2012

Results of the DMIP 2 Oklahoma experiments

Michael B. Smith
NOAA/NWS, michael.smith@noaa.gov

Victor Koren
NOAA/NWS

Ziya Zhang
NOAA/NWS

Yu Zhang
NOAA/NWS

Seann M. Reed
NOAA/NWS

See next page for additional authors

Authors
Michael B. Smith, Victor Koren, Ziya Zhang, Yu Zhang, Seann M. Reed, Zhengtao Cui, Fekadu Moreda, Brian A. Cosgrove, Naoki Mizukami, Eric A. Anderson, and DMIP 2 Participants

Journal of Hydrology 418–419 (2012) 17–48

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Results of the DMIP 2 Oklahoma experiments

Michael B. Smith [a,*], Victor Koren [a], Ziya Zhang [a], Yu Zhang [a], Seann M. Reed [a], Zhengtao Cui [a], Fekadu Moreda [b,1], Brian A. Cosgrove [a], Naoki Mizukami [a], Eric A. Anderson [a], and DMIP 2 Participants [2]

[a] Office of Hydrologic Development, NOAA/NWS, Silver Spring, MD, USA
[b] Water and Ecosystems Management, RTI International, 3040 Cornwallis Road, Research Triangle Park, NC 27709, USA

ARTICLE INFO

Article history:
Available online 3 September 2011

Keywords:
Distributed hydrologic modeling
Model intercomparison
Rainfall–runoff
Hydrologic simulation
Soil moisture
Channel routing

ABSTRACT

Phase 2 of the Distributed Model Intercomparison Project (DMIP 2) was formulated primarily as a mechanism to help guide the US National Weather Service (NWS) as it expands its use of spatially distributed watershed models for operational river, flash flood, and water resources forecasting. The overall purpose of DMIP 2 was to test many distributed models with operational quality data with a view towards meeting NWS operational forecasting needs. At the same time, DMIP 2 was formulated as an experiment that could be leveraged by the broader scientific community as a platform for testing, evaluating, and improving the science of spatially distributed models.

This paper presents the key results of the DMIP 2 experiments conducted for the Oklahoma region, which included comparison of lumped and distributed model simulations generated with uncalibrated and calibrated parameters, water balance tests, routing and soil moisture tests, and simulations at interior locations. Simulations from 14 independent groups and 16 models are analyzed. As in DMIP 1, the participant simulations were evaluated against observed hourly streamflow data and compared with simulations generated by the NWS operational lumped model. A wide range of statistical measures are used to evaluate model performance on both run-period and event basis. A noteworthy improvement in DMIP 2 was the combined use of *two* lumped models to form the benchmark for event improvement statistics, where improvement was measured in terms of runoff volume, peak flow, and peak timing for between 20 and 40 events in each basin.

Results indicate that in general, those spatially distributed models that are calibrated to perform well for basin outlet simulations also, in general, perform well at interior points whose drainage areas cover a wide range of scales. Two of the models were able to provide reasonable estimates of soil moisture versus depth over a wide geographic domain and through a period containing two severe droughts. In several parent and interior basins, a few uncalibrated spatially distributed models were able to achieve better goodness-of-fit statistics than other calibrated distributed models, highlighting the strength of those model structures combined with their *a priori* parameters. In general, calibration solely at basin outlets alone was not able to greatly improve relative model performance beyond that established by using uncalibrated *a priori* parameters. Further, results from the experiment for returning DMIP 1 participants reinforce the need for stationary data for model calibration: in some cases, the improvements gained by distributed models compared to lumped were not realized when the models were calibrated using inconsistent precipitation data from DMIP 1. Event-average improvement of distributed models over the combined lumped benchmark was measured in terms of runoff volume, peak flow, and peak timing for between 20 and 40 events. The percentage of model-basin pairs having positive distributed model improvement at basin outlets and interior points was 18%, 24%, and 28% respectively, for these quantities. These values correspond to 14%, 33%, and 22% respectively, in DMIP 1. While there may not seem to be much gain compared to DMIP 1 results, the DMIP 2 values were based on more precipitation–runoff events, more model-basin combinations (148 versus 51), more interior ungauged points (9 versus 3), and a benchmark comprised of two lumped model simulations.

