed to identify sensitive areas, such as the ensemble transform (Bishop and Toth 1999), ensemble Kalman filter (Hamill and Snyder 2002), ensemble transform Kalman filter (ETKF; Bishop et al. 2001). The idea of linear approximation is used more or less in the methods mentioned above. To overcome the limitation of the linear approximation, a nonlinear method, conditional nonlinear optimal perturbation (CNOP, Mu et al. 2003, 2010), was proposed and has been successfully applied to determine the sensitive area of targeted observation for high-impact ocean-atmospheric environmental events. So-called CNOP represents the mode of the most sensitive initial perturbation in the nonlinear regime. Kerswell et al. (2014) pointed out that the CNOP approach is a useful tool for analyzing nonlinear stability. To show the potential of the CNOP method in identifying the sensitive area of targeted observation, this paper will review the application of CNOP method to high-impact events such as TC, KLM, and ENSO.

2 Sensitive Area for the Targeted Observation of TC

For TC, Chou et al. (2011) have indicated that the targeted observation is an important strategy for improving its track forecast. Mu et al. (2009) further showed that the CNOP method is a valid one for identifying the sensitive area for the targeted observation of TC. Subsequently, Qin and Mu (2012) used the CNOP approach to determine the sensitive area. As an example, Fig. 1a shows the sensitive regions identified by CNOP (shaded), which is for typhoon Mirinae (200921) and calculated by MM5 and its adjoint model. The points form half an annulus around the typhoon centre at 24 h. Most of the areas with sensitivity are located in the right-half quadrant with respect to the storm motion. Figure 1b shows the

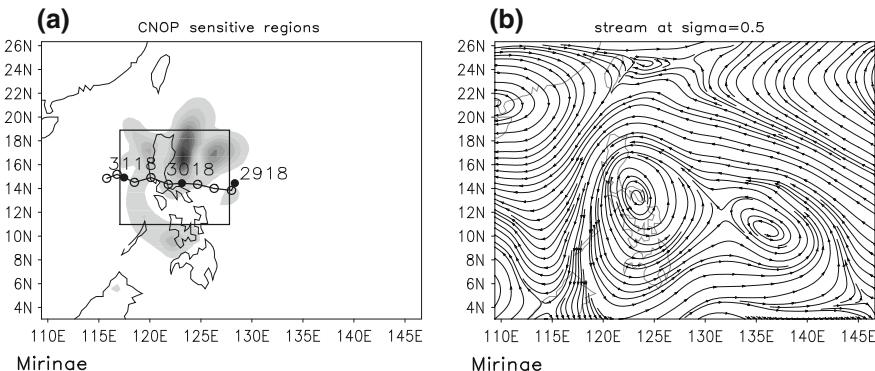


Fig. 1 Sensitive regions (shaded, normalized by respective maximum value) calculated by CNOPs (a) in the small rectangle over optimization period from '3018' (18UTC 30 October 2009) to '3118' for typhoon Mirinae. The dots and circles in the figure represent the position of the typhoon centre at 6-h intervals. The numbers from right to left (near each dot) represent the initial (0 h), targeting (24 h), and verification (48 h) times, respectively. b the stream lines at '3018' and $\sigma = 0.5$. (From Qin and Mu 2012)

streamlines for this case at 24 h and the vertical layer $\sigma = 0.5$. The CNOPs sensitive regions coincide well with the regions between south of the two subtropical highs: one centred over south China (110°E , 23°N) and the other over the western North Pacific. This finding indicates that the initial perturbations have large effects on the forecast in the verification region, implying that the steering flow between south of the two subtropical highs will affect the subsequent changes in Mirinae.

Although CNOP has not been utilized in the real-time field campaign, both observing system simulation experiments (OSSEs) and observing system experiments (OSEs) have been conducted to evaluate the effects of sensitive regions identified by CNOP. Chen et al. (2013) performed five experiments to investigate the usefulness of the sensitive area of targeted observations for each of 20 typhoon cases. The experiments are as follows: experiment 1 (EXP 1) was the control run that simulated the typhoon prediction using the NCEP reanalysis data, while experiments 2–5 (EXP 2–5) assimilated dropwindsonde observational data gathered from different sites in DOTSTAR at the initial time (0 h). EXP 2 assimilated all sonde data from each typhoon, but only the observational data from sensitive regions identified by CNOP (FSV) were used for EXP 3 (EXP 4). Finally, approximately the same number of dropwindsondes as used in EXP 3 and 4 was randomly selected, this time taking no account of the sensitive regions, to provide the data for EXP 5. The results of the experiments were shown in Fig. 2. The figure showed that using dropwindsonde data based on CNOP sensitivity can lead to improvements in typhoon track forecasting similar to, and occasionally better than, those achieved by assimilating all of the available data. Both approaches offered greater benefits than the other three alternatives averagely. It was proposed that CNOP provides a suitable approach to determining sensitive regions during targeted observation of typhoons.

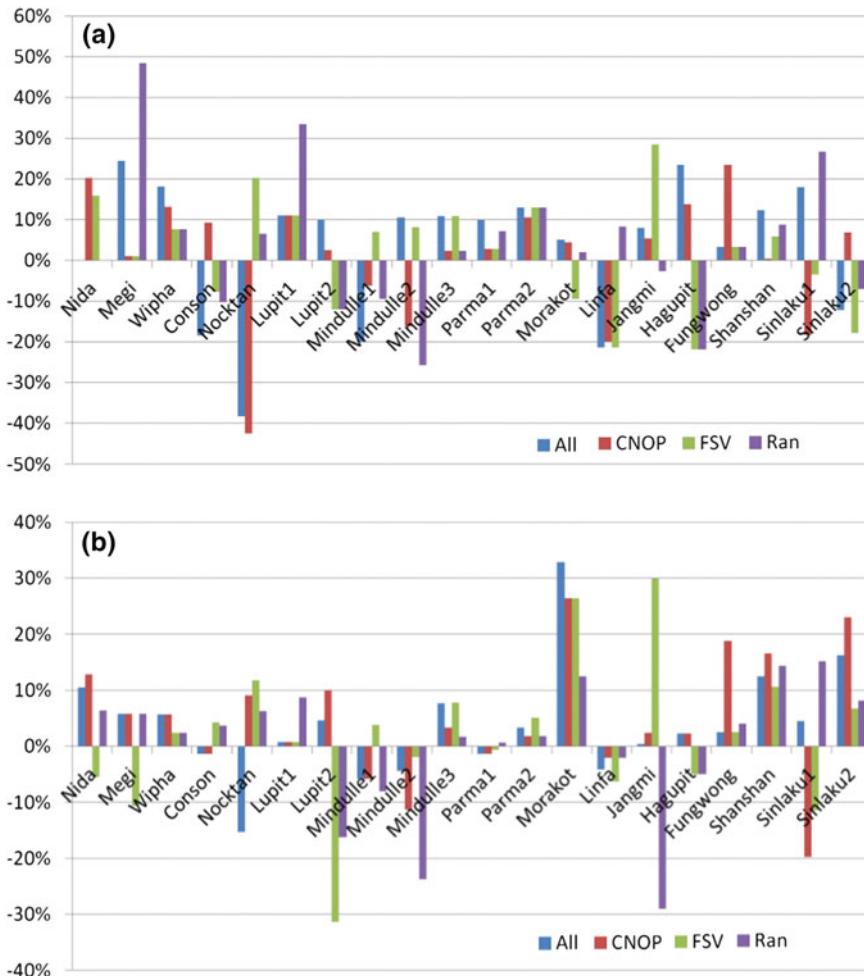


Fig. 2 Ratios of track forecast errors of EXP 2 to 5 (All, CNOP, FSV, and Ran) to EXP 1 at 24-h (a) and 36-h (b) using MM5 for twenty cases. Positive number means improvement, and negative number means deterioration. (From Chen et al. 2013)

3 Sensitive Area for the Targeted Observation of KLM

To show an application example of CNOP method to an ocean anomaly event, we will indicate the determination of sensitive area for the targeted observation of KLM in this section. As pointed out in the Sect. 1, the initial condition has important effects on the prediction of KLM. Therefore, Wang et al. (2013) investigated the impacts of initial perturbations on the predictability of KLM using the CNOP approach. They first calculated the optimal precursor (OPR) of the

occurrence of KLM within a 1.5-layer shallow-water model, where the OPR refers to the initial anomaly that can easiest develop into some weather or climate event under specific conditions. Figure 3a indicates the upper-layer thickness component of the OPR. We can see that the spatial structures of the OPR are very local and its large amplitude area is mainly located in the upstream region of KLM, namely the southeast of Kyushu.

Furthermore, Wang et al. (2013) also calculated the OGEs for the forecast of KLM. Here, the OGE represents the initial error which has the largest nonlinear evolution. They obtained two types of OGEs: type-1 OGE and type-2 OGE, where the type-1 OGE would cause the amplitude of the forecasted KLM to be underestimated, while the type-2 OGE would induce the forecasted KLM to be overestimated. The upper-layer thickness distributions of these two types of OGEs are shown in Fig. 3b, c. Figure 3 exhibits that the OPR is similar to the two types of OGEs and the similarity coefficient between the OPR and type-1 (type-2) OGE is negative (positive). They computed the similarity coefficients and found that their absolute values exceed 0.9. Besides, Wang et al. (2013) investigated the nonlinear evolution processes of OPR and OGEs. The results show that the evolution processes of the OPR are similar to those of the OGEs, which reflect that the evolutions of the OPR and the OGEs may share the similar mechanisms.

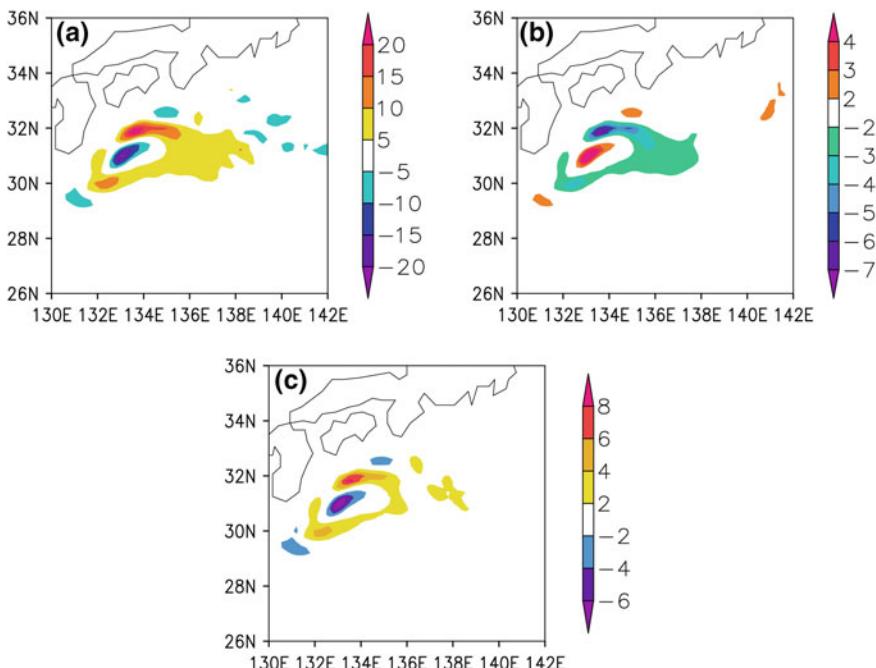
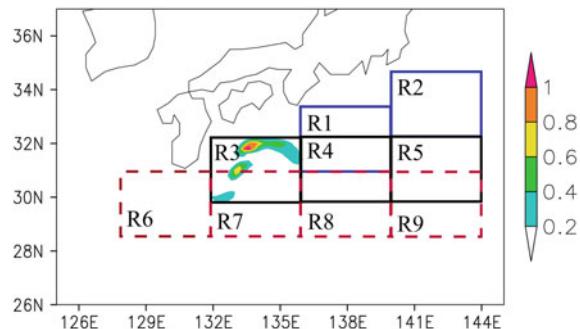


Fig. 3 The upper-layer thickness component of the OPR **a** for the Kuroshio large meander and **b** the type-1 and **c** the type-2 OGEs in the prediction of the large meander path (units: m). This figure is adapted from Wang et al. (2013)

Fig. 4 The spatial structure of the total energy for the type-1 OGE (shaded, units: $\text{m}^3 \text{s}^{-2}$). The nine regions were used for the ideal targeted observation experiments, and R3 was the sensitive area. The positions of these nine regions are listed in Table 1. This figure is from Wang et al. (2013)



The revealed similarity and localization features of OPR and OGEs inspire the authors to perform the targeted observation studies of the KLM forecast, because if the targeted observation carried out in one local area, it can not only find the precursor, but also reduce the initial error. This will help to improve the accuracy of the forecast and save the huge cost caused by the observation. Hence, Wang et al. (2013) determined the sensitive area for the targeted observation of the KLM forecast by computing the total energy distributions of type-1 OGE (Fig. 4). Because the spatial structures of OPR and OGEs are similar, the total energy distributions of the OPR and type-2 OGE are almost the same as those of the type-1 OGE. Figure 4 illustrates that large amplitude regions of the total energies are located mainly to the southeast of Kyushu, which is defined as the sensitive area R3.