In addition, we propose a set of statistical measures that can be used to guide the calibration of distributed and lumped models for operational forecasting.

Published by Elsevier B.V.

* Corresponding author. Address: Office of Hydrologic Development, NOAA National Weather Service, 1325 East West Highway, Room 8382, Silver Spring, MD 20910, USA. Tel.: +1 301 713 0640x128; fax: +1 301 713 0963.
E-mail address: michael.smith@noaa.gov (M.B. Smith).
[1] Present address.
[2] See Appendix A.

0022-1694/$ - see front matter Published by Elsevier B.V.
doi:10.1016/j.jhydrol.2011.08.056

1. Introduction

The US National Weather Service (NWS) continues to implement spatially distributed hydrologic models (hereafter called distributed models) for river, flash flood, and water resources forecasting. Since the conclusion of the first Distributed Model Intercomparison Project (DMIP 1; Reed et al., 2004; Smith et al., 2004a), the NWS has implemented a distributed modeling capability for basin outlet forecasts (e.g., Jones et al., 2009) as well as for generating gridded flash flood guidance over large domains (Schmidt et al., 2007). Indeed, distributed models are now routinely applied for operational forecasting in many parts of the world including, for example, Italy (Rabuffetti et al., 2009), Taiwan (Vieux et al., 2003), and Egypt (Koren and Barrett, 1994).

A companion paper (Smith et al., this issue) explains the motivation for, and design of, the Oklahoma experiments in the second phase of the Distributed Model Intercomparison Project (DMIP 2). It also describes the test basins, naming conventions, and data. Some experiments from DMIP 1 are repeated, albeit with more consistent radar-based precipitation estimates and with data from more streamflow gauges at interior points. A notable addition to DMIP 2 is the experiment designed to evaluate soil moisture simulations.

1.1. Participating institutions, models, and submissions

Fourteen participating groups submitted simulations for analysis and discussion at a DMIP 2 workshop convened in September, 2007; Table 1 lists the participants and the major characteristics of their models. Two of the groups (University of Arizona and the Danish Hydraulic Institute) submitted simulations from two models so that 16 models were run in DMIP 2 including the NWS lumped model. The references in Table 1 and the other papers in this issue provide further background on the specific models. Several groups had not participated in DMIP 1 and vice versa. Table 2 lists the participating institutions for both projects.

As with DMIP 1, the level of participation varied. Some participants submitted all requested simulations, while others submitted only a subset. Table 3 lists the simulations submitted by each DMIP 2 participant. Seven groups submitted the full set of model streamflow simulations: LMP, OHD, NEB, ARS, CEM, VUB, and EMC. The University of Alberta at Edmonton (UAE) submitted simulations for the Blue River well after the September, 2007 workshop but near the January 2008, deadline for the recalibrated results; we therefore include their simulations as a valuable contribution to the DMIP 2 results. With the UAE contribution, there were 14 participating groups and 16 models.

1.2. Benchmarks

Three benchmarks (e.g., Seibert, 2001) were used to assess model performance. Observed hourly streamflow data from the US Geological Survey (USGS) were used as 'truth' and two lumped models were used to provide hydrologic model benchmarks. The two models were the Sacramento Soil Moisture Accounting model (SAC-SMA; Burnash et al., 1973; Burnash, 1995), used by the NWS as its standard operational rainfall/runoff model (also used in DMIP 1), hereafter referred to as the LMP benchmark, and the GR4J lumped model contributed by the French participating group CEMAGREF, hereafter referred to as the CEM benchmark. In addition, to provide a combined benchmark for computing event improvement statistics, the LMP and CEM simulations were also averaged, hereafter called the LMP–CEM benchmark.

1.3. Definitions

As noted in the DMIP 1 discussion by Reed et al. (2004), there is no widely accepted definition of spatially distributed hydrologic modeling in the literature. For consistency with the DMIP 1 study, we therefore adopt the Reed et al. (2004) definition that a distributed model is one that (1) explicitly accounts for spatial variability of meteorological forcings and basin physical characteristics and (2) has the ability to produce simulations at interior points without explicit calibration at those points; please see Kampf and Burges (2007) for a detailed discussion of definitions and classifications regarding distributed hydrologic models.

Further, a parent basin is defined as a watershed for which explicit calibration is performed using basin outlet observed streamflow data. In our experiments, these parent basins represent the typical watershed sizes for which forecasts are generated by the NWS River Forecast Centers (RFCs). Interior points are locations within the parent basins where simulations are generated without explicit calibration. Hereafter, these are referred to as 'blind' simulations. Smith et al. (this issue) provide specific instructions on how the simulations for the parent basins and interior points were generated.

1.4. Calibration

Participants were free to calibrate their models using strategies and statistical measures of their choice, this process usually being model-dependent. As such, DMIP 2 simulations reflect participants' familiarity with the parameterization and calibration schemes of their models. Appendix B presents a summary of the calibration procedures used by the DMIP 2 participants. Additional information on parameterization and calibration strategies can be found in the other papers in this Special Issue.

An initial set of calibrated and uncalibrated simulations was submitted for analysis and review at the September, 2007 DMIP 2 workshop in Silver Spring, Maryland (MD). During this workshop, there was considerable discussion on statistical measures deemed appropriate by the NWS for model calibration. In particular, a decision was made to avoid using the Nash–Sutcliffe Efficiency statistic (NSE: Nash and Sutcliffe, 1970) for the evaluation of streamflow simulations, given that it summarizes model performance relative to an extremely weak benchmark – the observed mean output (Schaefli and Gupta, 2007) – and has no basis in underlying hydrologic theory (Gupta et al., 2008). Further, Gupta et al. (2009) and Martinez and Gupta (2010) have subsequently shown that use of the NSE does not ensure that a model is constrained to reproduce the mean and variability of the observed data; theoretical decomposition shows that the modeled water balance can be incorrect and the flow variability severely underestimated even though the NSE performance may be very high; for additional comments on NSE see Jain and Sudheer (2008) and Michel et al. (2006). Discussions consequently focused on the general concepts of fitting the shape, volume, and timing of observed hydrographs, and eventually the participants charged the NWS team with specifying three corresponding statistical calibration measures deemed important by the NWS for operational forecasting. It was suggested that such measures would be of great interest to the scientific community. In light of the specified statistical measures (listed below) participants were given the opportunity to submit recalibrated model simulations by the end of January, 2008. The University of Arizona and Wuhan University submitted recalibrated simulations.

In collaboration with the DMIP 2 participants, the NWS team selected the following two sets of statistical measures: see Appendix A of Smith et al. (2004a) for the equations.

Table 1
Participating groups and models in the DMIP 2 Oklahoma Region experiments. Note that some participants submitted simulations from more than one model.