To examine the validity of the sensitive area, Wang et al. (2013) investigated whether the targeted observations implemented over the sensitive area R3 can improve the forecast skill of KLM by a group of ideal experiments. To perform the experiments, nine local regions with the same size (240 grid points), including the sensitive area R3 and other eight arbitrarily chosen regions, were chosen (Fig. 4). These regions mainly cover the south of Japan and the upstream of Kuroshio extension. The experiment processes are as follows. First, they yielded 40 random initial errors over the whole model domain with the same amplitude of type-1 OGE. The nonlinear evolutions of these random errors were investigated. The averaged kinetic energy (denoted as J_1) of the forecast errors caused by these random errors was calculated. Then, suppose that the targeted observations are deployed in one of the nine regions, the random errors within this region will be eliminated, without changing the errors outside of that region. As a result, 40 random initial error fields were obtained for each region. Correspondingly, the evolutions of these errors were also examined and the averaged kinetic energy (denoted as J_2) of the prediction errors caused by them was computed. Finally, the relative improvement of the prediction due to the implementation of targeted observations over each region was measured by $(J_2 - J_1)/J_1$. According to the above definition, the minus represents the decrease of the forecast error. Table 1 shows the relative improvements of the KLM forecast for different areas. The reduction of the forecast error in the sensitive area is about 44 %, which is the largest in all regions, implying that the prediction

Table 1 The relative differences of the average kinetic energies of the forecast errors with and without implementing the targeted observations. From Wang et al. (2013)

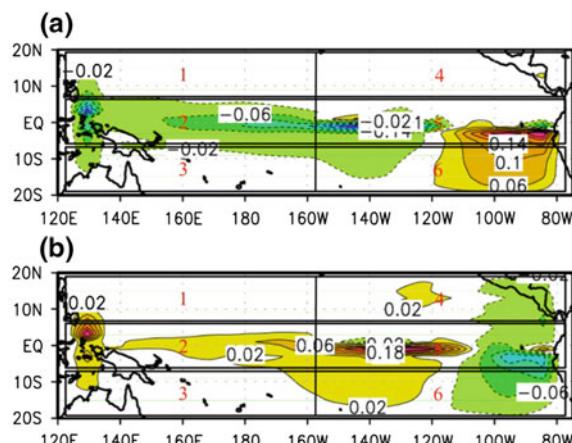
Region	R1	R2	R3	R4	R5	R6	R7	R8	R9
Relative differences (%) of forecast errors	-15.28	-1.68	-43.59	-11.85	-26.70	-0.04	-1.25	-1.53	0.73

of KLM will be greatly improved if the targeted observations are implemented over the sensitive area. It is worth noting that the similar results were also obtained by a recent study (Zou et al. 2015) based on the analyses of the more modeled KLM events. The above results show that if the observation network are deployed over the sensitive area determined by CNOP approach, the initial field in this region will be improved and the precursor signal of the occurrence of the KLM can also be well captured, which will greatly enhance the forecast skill of KLM.

4 Sensitive Area for the Targeted Observation of ENSO

In this section, we will show the determination of the sensitive area for the targeted observation of a high-impact climate event (ENSO) based on the CNOP approach. Similar to the KLM, Mu et al. (2014) calculated the OPR for ENSO onset and OGE for its forecast using the CNOP method and found the similarity between them. Specifically, with the Zebiak-Cane model, it was demonstrated that the CNOPs and local CNOPs can be regarded as the OPRs for El Niño and La Niña events, respectively (Fig. 5). The OPRs for El Niño/La Niña events exhibit a zonal dipole pattern for the SSTA component in the equatorial central and eastern Pacific and a

Fig. 5 The OPRs for **a** El Niño and **b** La Niña events with SST anomalies from Mu et al. (2014). Six rectangular domains denoted by Domain i ($i = 1, \dots, 6$) and Domain 5 is $157.5^{\circ}\text{W}-90^{\circ}\text{W}, 5^{\circ}\text{S}-5^{\circ}\text{N}$



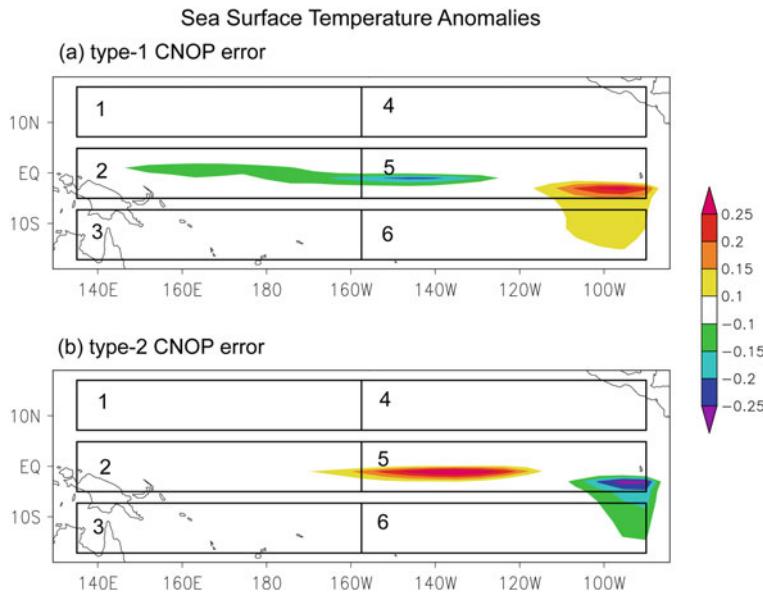


Fig. 6 The composite **a** type-1 and **b** type-2 OGE for El Niño events from Yu et al. (2012). Six rectangular domains denoted by Domain i ($i = 1, \dots, 6$). Large values of both types of CNOP error occur mainly in Domain 5 (157.5°W – 90°W , 5°S – 5°N)

basin wide deepening or shoaling along the equator for the thermocline depth anomaly component. For the El Niño events triggered by the OPRs in the ZC model, it was found that there are two types of OGEs causing the largest prediction errors of El Niño events (Fig. 6), and the spatial structures of OGEs share great similarities to the corresponding El Niño OPRs but with different signs. The large SSTA values for both OPRs and OGEs are mainly located in the equatorial eastern Pacific, which was also identified in realistic El Niño predictions with the coupled FGOALS-g model (Duan and Wei 2012). These regions with large values of initial perturbations may therefore represent the “sensitive area” for target observation of El Niño predictions. That is to say, if we implement the additional observation in these regions and assimilate them to the initial fields, the El Niño forecasting skill could be greatly improved.

Furthermore, Yu et al. (2012b) divided the tropical Pacific into six subsets equally and defined the equatorial eastern Pacific (157.5°W – 90°W , 5°S – 5°N) as the sensitive areas of target observation for El Niño prediction based on the OGEs, which is consistent with that identified by the OPRs due to their similarities (see Domain 5 in Figs. 5 and 6). In order to examine the validity of sensitive areas, Yu et al. (2012b) demonstrated that when only the initial errors of SSTA in the sensitive area are eliminated, El Niño prediction is more greatly improved compared to doing so in other regions with the same size (purple line in Fig. 7, Left). Meanwhile, the Niño-3 SSTA has the strongest growth after 1 year by superimposing the

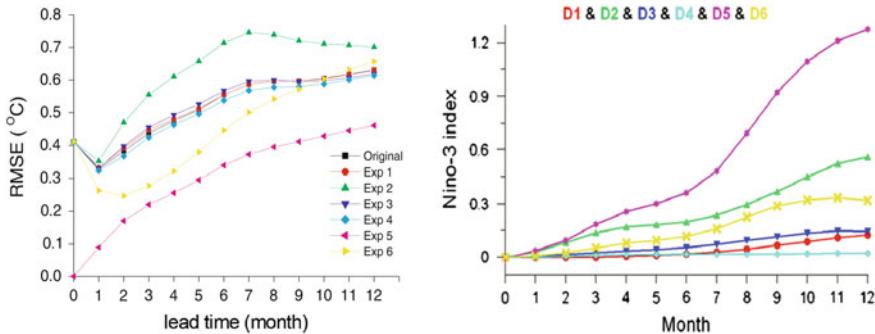


Fig. 7 *Left* Root-mean-square error for Niño-3 SSTA caused by the original initial errors (black line) and six other sets of initial errors. In Exp5, initial errors in sensitive areas (157.5°W – 90°W , 5°S – 5°N) from the original initial errors are eliminated. From Yu et al. (2012). *Right* The Niño-3 SSTA evolution caused by the six subsets of initial perturbations generated by adding in its components in one particular domain. D5 corresponds to the initial perturbations only in the sensitive areas (157.5°W – 90°W , 5°S – 5°N)

initial perturbation in the sensitive area, without adding any perturbation field in the other regions (purple line in Fig. 7, Right). However, the results from the ZC model mainly focused on the SSTA component and did not consider the role of subsurface anomalies due to the simplicity of the model. In fact, subsurface processes play an important role both in the evolution of the ENSO life cycle and in its predictions.

With the Community Earth System Model (CESM), Duan and Hu (2015) stated that there are two types of initial sea temperature errors often inducing large prediction errors of Niño-3 SSTA for El Niño events (Fig. 8). Similarly, they also identified the sensitive areas of target observation for El Niño prediction, which include the upper layer of the eastern equatorial Pacific and the lower layer of the western equatorial Pacific (i.e. the A, B and C regions in Fig. 8). Compared to the sensitive areas identified in the ZC model, the results in the CESM model highlight the influence of subsurface anomalies especially that in the subsurface layer of the western equatorial pacific. Using the CMIP5 model outputs, Zhang et al. (2015) confirmed that the optimal initial errors bear a strong resemblance to the optimal precursory disturbance for El Niño and La Niña events. It indicated that additional observations in the identified sensitive areas would be helpful not only to reduce the initial errors but also in detecting precursory signals for the development of ENSO events, which may greatly improve the ENSO prediction skill.

It is worth noting that a new type of El Niño events (CP-El Niño), in which warm SST is mainly concentrated in the central Pacific and different from the canonical events with maximum variations in the eastern equatorial Pacific (EP-El Niño), have occurred more and more frequently especially since 1990s (Ashok et al. 2007; Kao and Yu 2009; Kug et al. 2009). Unlike the EP-El Niño, the evolution of CP-El Niño events is mainly due to the zonal advective feedback rather than the thermocline feedback. Then, for such events, the sensitive areas of target observation should be further explored in the future.

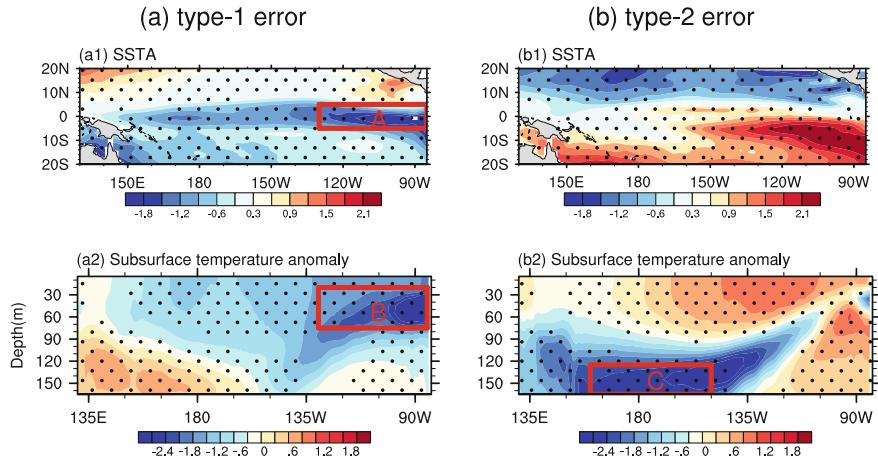


Fig. 8 Composite patterns of **a** type-1 and **b** type-2 initial errors in the CESM model. *Upper panels* show the SSTA component; *lower panels* show the equatorial (5°S – 5°N) subsurface temperature anomaly (units: $^{\circ}\text{C}$). Regions A, B and C represent are (5°S – 5°N , 150° – 85°W ; 0 – 5 m), (5°S – 5°N , 150° – 85°W ; 5 – 85 m) and (5°S – 5°N , 150°E – 135°W ; 120 – 165 m), respectively. Dotted areas indicate that the composites of SSTA and subsurface temperature anomaly errors exceed the 99 % significance level. This figure is from Duan and Hu (2015)

5 Summary and Discussion

We have reviewed the applications of CNOP method to the targeted observations of three high-impact ocean-atmospheric environmental events, including TC, KLM and ENSO. We first emphasized that initial conditions play important roles on the predictions of TC, KLM and ENSO based on the previous studies. As such, the CNOP method was used to calculate the optimal initial perturbation that can yield the largest impacts on the prediction of TC, KLM or ENSO, respectively. These initial perturbations exhibit obvious localization features. This inspired researchers to identify the sensitive areas of the targeted observations using the CNOP method.

In particular, for TC, the CNOP error was computed. The sensitive area was defined as the large amplitude area of the CNOP error. Subsequently, the observation system simulation experiments (OSSEs) and the observation system experiments (OSEs) were carried out to examine the usefulness of the sensitive area. The results showed that the additional observations over the sensitive area can greatly improve the prediction of TC. Similarly, for KLM and ENSO, the sensitive areas were also determined according to the localization features of the CNOP perturbations. To investigate the usefulness of the sensitive areas, a series of ideal experiments have been performed. The experiment results indicated that the forecast skills for KLM and ENSO can be improved if the targeted observations are deployed over the sensitive areas predetermined by the CNOP method. This implies that the CNOP method is a valid one for identifying the sensitive areas of the targeted observations of high-impact ocean-atmospheric environmental events.

It is worth noting that the similarity between the OPR and OGE in the prediction of KLM or ENSO has been revealed. This similarity feature shows that if the targeted observations are deployed over the only one sensitive area, the initial field in this region will be improved and the precursor signal of the occurrence of the KLM or ENSO can also be well captured. This will improve the forecast skill of KLM or ENSO. But for TC, the only OGE was calculated, so the similarity has not been revealed. More generally, whether do the similarity features underlie most high-impact ocean-atmospheric environmental events? To address this question, the OPR and OGE similarity for different high-impact events should be investigated.

It should point out that the validity of the targeted observation also depends on numerical model and data assimilation system. Therefore, we need develop the good enough numerical model and data assimilation system when performing the targeted observation studies of high-impact ocean-atmospheric environmental events. This requires the rising capabilities of computers, effective multidisciplinary study and cooperation of researchers from different fields. With the developments of numerical model, data assimilation system and targeted observations, it is expected that the predictions of high-impact ocean-atmospheric environmental events will be greatly improved.

Acknowledgements This word was supported by the National Natural Scientific Foundation of China (41230420, 41306023 and 41421005), the Strategic Priority Research Program of the Chinese Academy of Sciences (XDA11010303), and the NSFC Shandong Joint Fund for Marine Science Research Centers (U1406401).

References

- Ashok K, Behera SK, Rao SA, Weng HY, Yamagata T (2007) El Niño Modoki and its possible teleconnection. *J Geophys Res* 112:C11007. doi:[10.1029/2006JC003798](https://doi.org/10.1029/2006JC003798)
- Bishop CH, Toth Z (1999) Ensemble transformation and adaptive observations. *J Atmos Sci* 56:1748–1765
- Bishop CH, Etherton BJ, Majumdar SJ (2001) Adaptive sampling with the ensemble transform Kalman filter. Part I: Theoretical aspects. *Mon Weather Rev* 129(3):420–436
- Chen D, Cane MA, Kaplan A, Zebiak SE, Huang D (2004) Predictability of El Niño over the past 148 years. *Nature* 428:733–736
- Chen BY, Mu M, Qin XH (2013) The impact of assimilating dropwindsonde data deployed at different sites on typhoon track forecasts. *Mon Weather Rev* 141:2669–2682
- Chou KH, Wu CC, Lin PH, Aberson SD, Weissmann M, Harnisch F, Nakazawa T (2011) The impact of dropwindsonde observations on typhoon track forecasts in DOTSTAR and T-PARC. *Mon Weather Rev* 139(6):1728–1743
- Duan WS, Hu J (2015) The initial errors that induce a significant “spring predictability barrier” for El Niño events and their implications for target observation: results from an earth system model. *Clim Dyn.* doi:[10.1007/s00382-015-2789-5](https://doi.org/10.1007/s00382-015-2789-5)
- Duan WS, Wei C (2012) The ‘spring predictability barrier’ for ENSO predictions and its possible mechanism: results from a fully coupled model. *Int J Climatol* 33:1280–1292
- Duan WS, Zhang R (2010) Is model parameter error related to a significant spring predictability barrier for El Niño event? Result from theoretical model. *Adv Atmos Sci* 27:1003–1013

- Franklin JL (2009) 2008 National Hurricane Center forecast verification report. <http://www.nhc.noaa.gov/verification>
- Fujii Y, Tsujino H, Usui N, Nakano H, Kamachi M (2008) Application of singular vector analysis to the Kuroshio large meander. *J Geophys Res* 113:C07026. doi:10.1029/2007JC004476
- Hamill TM, Snyder C (2002) Using improved background-error covariances from an ensemble Kalman filter for adaptive observations. *Mon Weather Rev* 130(6):1552–1572
- Ishikawa Y, Awaji T, Komori N, Toyoda T (2004) Application of sensitivity analysis using an adjoint model for short-range forecasts of the Kuroshio path south of Japan. *J Oceanogr* 60(2):293–301
- Kao HY, Yu JY (2009) Contrasting eastern-Pacific and central-Pacific types of ENSO. *J Clim* 22:615–632
- Kagimoto T, Miyazawa Y, Guo X, Kawajiri H (2008) High resolution Kuroshio forecast system: description and its applications. In: High resolution numerical modelling of the atmosphere and ocean. Springer, New York, pp 209–239
- Kerswell RR, Pringle CCT, Willis AP (2014) An optimization approach for analysing nonlinear stability with transition to turbulence in fluids as an exemplar. *Rep Prog Phys* 77:085901
- Kug JS, Jin FF, An SI (2009) Two types of El Niño events: cold tongue El Niño and warm pool El Niño. *J Clim* 22:1499–1515
- Langland RH, Gelaro R, Rohaly GD, Shapiro MA (1999) Targeted observations in FASTEX: Adjoint-based targeting procedures and data impact experiments in IOP17 and IOP18. *Q J R Meteorol Soc* 125:3241–3270
- Moore AM, Kleeman R (1996) The dynamics of error growth and predictability in a coupled model of ENSO. *Q J R Meteorol Soc* 122:1405–1446
- Mu M (2013) Methods, current status, and prospect of targeted observation. *Sci China Earth Sci* 56:1997–2005
- Mu M, Duan WS, Wang B (2003) Conditional nonlinear optimal perturbation and its applications. *Nonlinear Process Geophys* 10:493–501
- Mu M, Zhou FF, Wang HL (2009) A method for identifying the sensitive areas in targeted observations for tropical cyclone prediction: conditional nonlinear optimal perturbation. *Mon Weather Rev* 137:1623–1639
- Mu M, Duan WS, Wang Q, Zhang R (2010) An extension of conditional nonlinear optimal perturbation approach and its applications. *Nonlinear Process Geophys* 17(2):211–220
- Mu M, Yu YS, Xu H, Gong TT (2014) Similarities between optimal precursors for ENSO events and optimally growing initial errors in El Niño predictions. *Theoret Appl Climatol* 115:461–469
- Mu M, Duan WS, Chen DK, Yu WD (2015) Target observations for improving initialization of high-impact ocean-atmospheric environmental events forecasting. *Natl Sci Rev* 2:226–236
- Nakamura H, Nishina A, Minobe S (2012) Response of storm tracks to bimodal Kuroshio path states south of Japan. *J Clim* 25(21):7772–7779
- Palmer TN, Gelaro R, Barkmeijer J, Buizza R (1998) Singular vectors, metrics, and adaptive observations. *J Atmos Sci* 55:633–653
- Pu ZX, Kalnay E, Sela J, Szunyogh I (1997) Sensitivity of forecast errors to initial conditions with a quasi-inverse linear method. *Mon Weather Rev* 125(10):2479–2503
- Qin XH, Mu M (2012) Influence of conditional nonlinear optimal perturbations sensitivity on typhoon track forecasts. *Q J R Meteorol Soc* 138:185–197
- Riehl H, Haggard WH, Sanborn RW (1956) On the prediction of 24-hour hurricane motion. *J. Meteor.* 13:415–420
- Wang Q, Mu M, Dijkstra HA (2012) Application of the conditional nonlinear optimal perturbation method to the predictability study of the Kuroshio large meander. *Adv Atmos Sci* 29:118–134
- Wang Q, Mu M, Dijkstra HA (2013) The similarity between optimal precursor and optimally growing initial error in prediction of Kuroshio large meander and its application to targeted observation. *J Geophys Res* 118:869–884
- Wu CC, Chen JH, Lin PH, Chou KH (2007) Targeted observations of tropical cyclone movement based on the adjoint-derived sensitivity steering vector. *J Atmos Sci* 64:2611–2626

- Xu H, Tokinaga H, Xie SP (2010) Atmospheric Effects of the Kuroshio Large Meander during 2004-05. *J Clim* 23(17):4704–4715
- Yu YS, Mu M, Duan WS (2012a) Does model parameter error cause a significant spring predictability barrier for El Niño events in the Zebiak-Cane model? *J Clim* 25:1263–1277
- Yu YS, Mu M, Duan WS, Gong TT (2012b) Contribution of the location and spatial pattern of initial error to uncertainties in El Niño predictions. *J Geophys Res* 117:C06018. doi:[10.1029/2011JC007758](https://doi.org/10.1029/2011JC007758)
- Zebiak SE, Cane MA (1987) A model El Niño Southern oscillation. *Mon Weather Rev* 115:2262–2278
- Zhang J, Duan WS, Zhi XF (2015) Using CMIP5 model outputs to investigate the initial errors that cause the “spring predictability barrier” for El Niño events. *Sci China: Earth Sci* 58:685–696
- Zou GA, Wang Q, Mu M (2015) Identifying the sensitive areas of adaptive observations for the prediction of Kuroshio large meander in a shallow-water model. *Chin J Oceanol Limnol* (Accepted)

Responses of Terrestrial Ecosystem to Climate Change: Results from Approach of Conditional Nonlinear Optimal Perturbation of Parameters

Guodong Sun and Mu Mu

Abstract Climate change is an important factor influencing the structure and carbon cycle of a terrestrial ecosystem. The estimation of a terrestrial ecosystem may be uncertain due to the unknown change about the climate in the future. In this chapter, we review the recent progress of the authors on the variations in a terrestrial ecosystem due to climate change using conditional nonlinear optimal perturbation of parameters (CNOP-P) approach. The stability of a grassland ecosystem to climate change is discussed first. A five-variable theoretical model of a grassland ecosystem is employed. The type of climate change described by the CNOP-P approach is called the CNOP-P or nonlinear type of climate change. Two linear types of climate change are used to compare stability of a grassland ecosystem. The results show that when it is affected by the CNOP-P-type climate change, a grassland ecosystem abruptly becomes a desert ecosystem. However, the two linear types of climate change do not lead to this abrupt change within a specific amplitude range. Similar results are found when a desert ecosystem is used as the reference state. Second, the maximum impact of climate change on the soil carbon in China is explored using the CNOP-P approach with the Lund-Potsdam-Jena (LPJ) model. The variations in the amount of soil carbon due to CNOP-P-type temperature or precipitation perturbation are compared with those caused by a linear perturbation of temperature or precipitation. The CNOP-P-type temperature or precipitation perturbation could lead to the variation of variability of temperature or precipitation. However, the linear type of temperature or precipitation perturbation fails. The numerical results demonstrate that, in southern China, the amount of soil carbon augments as a result of the CNOP-P-type temperature perturbation, and the variation in the amount of soil carbon due to the linear temperature perturbation is minor. The pool of fast decomposing soil carbon is an important factor to lead to the difference. The above

G. Sun · M. Mu (✉)

State Key Laboratory of Numerical Modeling for Atmospheric Sciences and Geophysical Fluid Dynamics (LASG), Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
e-mail: mumu@qdio.ac.cn

M. Mu

Institute of Atmospheric Sciences, Fudan University, Shanghai 200433, China

numerical results imply that the CNOP-P approach is an available tool to explore the response of terrestrial ecosystem to climate change.

Keywords Conditional nonlinear optimal perturbation of parameters • Climate change • Grassland ecosystem • Terrestrial ecosystem

1 Introduction

It is clear that the global climate system will change in the future due to human activities and fossil fuel use (Gerber et al. 2004). Variations will occur not only in the climatology but also in the climate variability. The variation in the climate variability is consequence of extreme heat, heat waves and heavy precipitation events, and so on. Therefore, in the future, the climate variability will vary with the increasing of extreme events (Piao et al. 2007).

Climate change plays a key role in influencing the structure of the terrestrial ecosystem and its carbon cycle (Zeng et al. 2004). Numerous observations show that the temperature and the precipitation are the primary forces driving variations in the terrestrial ecosystem (Liu et al. 2006). For example, an important desertification mechanism is the positive interaction between vegetation and climate in Sahel (Claussen et al. 1999). Recently, many studies have found that climate variability plays an indispensable role in the variation of the terrestrial ecosystem (Notaro et al. 2006). Mitchell and Csillag (2001) demonstrated that uncertainty in estimates of the net primary production (NPP) of grasslands is caused by the climate variability. Botta and Foley (2002) indicated that climate variability results in changes in an ecosystem's structure and the amounts of soil and vegetation carbon. Therefore, these results show that variations in the climatology and climate variability are the primary forces driving variations in the terrestrial ecosystem. And, it is necessary to explore the response of the terrestrial ecosystem to variations in the climatology and climate variability.

A common way of exploring the impact of climate change on the terrestrial ecosystem is to employ the output of general circulation models (GCMs, Ni et al. 2004; Eglin et al. 2010). However, some studies have indicated that the climate data simulated by GCMs are uncertain (Kharin et al. 2007). Therefore, the estimations of climate mean state and climate variability may be uncertain, and the response of the terrestrial ecosystem to the climate data is uncertain (Trumbore and Czimeczik 2008). Although GCMs have produced multiple climate scenarios, the maximum extent of the uncertainty in the simulated terrestrial ecosystem due to these outputs could not be measured.

In the studies of Mu et al. (2003), the approach of conditional nonlinear optimal perturbation related to initial perturbation (CNOP) is proposed. The approach has been used to investigate the dynamics of the ENSO's predictability, the prediction error (Duan and Mu 2006; Mu et al. 2007a), the nonlinear stability of steady states

of thermohaline circulation (Mu et al. 2004), adaptive observation (Mu et al. 2007b), ensemble prediction (Mu and Jiang 2008), the grassland ecosystem (Mu and Wang 2007) and the predictability study of the Kuroshio large meander (Wang et al. 2012). Mu et al. (2010) extended this approach to find the optimum integrated mode of the initial perturbations with the model parameters. The CNOP approach relating to initial perturbations is called the CNOP-I approach, and that relating to perturbations of the model's parameters is called the CNOP-P approach. The CNOP-P approach has been used in discussions of the ENSO's predictability (Mu et al. 2010). All of these applications show that CNOP approach are useful tools for studying nonlinear systems and indicate that they may also be effective for exploring the response of the terrestrial ecosystem to climate change. Therefore, there are some studies to explore the maximum amount of variation in the terrestrial ecosystem caused by climate change, including the climate and its variability, using the CNOP-P approach (Sun and Mu 2011, 2012). In this chapter, our goal is to review these studies.

2 The Models and Method

2.1 *The Theoretical Five-Variable Grassland Ecosystem Model*

Firstly, a simple model (a five-variable grassland ecosystem model) is used to show the variation of terrestrial ecosystem due to climate change based on the work of Sun and Mu (2011). The five-variable grassland ecosystem model includes a three-variable ecosystem model and a three-layer land surface hydrological model for one species of grass (Zeng et al. 2005) that is native to Inner Mongolia. The model is as follows:

$$\begin{aligned}
 \frac{dM_c}{dt} &= \alpha^* (G(M_c, W_r) - D_c(M_c, W_r) - C_c(M_c)), \\
 \frac{dM_d}{dt} &= \alpha^* (\beta' D_c(M_c, W_r) - D_d(M_d) - C_d(M_d)), \\
 \frac{dW_c}{dt} &= P_c(M_c) + E_r(M_c, W_r) - E_c(M_c, W_r) - R_c(M_c), \\
 \frac{dW_s}{dt} &= P_s(M_c) + E_s(M_c, W_s, M_d) + R_c(M_c) - Q_{sr}(W_s, W_r) - R_s(M_c, W_s, M_d), \\
 \frac{dW_r}{dt} &= P_r(M_c) + \alpha_r R_s(M_c, W_s, M_d) + E_r(M_c, W_r) + Q_{sr}(W_s, W_r) - R_r(M_c, W_r).
 \end{aligned}$$

The model includes five variables. They are the amount of living biomass (M_c), the amount of wilted biomass (M_d), the water content of the vegetation canopy (W_c), the water content of the thin surface layer of soil (W_s) and the water content of the

root layer (W_r). More details of the model's parameters and their physical explanations can be found in Zeng et al. (2006) and Sun and Mu (2009). Although the model is simple, it clearly and concisely represents the essential features of the complex atmosphere-ecosystem-soil system, including multiple equilibria, bifurcations and abrupt changes. The model has also been used to investigate the nonlinear stability of grassland equilibrium in response to human activity (Sun and Mu 2009).

2.2 *The Lund-Potsdam-Jena (LPJ) Model*

The simple model could show the essential characters of variation of terrestrial ecosystem. Meanwhile, the complex model, such as dynamic global vegetation model (DGVM), provides an effective way to explore the variations of plant functional type (PFT) and terrestrial cycle due to the climate change (Bonan et al. 2002). The LPJ DGVM is used in the study of Sun and Mu (2012) due to its broad applicability to the terrestrial carbon and hydrological cycles. This model, which originates in the biome family of models (Prentice et al. 1992), simulates the distribution of functional types of plants and combines process-based representations of terrestrial vegetation dynamics and land-atmosphere carbon and water exchanges. The LPJ DGVM explicitly includes photosynthesis, mortality, fire disturbances, soil heterotrophic respiration, and other factors. A detailed description and evaluation of the model was supplied by Sitch et al. (2003).

The data, which is applied to drive the model, are the monthly precipitation, temperature, wet frequency and cloud cover from the Climatic Research Unit (CRU) for the period from 1901 to 1998 (Mitchell and Jones 2005). Data on global atmospheric CO_2 concentration based on the carbon cycle model, ice-core measurements and atmospheric observations (Kicklighter et al. 1999). The soil texture data come from the FAO soil data set (Zobler 1986).

2.3 *Conditional Nonlinear Optimal Perturbation of Parameters (CNOP-P)*

In the study of Mu et al. (2010), the CNOP-P approach is proposed based on types of predictability. The CNOP-P is the parameter perturbation, which could result in the maximal cost function at the final time step. In this section, we review the derivation of this approach for readers' convenience. The nonlinear differential equations are as follows:

$$\begin{cases} \frac{\partial U}{\partial t} = F(U, P) & U \in R^n, t \in [0, T] \\ U|_{t=0} = U_0 \end{cases} \quad (1)$$

where F is a continuous nonlinear operator, P is a parameter vector, and U_0 is the initial value. Let M_τ be the propagator of the nonlinear differential equations from the initial time, 0, to time τ . u_τ is a solution to the nonlinear equations at time τ that satisfies $u(\tau) = M_\tau(u_0, p)$.

Let $U(T; U_0, P)$ and $U(T; U_0, P) + u(T; U_0, p)$ be solutions to the set of nonlinear differential equations Eq. (1) based on P and $P + p$, respectively, where P and p are parameter vectors. $u(T; U_0, p)$ describes the departure from the reference state $U(T; U_0, P)$ caused by p . The solutions satisfy

$$\begin{cases} U(T; U_0, P) = M_T(U_0, P) \\ U(T; U_0, P) + u(T; U_0, p) = M_T(U_0, P + p) \end{cases}.$$

For the proper norm, $\|\cdot\|$, a parameter perturbation p_δ is called the CNOP-P if and only if

$$J(p_\delta) = \max_{p \in \Omega} J(p), \quad (2)$$

where

$$J(p) = \|M_T(U_0, P + p) - M_T(U_0, P)\| \quad (3)$$

P is a reference state, and p is a perturbation of the reference state. $p \in \Omega$ is a constraint.

2.4 Experimental Design

There are different experimental designs for the two models. For the theoretical grassland ecosystem model, $\|\cdot\|$ is L_2 norm. The variations of unknown variables are constrained by δ ($\delta = 0.1, 0.2, 0.3$ and 0.4). And, two linear stable equilibriums are considered as the target variables.

However, for the DGVM, the experimental designs are introduced. There are lots of studies to study the impact of climate change on terrestrial ecosystems (Gao et al. 2003; Gerber et al. 2004; Matthews et al. 2005). As usually, a fixed perturbation series is added into the reference series as follows:

$$\frac{\sum_{i=1}^n (X_i + \delta)}{n} = \frac{\sum_{i=1}^n (X_i)}{n} + \delta \quad (4)$$

where $\{X_i\}_{i=1, n}$ is the annual temperature or precipitation during the time period under study. δ represents the fixed perturbation. One character of this approach considers variations in the climatology of the temperature or the precipitation. This

above method supplies a type of climate scenario, which is called as the linear climate change scenario. In fact, the variations of annual perturbation of temperature or precipitation are different. It is not reasonable to the fixed constant is added into the temperature or precipitation series. Hence, to consider the annual perturbation, the experiment described below is designed.

In left side of Eq. (4), the fixed constant is replaced by unknown variable x_i . x_i represents the unknown annual perturbation, and satisfies

$$\frac{\sum_{i=1}^n (X_i + x_i)}{n} = \frac{\sum_{i=1}^n (X_i)}{n} + \delta \quad (5)$$

$$0 \leq x_i \leq \sigma \quad (6)$$

From Eq. (5), we find x_i not only leads to the variation of the mean of the temperature or the precipitation that are similar to Eq. (4), but also leads to the variation of variability of annual temperature or precipitation perturbations labeled by standard deviation. Moreover, x_i should be bounded by σ in Eq. (6). Hence, the above formulae Eqs. (5) and (6) also show a type of climate scenario, which is time-dependent and nonlinear. This type of climate scenario is known as the nonlinear type climate change scenario. In the study of Sun and Mu (2012), the choices of δ and σ are shown in Table 1.

Soil carbon is an important part in the terrestrial ecosystem carbon cycle. In our study, the amount of soil carbon is considered as the research subject. Before the model is applied, the soil carbon pool is absence. It is unjustified to run the model. So, the LPJ model need be run using CRU data for the period from 1901 to 1930 for 1000 model years so that it reaches an approximate equilibrium in terms of the number of carbon pools and amount of vegetation cover.

To compute the optimal value of the optimization problem (2), the two optimization algorithms are employed. For the theoretical grassland ecosystem model, because the information of gradient about the unknown variables could be obtained, the spectral projected gradients (SPG2) algorithm (Birgin et al. 2000) is applied. However, since the gradient about the unknown variables may be non-differentiable for the DGVM, differential evolution (DE; Storn and Price 1997) is employed. The advantage of DE is that it could discover the optimal value without information of gradient.

Table 1 The choices of δ and σ in Eqs. (5) and (6)

Variable	δ	σ
Temperature	2°	3°
Precipitation	$\bar{P} \times 20\%$ ^a	$P_{\max} \times 20\%$ ^b

^a \bar{P} is the mean precipitation

^b P_{\max} is the maximum of precipitation in study period

3 The Numerical Results for a Grassland Ecosystem

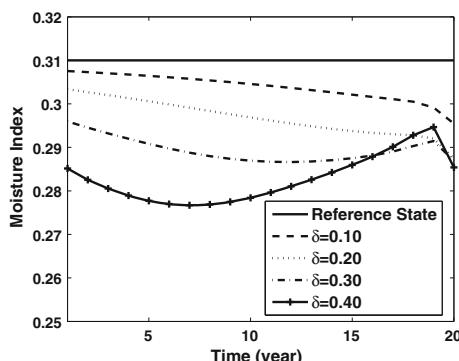
3.1 The Response of a Grassland Ecosystem to CNOP-P-Type Climate Change

Within the theoretical grassland ecosystem model, an important factor is μ , which is the ratio of precipitation to potential evapotranspiration and represents climate change. In the study of Sun and Mu (2011), the stability of grassland ecosystems to climate change or moisture index change ($\mu = 0.31$) is discussed. There are two linearly stable equilibrium states (grassland and desert) when the moisture index is 0.31. These are considered as the target subjects.

Figure 1 shows the annual variation in the moisture index, μ , during the time period when the CNOP-P are added into the reference state for the different constraint conditions ($\delta = 0.1, 0.2, 0.3, 0.4$) for the grassland ecosystem. The numerical results show that the moisture index, which could cause the maximal variation of the grassland ecosystem equilibrium, decreases compared to the reference state. This means that the drought maybe occur during the study period. Under the CNOP-P-type climate change, the variation of grassland equilibrium state is analyzed in Fig. 2. It is found that the living biomasses decrease due to the drought for the four CNOP-P-type climate changes. However, there are different results at the final time. When $\delta = 0.1, 0.2$, and 0.3 , although the living biomasses decrease at the initial 20 years, the living biomasses increase due to the relief of drought. When $\delta = 0.4$, the living biomass will unceasingly decrease despite the relief of drought. The numerical results indicate that the grassland ecosystem will evolves into a desert if there is small quantity of living biomass.

When the desert ecosystem is regarded as the target, the moisture index will increase due to the CNOP-P-type perturbation for four constraint conditions ($\delta = 0.1, 0.2, 0.3, 0.4$). This means that the warm and humid climate condition could lead to the maximal variation of the desert ecosystem (Fig. 3). Figure 4 shows the variation of the desert ecosystem under the CNOP-P-type climate change. The desert ecosystem fails to change into the grassland ecosystem at the

Fig. 1 Annual variations of moisture indices caused by the CNOP-Ps superimposed upon the reference state ($\mu = 0.31$) for the different constraint conditions for the grassland ecosystem, with an optimization time of 20 years ($T = 20$) (From Sun and Mu 2011)



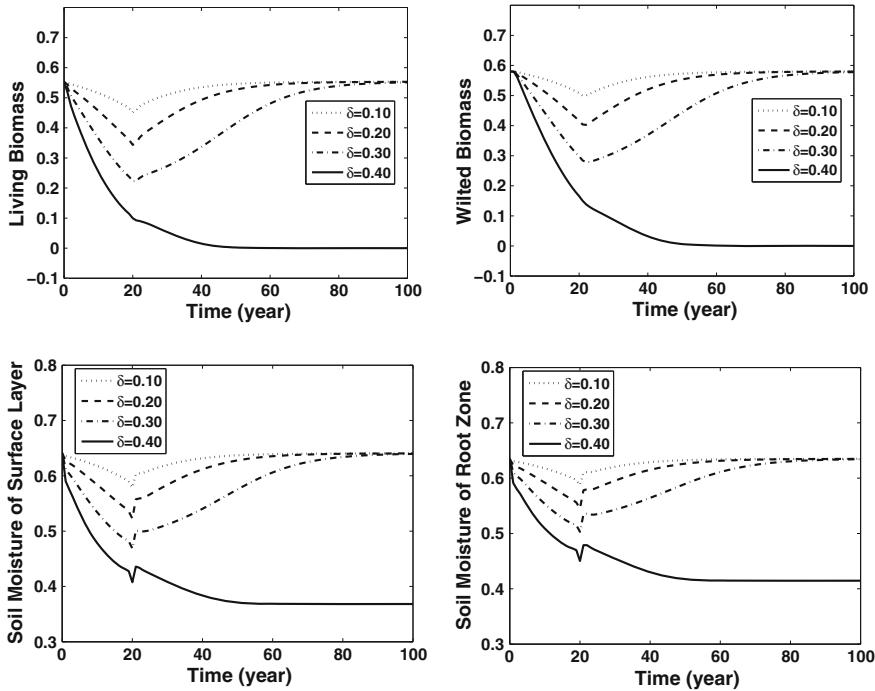
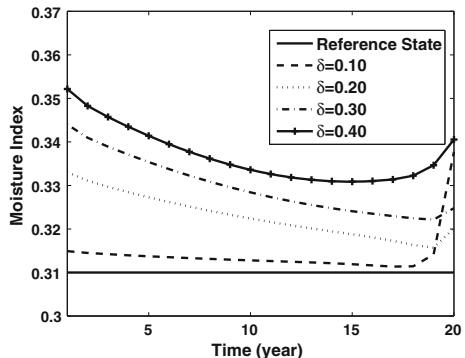


Fig. 2 The nonlinear evolution of four components of the grassland ecosystem influenced by CNOP-P-type climate change and their evolution when the moisture index recovers to the reference state (From Sun and Mu 2011)

Fig. 3 Same as Fig. 1, but for the desert ecosystem (From Sun and Mu 2011)



final time for slight climate warm and humid. With the increasing of the constrain condition, the desert ecosystem influenced by the CNOP-P-type climate change will translate into the grassland equilibrium state when $\delta = 0.3$ or 0.4.

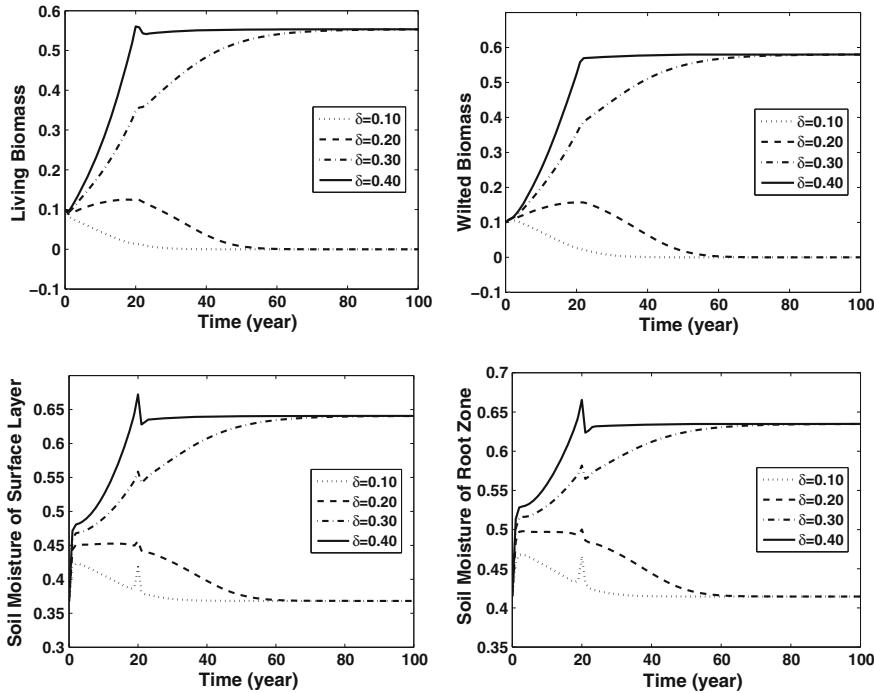


Fig. 4 Same as Fig. 2, but for the desert ecosystem (From Sun and Mu 2011)

3.2 A Comparison of the Responses to Nonlinear and Linear Climate Change

As shown above, the CNOP-P-type climate change represents the nonlinear type of climate change. To show the nonlinear character of the two equilibria, two linear climate perturbations, which can be classified by their slopes, are employed. Firstly, the variations of two ecosystem equilibria at final time are shown in Tables 2 and 3 for $\delta = 0.1, 0.2, 0.3$ and 0.4 . It is shown that the CNOP-P-type climate change causes greater variations than two type linear climate change for the grassland and desert ecosystems. Secondly, to discuss the abrupt changes of grassland and desert ecosystems, the linear types of climate change with the certain constrain conditions $\delta = 0.339$ for the grassland ecosystem and $\delta = 0.24$ for the desert ecosystem are applied (Fig. 5). The numerical results show that the grassland ecosystem or the desert ecosystem will evolve into the desert ecosystem or the

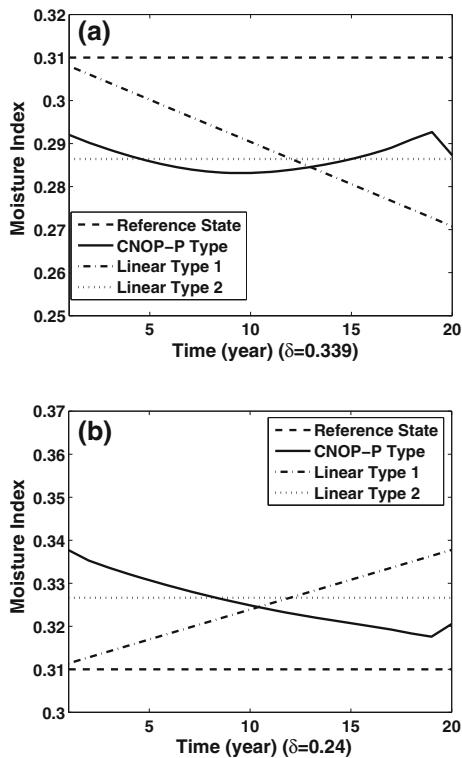
Table 2 The relative change of the grassland ecosystem caused by three kinds of climate change for $T = 20$ (From Sun and Mu 2011)

δ	CNOP-P-type	Linear type 1	Linear type 2
0.1	0.121	0.120	0.111
0.2	0.252	0.249	0.243
0.3	0.402	0.379	0.399
0.4	0.566	0.504	0.557

Table 3 Same as in Table 1, but for the desert ecosystem (From Sun and Mu 2011)

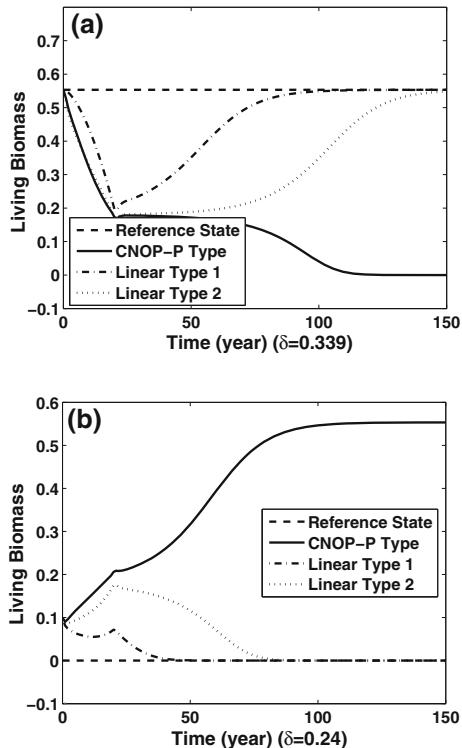
δ	CNOP-P-type	Linear type 1	Linear type 2
0.1	0.141	0.106	0.126
0.2	0.425	0.219	0.368
0.3	1.005	0.423	0.896
0.4	1.556	0.769	1.491

Fig. 5 The annual variation in moisture indices for the different types of climate change. **a** the grassland ecosystem; **b** the desert ecosystem (From Sun and Mu 2011)



grassland ecosystem with the CNOP-P-type climate change. However, the grassland ecosystem or the desert ecosystem fails with the two linear types of climate change (Fig. 6).

Fig. 6 The nonlinear evolution of the living biomass influenced by different climate change types and their evolution when the moisture index recovers to the reference state. **a** the grassland ecosystem; **b** the desert ecosystem (From Sun and Mu 2011)



4 The Impact of Climate Change on Soil Carbon

4.1 Variations in the Amount of Soil Carbon Due to Temperature Changes

4.1.1 The Impact of CNOP-P-Type Temperature Changes on the Soil Carbon

The theoretical studies about the terrestrial ecosystem to climate change are introduced in the above section. In this section, we review the variations of the terrestrial ecosystem due to climate change with the DGVM (Figs. 7 and 8). Figure 8a shows the temporal changes in the temperature based on from 1961 to 1970 using the CNOP-P-type perturbation. The spatial variations of soil carbon are shown in Fig. 8a. Compared to the reference state, the soil carbon increases in northeastern China and parts of southern and northern China, whereas the soil carbon in the arid and semi-arid regions of China decreased. Temperate broad-leaved evergreen (TeBE), temperate broad-leaved deciduous (TeBS), and boreal needle-leaved evergreen (BoNE) trees and temperate herbaceous plants

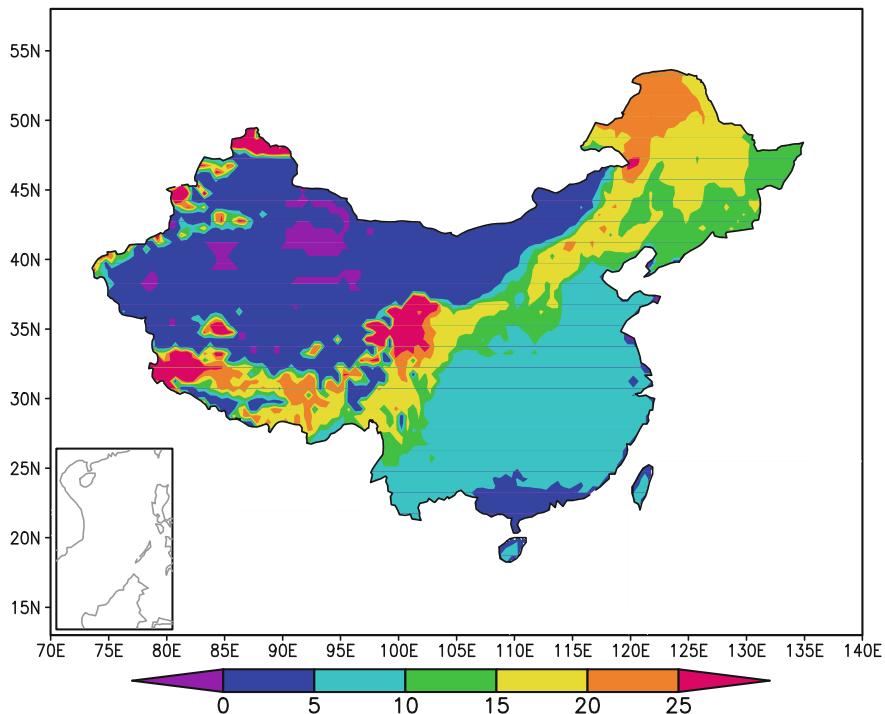


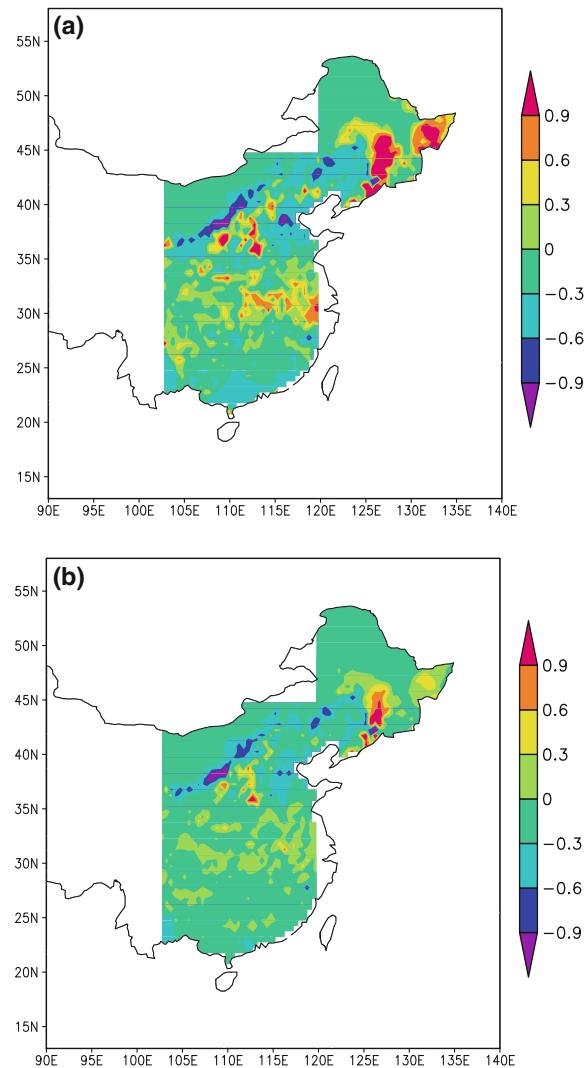
Fig. 7 The spatial mean soil carbon density in China from 1981 to 1998 ($\text{Kg C m}^{-2} \text{ year}^{-1}$) (From Sun and Mu 2012)

(TeH) are the main PFTs in the study region. The soil carbon increases for the BoNE trees (by approximately $62 \text{ g C m}^{-2} \text{ year}^{-1}$) and decreased for the TeBE, TeBS, and TeH (Fig. 8).

4.1.2 Variations in the Amount of Soil Carbon Resulting from Different Types of Temperature Perturbation

In the previous studies, to explore the response of terrestrial ecosystem to climate change, a linear type climate change scenario, which is 2° applied to the reference temperature change scenario. The linear type temperature change just considers the variation of climatology. For the CNOP-P-type temperature, the variations of climatology and climate variability are taken into account. Figure 8 shows that the soil carbon responds weakly to linear temperature changes, whereas the amount of soil carbon was intensely augmented as a result of the CNOP-P-type temperature change. Moreover, in Southern China, the variation of soil carbon due to the CNOP-P-type temperature change is significant greater than that due to the linear type temperature change. The augment of soil carbon implies that the carbon sink

Fig. 8 The variation in soil carbon density when compared to the reference state ($\text{Kg C m}^{-2} \text{ year}^{-1}$). **a** the CNOP temperature change; **b** the linear temperature change (From Sun and Mu 2012)



could occur in Southern China due to the nonlinear type temperature change and variability. Furthermore, in Table 4, the absolute variation in comparison with the reference state as a result of the CNOP-P-type temperature change is greater than the variation caused by the linear temperature change for the different PFTs.

Within the LPJ model, the soil carbon is calculated using belowground litter and pools of quickly and slowly decomposing soil carbon. To discuss the variations of three components due to the different types of temperature change, Fig. 9 is shown. It is shown that the variation of quickly decomposing soil carbon is main contribution for the variation of soil carbon. The amounts of belowground litter and slowly decomposing soil carbon are limited for the variation of soil carbon.

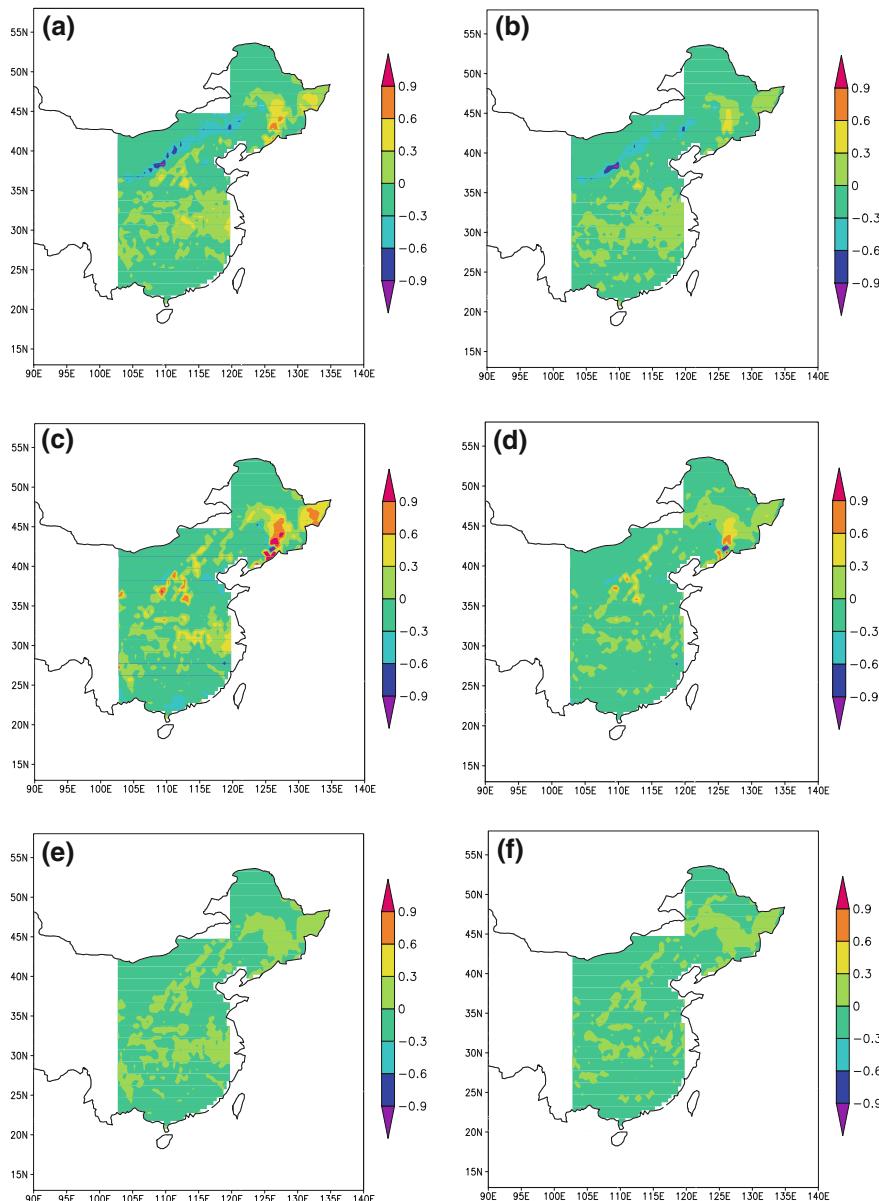
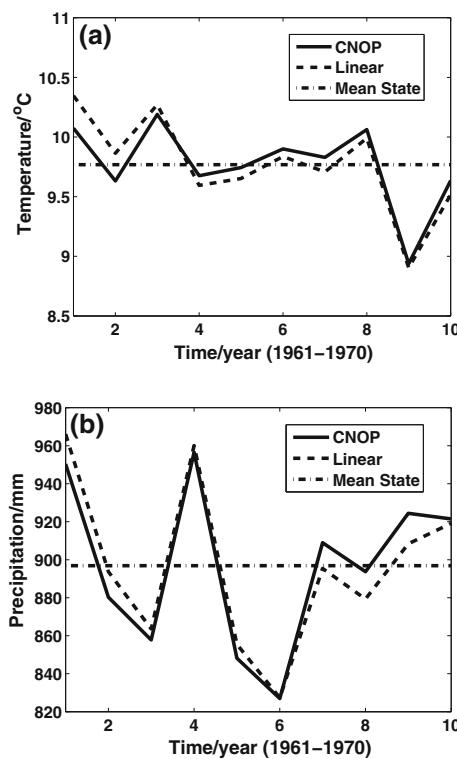


Fig. 9 Variation in the three components of the soil carbon density when compared with the reference state ($\text{Kg C m}^{-2} \text{ year}^{-1}$). The left column represents the CNOP temperature change, and the right column is the linear temperature change. **a** and **b** belowground litter; **c** and **d** the fast-decomposing soil carbon pool; **e** and **f** the slow-decomposing soil carbon pool (From Sun and Mu 2012)

Table 4 The absolute variation in the plant functional types (PFTs) because of the different types of temperature change ($\text{g C m}^{-2} \text{ year}^{-1}$). TeBE: temperate broad-leaved evergreen, TeBS: temperate broad-leaved summergreen, BoNE: boreal needle-leaved evergreen, TeH: temperate herbaceous (From Sun and Mu 2012)

PFTs	CNOP-P-type	Linear type
TeBE	247.06	141.74
TeBS	324.58	155.06
BoNE	351.14	187.04
TeH	514.74	455.98

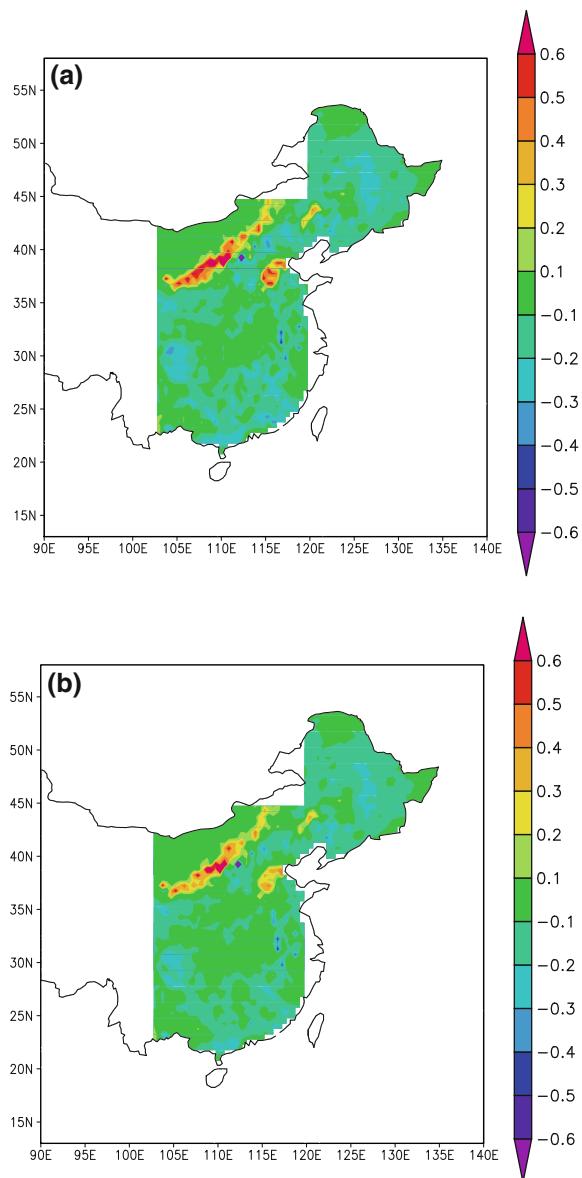
Fig. 10 The temporal variations in the climate conditions because of CNOP and linear perturbations. **a** temperature; **b** precipitation (From Sun and Mu 2012)



4.2 Variations in the Soil Carbon Resulting from Changes in Precipitation

Similarly, the variation of soil carbon due to the CNOP-P-type precipitation is also analyzed. Figure 10b shows the temporal precipitation resulting from CNOP-P-type perturbation. The variation in the soil carbon is minor in China due to the nonlinear and linear types of precipitation changes (Fig. 11). The numerical results imply that the response of the soil carbon to variations in the annual precipitation is minor.

Fig. 11 The variation in soil carbon density when compared to the reference state ($\text{Kg C m}^{-2} \text{ year}^{-1}$). **a** the CNOP precipitation change; **b** the linear precipitation change ($\text{Kg C m}^{-2} \text{ year}^{-1}$) (From Sun and Mu 2012)



5 Summary

In this chapter, we review the progress of the authors' research on variations in the terrestrial ecosystem due to climate change. First, the nonlinear response of a grassland ecosystem to climate change is considered using the CNOP-P approach. With sufficiently large finite-amplitude perturbations of its parameters, a grassland (desert) ecosystem with the CNOP-P-type climate change is nonlinearly unstable. And, the variation of the grassland (desert) ecosystem to nonlinear climate change is more intense than its response to linear climate change for the same amplitude. The abrupt change occurs for the grassland (desert) ecosystem with the CNOP-P-type climate change. During this transition, the moisture content of the soil in the root zone and the amount of wilted biomass play crucial roles in the grassland ecosystem (Sun and Mu 2011).

Second, to further examine the response of the soil carbon in China to climate change using a complex model, the LPJ model is employed. The difference between the CNOP-P-type and linear perturbations is the variation of the variability. The variations in the soil carbon density for CNOP-P-type temperature change are greater than those for the linear temperature change. In the Southern China, the soil carbon increases due to the CNOP-P-type temperature. The soils maybe play a role of sink due to the pools of quickly decomposing soil carbon. The intrinsic difference in the responses to the two types of temperature perturbations in southern China suggests that quickly decomposing soil carbon is sensitive to changes in the temperature variability. The sensitivity of the amount of soil carbon to different types of changes in the amount of precipitation is weak. This may be due to the absence of extreme precipitation events (Sun and Mu 2012).

6 Future Prospects

In the above reviews, the new climate change scenario called as the CNOP-P-type climate change scenario by authors is established. The CNOP-P-type climate change, which supplies a possible climate change scenario based on the reasonable climate change range, explores the nonlinear stability of the grassland ecosystem and maximal variations in the terrestrial ecosystem. This type of climate change not only considers the variation of climatology, but the variation of climate variability. However, the above studies just employed the representative climate change range, such as the increasing by 2 °C for the temperature mean state in whole study region. The regional and seasonal character of the climate change fails to be considered. The approach also could provide a way to investigate the maximal possible variations of the terrestrial ecosystem to take into account the regional and seasonal character of the climate change. The regional and seasonal character of the climate change could be evaluated in view of the GCMs model output (Pan et al. 2010). So, the CNOP-P approach could supply the possible climate change based on the

outputs from the GCMs to discuss the maximal response of terrestrial ecosystem to climate change.

On the other hand, variations in the amount of soil carbon as part of carbon cycle in a terrestrial ecosystem are explored. In addition to the amount of soil carbon, the net primary production (NPP) and the net ecosystem production (NEP) are important parts of the carbon cycle, particularly for estimating the amount of carbon sources and sinks. For example, Berthelot et al. (2005) used 14 oceanic and atmospheric general circulation models (OAGCMs) to estimate the impact of climate change on terrestrial carbon pools and fluxes. Although there were coinciding trends in the variation of the net ecosystem production (NEP), the standard deviation of these models was 2.7 Gt C yr⁻¹. And, the uncertainty of the amount of global carbon was between -73.9 and -6.7 Gt C yr⁻¹. In the future, it will be interesting to discuss the maximum uncertainty in the NPP and NEP due to CNOP-P-type climate change. The interaction between climate change and terrestrial ecosystem is a key issue. Our current studies just analyze the impact of climate change on the terrestrial ecosystem. More importantly, the feedback from the terrestrial ecosystem to atmosphere is worth to being discussed.

Appendix A

A linear dynamical system is introduced to help readers to understand the sensitivity of system to parameter using the CNOP-P approach (Eq. A1):

$$\begin{cases} \frac{dx}{dt} = -2x - 5y \\ \frac{dy}{dt} = -x - 3y + f(t) \end{cases} \quad (\text{A1})$$

x and y are state variables of the linear dynamical system. $f(t)$ is the parameter of the linear dynamical system and time-dependent. The Eq. (A1) could be discrete using the Euler scheme. The discrete scheme is

$$\begin{cases} \frac{x^{k+1} - x^k}{\Delta t} = -2x^k - 5y^k \\ \frac{y^{k+1} - y^k}{\Delta t} = -x^k - 3y^k + f^k \end{cases} \quad (\text{A2}).$$

The initial values of x and y are respectively 0.3 and 0.2, and the time step is 0.1, and the integral step number is 10. The parameter $f(t)$ is conveniently chosen as a constant ($f=0.31$) in Eq. (A1), and as the reference parameter. To explore the most sensitive response of system to temporal parameter $f(t)$, the parameter of each time step in discrete scheme is considered as the study objects with the CNOP-P approach. 10 unknown parameters (f^1, f^2, \dots, f^{10}) according to the integral step number, which are constrained by L2 norm and whose extent is $\delta=0.3$, need be computed using the CNOP-P method to discuss the sensitivity of system to parameters. Figure A.1 shows the reference parameter, and the CNOP-P-type

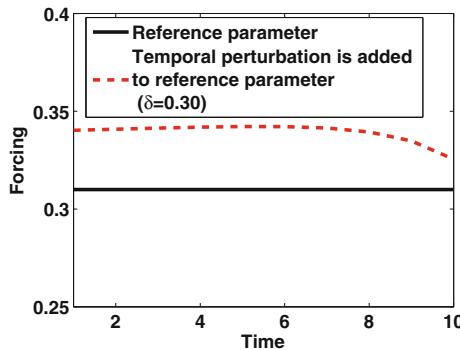


Fig. A.1 The reference parameter and the CNOP-P-type parameter

parameter that the temporal perturbation (Table A.1) obtained by the CNOP-P approach is added to the reference parameter. Figure A.2 shows the system evolvement due to the reference parameter and the CNOP-P-type parameter. The numerical results show there are larger variations of x and y due to the CNOP-P-type parameter compared to the reference parameter. This illustrates that the CNOP-P-type parameter could result in most unstable or maximal variations of the linear dynamical system. Although the linear dynamical system is applied in the Appendix, the CNOP-P approach could also be employed to discuss nonlinear dynamical systems, because the CNOP-P could be calculated without the restrictions to the linear or nonlinear dynamical system.

Table A.1 The temporal perturbation of parameter using the CNOP-P approach

The parameter	f^1	f^2	f^3	f^4	f^5	f^6	f^7	f^8	f^9	f^{10}
The perturbation value	0.098	0.100	0.101	0.103	0.104	0.104	0.101	0.095	0.080	0.005

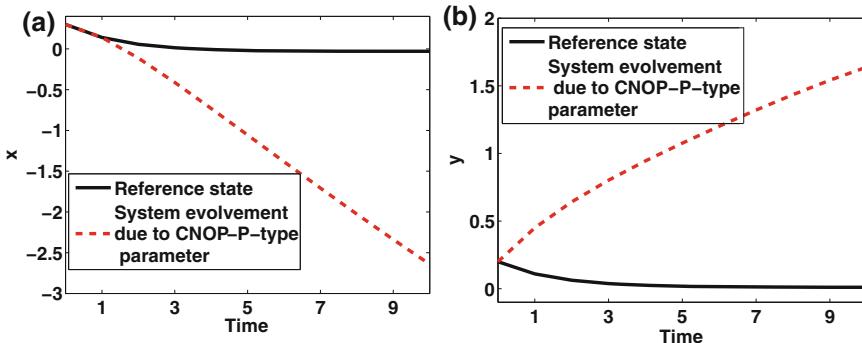


Fig. A.2 The evolvement of system due to the reference parameter and the CNOP-P-type parameter: **a** x ; **b** y

References

- Berthelot M, Friedlingstein P, Ciais P, Dufresne J-L, Monfray P (2005) How uncertainties in future climate change predictions translate into future terrestrial carbon fluxes. *Glob Chang Biol* 11:959–970. doi:[10.1111/j.1365-2486.2005.00957.x](https://doi.org/10.1111/j.1365-2486.2005.00957.x)
- Birgin EG, Martinez JM, Raydan M (2000) Nonmonotone spectral projected gradient methods on convex sets. *SIAM J Optim* 10:1196–1211
- Bonan GB, Levis S, Kergoat L, Oleson KW (2002) Land-scapes as patches of plant functional types: an integrated concept for climate and ecosystem models. *Glob Biogeochem Cycles* 16 (2):1021. doi:[10.1029/2000GB001360](https://doi.org/10.1029/2000GB001360)
- Botta A, Foley JA (2002) Effects of climate variability and disturbances on the Amazonian terrestrial ecosystem dynamics. *Glob Biogeochem Cycles* 16(4):1070. doi:[10.1029/2000GB001338](https://doi.org/10.1029/2000GB001338)
- Claussen M, Kubatzki C, Brovkin V, Ganopolski A, Hoelzmann P, Pachur HJ (1999) Simulation of an abrupt change in Saharan vegetation in the mid-holocene. *Geophys Res Lett* 26 (14):2037–2040
- Duan WS, Mu M (2006) Investigating decadal variability of El Niño-Southern Oscillation asymmetry by conditional nonlinear optimal perturbation. *J Geophys Res* 111:C07015. doi:[10.1029/2005JC003458](https://doi.org/10.1029/2005JC003458)
- Eglin T, Ciais P, Piao SL, Barre P, Bellassen V, Cadule P, Chenu C, Gasser T, Koven C, Reichstein M, Smith P (2010) Historical and future perspectives of global soil carbon response to climate and land-use changes. *Tellus B* 62(5):700–718. doi:[10.1111/j.1600-0889.2010.00499.x](https://doi.org/10.1111/j.1600-0889.2010.00499.x)
- Gao XJ, Luo Y, Lin WT, Zhao ZC, Filippo G (2003) Simulation of effects of land use change on climate in China by a regional climate model. *Adv Atmos Sci* 20(4):583–592
- Gerber S, Joos F, Prentice IC (2004) Sensitivity of a dynamic global vegetation model to climate and atmospheric CO₂. *Glob Change Biol.* 10(8):1223–1239. doi:[10.1111/j.1529-8817.2003.00807.x](https://doi.org/10.1111/j.1529-8817.2003.00807.x)
- Kicklighter DW et al (1999) A first-order analysis of the potential role of CO₂ fertilization to affect the global carbon budget: a comparison of four terrestrial biosphere models. *Tellus B* 51 (2):343–366. doi:[10.1034/j.1600-0889.1999.00017.x](https://doi.org/10.1034/j.1600-0889.1999.00017.x)
- Kharin VV, Zwiers FW, Zhang X, Hegerl GC (2007) Changes in temperature and precipitation extremes in the IPCC ensemble of global coupled model simulations. *J. Clim* 20:1419–1444. doi:[10.1175/JCLI4066.1](https://doi.org/10.1175/JCLI4066.1)
- Liu Z, Notaro M, Kutzbach J, Liu N (2006) Assessing global vegetation-climate feedbacks from observations. *J. Clim* 19:787–814
- Matthews HD, Weaver AJ, Meissner KJ (2005) Terrestrial carbon cycle dynamics under recent and future climate change. *J. Clim* 18:1609–1628
- Mitchell TD, Jones PD (2005) An improved method of constructing a database of monthly climate observations and associated high-resolution grids. *Int J Climatol* 25(6):693–712
- Mitchell SW, Csillag F (2001) Assessing the stability and uncertainty of predicted vegetation growth under climatic variability: northern mixed grass prairie. *Ecol Model* 139(2–3):101–121
- Mu M, Duan WS, Wang B (2003) Conditional nonlinear optimal perturbation and its applications. *Nonlinear Process Geophys* 10:493–501
- Mu M, Sun L, Dijkstra HA (2004) The sensitivity and stability of the ocean's thermohaline circulation to finite amplitude perturbations. *J Phys Oceanogr* 34:2305–2315
- Mu M, Wang B (2007) Nonlinear instability and sensitivity of a theoretical grassland ecosystem to finite-amplitude perturbations. *Nonlinear Process Geophys* 14:409–423
- Mu M, Duan WS, Wang Q, Zhang R (2010) An extension of conditional nonlinear optimal perturbation approach and its applications. *Nonlinear Process Geophys* 17:211–220. doi:[10.5194/npg-17-211-2010](https://doi.org/10.5194/npg-17-211-2010)

- Mu M, Duan WS, Wang B (2007a) Season-dependent dynamics of nonlinear optimal error growth and El Nino- Southern Oscillation predictability in a theoretical model. *J Geophys Res* 112: D10113. doi:[10.1029/2005JD006981](https://doi.org/10.1029/2005JD006981)
- Mu M, Wang HL, Zhou FF (2007b) A Preliminary application of conditional nonlinear optimal perturbation to adaptive observation. *Chin J Atmos Sci (in Chinese)* 31(6):1102–1112
- Mu M, Jiang ZN (2008) A new approach to the generation of initial perturbations for ensemble prediction: Conditional nonlinear optimal perturbation. *Chin Sci Bull* 53(13):2062–2068
- Ni J (2004) Estimating grassland net primary productivity from field biomass measurements in temperate northern China. *Plant Ecol* 174(2):217–234
- Notaro M, Liu Z, Williams JW (2006) Observed vegetation-climate feedbacks in the United States. *J. Clim* 19:763–786
- Pan Z, Andrade D, Segal M, Wimberley J, McKinney N, Takle ES (2010) Uncertainty in future soil carbon trends at a central U.S. site under an ensemble of GCM scenario climates. *Ecol Model* 221:876–881. doi:[10.1016/j.ecolmodel.2009.11.013](https://doi.org/10.1016/j.ecolmodel.2009.11.013)
- Piao SL, Fang JY, Zhou L, Tan K, Tao S (2007) Changes in biomass carbon stocks in China's grasslands between 1982 and 1999. *Glob Biogeochem Cycles*, 21, GB2002. doi:[10.1029/2005GB002634](https://doi.org/10.1029/2005GB002634)
- Prentice IC, Cramer W, Harrison SP et al (1992) A global biome model based on plant physiology and dominance, soil properties and climate. *J Biogeogr* 19:117–134
- Sitch S et al (2003) Evaluation of ecosystem dynamics, plant geography and terrestrial carbon cycling in the LPJ Dynamic Vegetation Model. *Glob Change Biol* 9(2):161–185
- Storn R, Price K (1997) Differential Evolution-a simple and efficient heuristic for global optimization over continuous spaces. *J Glob Optim* 11(4):341–359
- Sun GD, Mu M (2009) Nonlinear feature of the abrupt transitions between multiple equilibria states of an ecosystem model. *Adv Atmos Sci* 26(2):293–304. doi:[10.1007/s00376-009-0293-8](https://doi.org/10.1007/s00376-009-0293-8)
- Sun GD, Mu M (2011) Response of a grassland ecosystem to climate change in a theoretical model. *Adv Atmos Sci* 28(6):1266–1278. doi:[10.1007/s00376-011-0169-6](https://doi.org/10.1007/s00376-011-0169-6)
- Sun GD, Mu M (2012) Responses of soil carbon variation to climate variability in China using the LPJ model. *Theor Appl Climatol* 110:143–153. doi:[10.1007/s00704-012-0619-9](https://doi.org/10.1007/s00704-012-0619-9)
- Trumbore SE, Czimczik CI (2008) An uncertain future for soil carbon. *Science* 321:1455–1456
- Wang Q, Mu M, Dijkstra HA (2012) Application of the conditional nonlinear optimal perturbation method to the predictability study of the Kuroshio large meander. *Adv Atmos Sci* 29(1):118–134. doi:[10.1007/s00376-011-0199-0](https://doi.org/10.1007/s00376-011-0199-0)
- Zeng XD, Shen SSP, Zeng XB, Dickinson RE (2004) Multiple equilibrium states and the abrupt transitions in a dynamical system of soil water interacting with vegetation. *Geophys Res Lett* 31:5501. doi:[10.1029/2003GL018910](https://doi.org/10.1029/2003GL018910)
- Zeng XD, Zeng XB, Shen SSP, Dickinson RE, Zeng QC (2005) Vegetation-soil water interaction within a dynamical ecosystem model of grassland in semi-arid areas. *Tellus* 57B:189–202
- Zeng XD, Wang AH, Zeng QC, Dickinson RE, Zeng XB, Shen SSP (2006) Intermediately complex models for the hydrological interactions in the atmosphere-vegetation-soil system. *Adv Atmos Sci* 23(1):127–140
- Zobler L (1986) A world soil file for global climate modeling. NASA Technical Memorandum, 87802, NASA, Washington, D.C., pp 32

Index

Numbers

- 3D-Var, 75, 436, 437
3DVAR, 221, 223–226, 238, 265, 269, 271, 285–290, 292, 293, 295, 297, 301, 303–305, 309, 313, 315–321, 405, 406, 408, 411, 416, 418, 421, 425, 447
4dEnVar, 85, 88, 107, 108
4D-PSAS, 77, 78
4dVar, 83, 85, 89–100, 102, 103, 106–110
4D-Var, 35, 37–39, 42, 43, 45, 71, 77, 78, 122, 361, 362, 364, 368, 371, 374, 378, 379, 385, 444, 468, 483–486, 488, 490
4DVAR, 37, 83–99, 107–109, 262, 263, 265, 287, 320, 383, 389, 393, 398, 402, 405, 406, 409, 410, 412–414, 416, 418, 421, 423–425, 445, 446

A

- ACE, 446
ADCIRC-2D, 74
Adjoint-based method, 37, 75, 108, 485
Adjoint-free, 83–85, 88, 107
Adjoint model, 8, 22–24, 27, 31, 33, 34, 37, 39, 42, 48, 73, 74, 76, 87, 90, 356, 410, 467–473, 475, 480, 483–485, 488, 490, 491, 493, 496–498, 502–506, 508, 509, 515
Adjoint sensitivity, 361, 378, 488, 490, 515
ADM-Aeolus, 260, 266, 447
AIRS, 75, 362, 363, 368, 373–378, 445, 506, 508
AMSR-E, 198, 204, 207
AMSU-A, 75, 337, 341, 344, 345, 347, 349, 351, 353, 355, 356, 358, 430, 439, 442, 444, 506
AMSU-B, 337, 341, 344, 345, 347, 350, 351, 353–356, 358, 506
A priori/a posteriori, 75, 136, 211, 223, 292, 338, 340, 361–364, 368–370, 375, 376, 379
ASCAT, 198, 204

Assimilation of images, 49

- Atmosphere-chemistry data assimilation, 67
Augmented Lagrangian methods, 15
AXBT, 405, 408, 413, 417, 419–421, 425

B

- Background error covariances, 33, 36, 37, 107, 179, 320, 409, 429, 435, 438, 439, 446
Background error variance, 180, 190–192, 439
Bathymetry, 93, 242, 243, 247, 249–251, 411
Bed level, 242
Bias estimation, 443
Bottom topography mapping, 255, 256
Boundary condition, 6, 12–14, 20, 36, 41, 42, 59, 72, 89, 92, 141, 197, 231, 243, 246, 250, 269, 275, 293, 296, 309, 331, 342, 383, 388, 390, 406, 407, 412, 431, 449, 459, 469
Brightness temperature, 195, 197, 199, 210, 211, 337, 339, 340, 342, 344, 345, 347–349, 351–354, 356–358

C

- Calculus of variations, 2
Canadian Environment Service, 448
Carbon cycle, 60, 527, 528, 530, 532, 544
Cauchy problem, 5
CESM, 522, 523
CLSM, 206, 208
CMC, 430
CNOP, 513, 515–517, 520, 521, 523, 528, 529
CNOP-P, 527, 529, 530, 533, 534, 536, 538, 543, 544
COAMPS, 407, 408, 468–471, 473, 480
Coastal flow, 71, 72, 76, 78, 211, 394, 396, 405–407, 409, 473
Complex terrain, 219–222, 224, 226, 227, 229, 235, 237, 238

- Constrained optimization, 14, 15, 32, 115, 116, 186, 197, 445, 531
- Control of the boundary, 12
- Control variable, 2, 8–10, 14, 16, 17, 20, 33, 59–62, 65–68, 92, 110, 288, 294, 456, 459
- Correlated observation errors, 177, 180, 362
- Cost function, 8, 9, 12, 16, 17, 21, 26, 30, 32, 33, 35, 36, 40, 44, 63, 64, 71, 74, 83, 84, 87–89, 94, 95, 97, 100, 103, 107, 108, 179, 223, 246, 285, 286, 288, 289, 292, 294, 319, 340, 356, 445, 455, 457, 458, 462, 463, 530
- Coupled data assimilation, 55–57, 59–65, 67–69
- Coupled error covariance, 63, 65
- Coupled modeling systems, 55, 56, 58, 63, 66, 68
- Cross-component correlations, 61–68
- CRTM, 337, 339, 340, 343, 345, 352, 355–358
- D**
- DART, 84, 92, 93, 221, 223, 235
- DAS, 203, 361, 362, 364, 369, 370, 372, 378, 379, 437, 439
- Data analysis, 1, 297, 408
- Data assimilation
- for continental waters, 47
 - for plant growth, 48
 - in agronomy, 47
 - in medicine, 49
- Data thinning, 121, 122, 128, 134, 179, 395
- DFS, 455, 462
- DGVM, 530–532, 537
- Diffusive wave, 245, 246, 440
- Dirac's measures, 11
- Directional derivative, 9, 11–13, 16, 20, 340, 356
- Direct methods, 211
- DMSP, 75, 447
- Doppler radar, 286, 288, 296, 310, 327, 333, 335
- Dry adjoint, 62, 491, 493, 496, 497, 503, 506, 507, 509
- Duality methods, 14
- DWL, 259–263, 265, 266, 268, 269, 271, 272, 274, 276, 278
- Dynamical system, 32, 48, 121, 122, 126, 135, 141, 251
- Dynamic optimization, 126
- E**
- EAKF, 221
- ECMWF, 31, 34, 35, 37, 45, 87, 223, 266–269, 275, 276, 374, 430, 441, 483, 485–490, 499, 504, 509
- EKF, 199
- Empirical gramian, 122, 128, 130
- Energy norm, 371, 483, 485, 486, 503, 508, 509
- EnKF, 84, 198, 199, 202–205, 210, 212, 221, 223–226, 235–238, 287, 288, 320, 328, 445, 447
- Ensemble methods, 55, 83, 84, 108, 109
- ENSO, 513–515, 520, 522, 523, 528, 529
- Envisat, 102–105, 413, 446
- EOFs, 85, 94, 109, 180
- ERA-40ERA-interim, 223, 431, 438, 443, 444
- Error correlations, 199, 211, 361–371, 374–378, 410, 445
- Error covariance tuning, 37, 56, 58–68, 72, 177–182, 185, 186, 190, 192, 199, 223, 224, 285, 289, 292, 294, 319, 358, 363, 364, 408–410, 439
- Estuary modeling, 241, 248, 249
- Euclidean norm, 11, 370
- Euler-Larrange equations, 3, 7, 13, 21, 247
- Extended Kalman filter, 199, 241, 245, 246
- F**
- FNMOC, 407
- Forecast error covariance, 56, 59, 60, 63, 65, 66, 68, 182, 456
- Forecast sensitivity, 37, 75, 364, 365, 372, 376, 483–486, 492, 505, 508
- Forecast uncertainty, 142, 143, 199, 402
- Forward operator/observation operator, 16, 20, 32, 33, 45, 59, 62, 63, 72, 73, 89, 102, 135, 158, 178, 179, 183, 184, 187, 211, 224, 287–290, 292, 296, 297, 303, 315, 316, 319, 333, 340, 409, 414, 444, 457, 459, 486
- FSOI, 485, 486, 490, 505, 507, 508
- FSR, 361, 363–366, 369–373, 375–379
- G**
- Gain matrix, 178, 184, 224, 364, 486
- Gaussian/non-Gaussian error, 59, 62, 68, 86, 287
- Gauss-Newton method, 34
- GCMs, 528, 543
- GCV, 4
- Generalized cross-validation, 4
- GEONET, 383–385, 393–395, 402
- GEOS, 200, 362, 448
- GFS, 221, 271, 459
- GLOW, 262, 263
- GNSS, 75, 383, 384

- GPS, 46, 248, 249, 275, 384, 385, 394–396, 398–400, 402, 432, 442, 446, 506
- Gramian, 122, 128, 130, 133, 134, 136, 138
- Gravimetric method, 195, 197, 211
- Gravity wave drag (GWD), 435, 440, 444–446, 449, 483, 485, 489, 491, 492, 500, 504
- H**
- Heavy rainfall, 327, 383–386
- Hermite polynomial, 142, 144, 164, 165, 169, 171
- Hessian, 21, 22, 36, 43, 84, 87, 91, 107, 108, 110, 294, 362, 457
- HFA, 104–106
- Hilbert space, 8–10, 166, 170
- History matching, 46
- HWRF, 260, 261, 274, 275, 278
- Hybrid data assimilation, 37, 55
- HYCOM, 407, 409, 412, 415
- I**
- IASI, 75, 361–363, 368, 372–378
- I4D-Var, 77, 78
- Impact of industrial pollution, 4
- Incremental 4D VAR, 31, 32, 35, 77
- Information quantification, 57, 136
- Initial condition, 2, 4–6, 8, 10, 11, 14, 17, 20, 27, 29, 30, 36, 38, 41, 42, 44, 45, 55, 56, 59, 61, 72, 83, 90, 92, 94, 102, 125, 127–129, 141–145, 147, 149–151, 153, 154, 159, 162, 220, 233, 235, 269, 271, 275, 287, 305, 320, 328, 383–385, 394, 398, 402, 410, 424, 442, 462, 463, 467, 469, 483, 513–515, 517, 523
- Inner-loop, 32–35, 401, 488
- Inter-channel error correlations, 362, 363, 368, 375–378
- Inverse methods, 241–243, 255
- Iterative method, 9, 15, 36, 180
- J**
- JMA, 327, 329–331, 333, 335, 383–390, 394, 398, 402
- JMA-NHM, 328
- JNoVA, 383, 385, 389, 393–395, 398, 401
- JTWC, 264, 265
- K**
- Kalman-Bucy filtering, 4
- Kalman filter equations, 178
- Kalman gain, 124, 126, 178, 181, 186, 224, 364, 440, 486
- Kinematic wave, 245, 246, 250, 253, 254
- KLM, 513–515, 517–520, 523
- L**
- Lagrange multiplier, 14, 15
- Land-atmosphere coupling, 65
- Laplace equation, 12
- Large scale meteorology, 2
- LDAS, 198–201, 205–207, 209, 212
- Least-squares, 71, 74, 86
- LETKF, 328, 332, 334, 335
- Linearized model, 74, 88, 224, 294, 410, 412, 488, 490, 509
- LIS, 200
- Lorenz model, 370
- LPJ, 527, 530, 532, 539, 543
- LSMs, 197–199, 203–205
- LSV, 515
- M**
- MATERHORN, 219, 221, 222, 228–230, 235, 238
- Maximum likelihood estimation, 72, 180, 223
- Max-min problem, 128, 129
- Mesoscale meteorology, 2
- Meteo-France, 35, 45, 88
- Met Office, 37, 45, 88, 339, 362, 374, 430
- MHS, 75, 506
- Middle atmospheric dynamics, 443
- MIPAS, 446, 447
- MLEF, 65, 455, 457, 458, 463
- MLS, 76, 438, 446
- MODAS, 412, 415–417
- Moist adjoint, 483, 485, 491, 493, 495, 498, 502, 505–508
- Monte Carlo, 151, 154, 199, 249
- MOZART, 459
- Mutual information, 57–59
- N**
- NAM, 221, 231, 233, 442
- NASA, 46, 65, 197, 198, 200, 204–206, 209, 280, 362, 448, 458, 463
- NAVDAS-AR, 71, 72, 74, 75, 361–363, 371–376, 378
- NAVGEM, 72, 74–76, 361, 363, 371, 378, 407
- NAVOCEANO, 406, 409, 412, 413, 423
- NCEP, 221, 223, 231, 235, 260, 261, 271, 275, 303, 342, 459, 516
- NCODA, 405–408, 411, 412, 414, 417, 419, 421–423, 425
- NCOM, 71, 72, 74, 76, 78, 83, 85, 92, 93, 95, 96, 99, 405–407, 409–414, 417, 419, 423–425
- NCOM 4D-Var, 73, 76, 77
- NEMO, 87

- Nesting, 143, 342, 467, 468, 470, 471, 473, 480
 NLDAS, 211
 NOAA, 275, 285, 303, 319–321, 339, 341, 413
 NOGAPS, 75, 407, 412
 Non-orographic gravity wave, 435, 485, 488
 Nowcasting, 1
 NRL, 78, 264, 378, 406, 408, 420, 426, 468, 469, 480
 NWP, 31, 32, 35, 45, 75, 78, 121, 129, 135, 219, 220, 223, 259, 261, 266, 278, 285–289, 291, 305, 313, 319–321, 337, 338, 357, 362, 373, 379, 384, 467, 468, 483
- O**
 Objective function, 38, 39, 43, 130, 248, 486, 490
 Observability, 121–124, 126–128, 130, 131, 133–137, 255
 Observability optimization, 128, 135
 Observation error covariance, 33, 38, 43, 63, 72, 75, 86, 177–181, 185, 186, 192, 289, 361–366, 371, 372, 378, 409, 459
 Observation error variance, 180, 190–192, 290, 361, 362, 364, 375–378
 Observation impact, 74, 102, 121, 122, 124–126, 130, 131, 135, 136, 371, 372, 483, 485, 505, 507, 508
 Observation sampling error, 180
 Observation sensitivity, 124, 125, 372, 376, 379, 484, 485
 Observation space approach, 87, 178, 186, 408, 439
 OI, 99, 226
 OMI, 66–68
 ONR, 261, 263, 379
 Optimal control techniques, 8
 Optimal interpolation, 87, 103, 415
 Optimality system, 14, 16, 19, 21, 24–27
 Optimal sensor placement, 121
 Optimization, 83, 88, 96, 107, 109, 122, 128, 129, 135, 249, 516, 532, 533
 Optimization algorithms, 109, 532
 OSE, 371, 379, 484, 505, 516, 523
 OSSE, 221, 223, 259–261, 266, 268–270, 274, 275, 277, 278, 286, 287, 516, 523
 Outer-loop, 94, 95, 99
 Ozone observation, 67, 68
- P**
 Parameter estimation, 245, 445, 446
 Parametric FORTRAN, 73, 76
 Partial observability, 126
 PDE, 126
- Penalty function, 115, 116
 Penalty methods, 115
 Perturbation calculation, 470
 PFT, 530, 538, 539, 541
 Plant growth, 48
 Pointwise data, 1
 POM, 407
 Precipitation forcing, 205, 206, 208
 PWV, 383–385, 394, 395, 398, 400, 402
- Q**
 QPF, 263, 384
- R**
 Radar data assimilation, 285, 286, 288, 289, 305, 319
 Radial velocity forward operator, 289, 303
 Radiative effect of hydrometeors, 337, 339, 344, 345, 347, 352, 354, 357, 358
 R4D-Var, 77
 Reflectivity forward operator, 287, 288, 290, 316
 Relo NCOM, 405–407, 409, 412, 425
 Representation error, 177, 179–181, 183, 185, 187, 188, 190
 Representer, 71, 72, 74, 76–78, 362, 405, 409, 410
 Representer-based systems, 71
 Response function, 338, 352, 356–358, 474, 476, 480
 River modeling, 245, 255
 ROM, 72, 74, 77, 78
 ROMS 4D-Var, 71, 72, 77
 Root-zone soil moisture, 195, 196, 203, 204, 211
 RRTM, 229, 268, 270, 337, 339, 340, 342, 344, 355–358
 RTTOV, 337, 339, 341, 344, 352, 354, 356–358, 430
- S**
 SABER, 76, 437, 441, 447
 Saddle point of the Lagrangian, 14
 Saint-Venant equations, 22, 241, 245, 253, 256
 S4D-Var, 72
 Second order adjoint, 22, 28
 Sensitivity analysis, 362, 379
 Sensitivity gradient, 483, 485, 486, 491, 493, 500, 508
 Sensitivity patterns, 488, 496, 508, 509
 Sensor placement, 121
 SEOM-2D, 74
 Shallow water equations, 122, 127, 128, 130, 241, 253

- Shannon information theory, 57
Sherman-Morrison-Woodbury, 87, 117, 118
SMAP, 197, 198
Smoothing splines, 4
SMOS, 197, 198
SMOSREX, 203
SNAP, 448
Soil moisture data assimilation, 195, 199, 200, 207, 210, 212
Soil water content, 196, 197, 211
Space of observations, 87, 178, 184, 186, 198, 260, 408, 439
Space of the state variable, 9, 11, 26
SPARC, 431, 448, 449
Spatiotemporal scales, 59, 62, 66, 68, 401
SREM2D, 206–208
S-RIP, 431
SSH, 77, 78, 406, 408, 410, 412, 414–416, 418, 419, 422, 425
SSMIS, 75, 76, 447, 451
State space approach, 87, 88, 177, 184
State variable, 59, 199, 203, 206, 292, 340, 356, 445–447
Stochastic Collocation, 163
Stochastic Galerkin, 163
Stopping criterion, 103
Stratospheric and mesospheric data assimilation, 448
Stratospheric jets, 434
Strong constraint formalism, 14, 72, 77, 115, 293
Strongly coupled data assimilation, 56
SVAT, 203
SWH, 99–103, 106
SWOT, 242, 244
- T**
Tangent-linear approximation, 424, 488
Tangent linear model, 74, 76, 122, 129, 410, 412, 467, 469, 471
Target observations, 513, 521, 522
TC, 455, 457, 462, 463, 513–515, 523
- TE norm, 485, 493, 499, 502
Terrestrial ecosystem, 527–532, 537, 538, 543
TRMM, 198
Tropical cyclone, 261, 263, 267, 269, 270, 275, 384, 513
TVD, 329, 331
TVR, 327–329, 331, 334
TWiLiTE, 261
- U**
UAV, 261
Uncertainty quantification, 142, 143
Uncertainty reduction, 55, 462
Underwater morphology, 244
Underwater topography, 241, 242, 244, 255
- V**
Variational data assimilation, 56, 61, 62, 71, 72, 74, 75, 78, 203, 212, 221, 263, 265, 285–288, 319, 361, 362, 406
Variational formalism, 13
Variational methods in meteorology, 2, 5
VDA, 13, 15, 16, 47
- W**
WAM, 83, 100–102
Wave dynamics, 251, 253
W4D-Var, 71–74, 76
Weak constraint formalism, 7, 13, 15, 16, 73, 77, 115, 116, 285, 286, 288, 292, 295, 305, 319, 410
Weakly coupled data assimilation, 56
Weak solutions, 115
Wiener Chaos, 141, 143, 144, 149, 153
Wind profiles, 234, 259–264, 266, 268, 269, 274, 276, 277
WRF, 65, 219–223, 229–231, 233, 235, 237, 260, 261, 263, 265, 267, 269–271, 275, 278, 289, 290, 305, 320, 337, 339, 342, 357, 457, 463
WRF-Chem, 66, 67, 455, 457–459, 463