Participant and acronym	Modeling system name	Primary reference(s)	Primary application	Spatial unit for rainfall–runoff calculations	Rainfall–runoff/vertical flux model	Channel routing method
Agricultural Research Service (ARS)	SWAT	Arnold and Fohrer (2005)	Land management/agricultural	Hydrologic response unit (HRU) (6–7 km^2)	Multi-layer soil water balance	Muskingum
University of Arizona (AZ1)	DHM-UA	Pokhrel et al. (2008)	Streamflow forecasting	16 km^2 grid cells	SAC-SMA	Muskingum
University of Arizona (AZ2)	HL-RDHM	Koren et al. (2004)	Streamflow, water resources forecasting	16 km^2 grid cells	SAC-SMA	Kinematic wave
Danish Hydraulics Institute (DH1)	Mike 11	Butts et al. (2004)	Forecasting, design, water management	Subbasins (~150 km^2)	NAM	Full dynamic wave solution
Danish Hydraulics Institute (DH2)	Mike SHE	Butts et al. (2004)	Forecasting, design, water management	Grids	Various	Various
Environmental Modeling Center (EMC)	NOAH Land Surface Model	<http://www.emc.ncep.noaa.gov/mmb/gcp/noahlsm/README_2.2.htm>	Land–atmosphere interactions for climate and weather prediction models, off-line runs for data assimilation and runoff prediction	~160 km^2 (1/8th degree grids)	Multi-layer soil water and energy balance	Linearized St. Venant equation
CEMAGREF (CEM)	GR4J	Perrin et al. (2003)	Streamflow forecasting	Lumped		Unit Hydrograph
NWS Office of Hydrologic Development (OHD)	HL-RDHM	Koren et al. (2004)	Streamflow, water resources forecasting	16 km^2 grid cells	SAC-SMA modified with heat transfer component for frozen ground effects (Koren et al., 2006, 2007)	Kinematic wave
University of Oklahoma (UOK)	Vflo™	Vieux (2004)	Streamflow forecasting	1 km^2 or smaller	Event based Green–Ampt infiltration	Kinematic wave
Imperial College of London (ICL)	Semi-distributed	Moore (1985)	Streamflow forecasting	Semi-distributed	Probability distributed soil moisture	
U. Nebraska at Lincoln (NEB)	HSPF	Bicknell et al. (1997) and Ryu (2009)	Streamflow and water quality forecasting	Sub-basins	Conceptual	Muskingum
Wuhan University (WHU)	LL-III	Li (2001a,b)	Streamflow and water resources forecasting	4-km grid	Multi-layer finite difference model	Full dynamic wave solution
U. Illinois (ILL)	THREW	Tian et al. (2006)	Streamflow forecasting	Sub basin REWs		
U. California at Irvine (UCI)	Semi-distributed SAC-SMA	Khakbaz et al. (this issue)	Streamflow forecasting	Sub basin (avg. size ~100 km^2)	SAC-SMA	Kinematic wave
U. Alberta, Edmonton (UAE)	DPHM-RS	Biftu and Gan (2001)	Streamflow forecasting	Sub-basin	Multi-layer water and energy balance	Muskingum–Cunge
Vrije U. Brussels (VUB)	WetSpa	Liu and De Smedt (2004)	Streamflow and water resources forecasting	50 m grid	Root zone soil water balance	Kinematic wave

Table 2
Comparison of participants in DMIP 1 and 2.

DMIP 1	DMIP 2
	U. Alberta at Edmonton Canada (DPHM-RS)
Agricultural Research Service (ARS) SWAT	ARS (SWAT)
U. Arizona (SAC-SMA)	U. Arizona (HL-RDHM and DHM-UA)
	CEMAGREF (GR4J)
	U. of California at Irvine (Semi-distributed SAC-SMA)
DHI Water and Environment (DHI) Mike 11	DHI (Mike 11 and Mike SHE)
NOAA Environmental Modeling Center (EMC) (Noah LSM)	EMC (Noah LSM)
Hydrologic Research Center (HRCDHM)	
Mass. Institute of Technology (tRIBS)	
	U. Illinois (THREW)
	Imperial College of London
	U. Nebraska (HSPF)
NWS Office of Hydrologic Development (OHD) HL-RDHM and Lumped SAC-SMA	OHD HL-RDHM and Lumped SAC-SMA
U. Oklahoma (r.water.fea)	U. Oklahoma (Vflo™)
U. California at Berkeley (VIC)	
Utah State University (TOPNET)	
	Vrije U. Brussels (WetSpa)
U. Waterloo Ontario (WATFLOOD)	
Wuhan U. China (LL-II)	Wuhan U. China (LL-III)

1.4.1. Overall run period measures (computed over calibration, verification, or combined periods)

(i) Modified correlation coefficient, r_{mod} (McCuen and Snyder, 1975; Smith et al. (2004a)). This measure was included to provide consistency with the DMIP 1 results shown in Reed et al. (2004).
(ii) %Bias (Smith et al., 2004a)
(iii) Squared error measure like RMSE to emphasize high flow prediction. This could be %RMSE. Units could be in $m^3 s^{-1}$ or mm/h. The latter allows for analyses to be independent of basin size.
(iv) Long term mass balance index. For this we generated comparison plots of runoff coefficient versus P/PE ratio.

1.4.2. Specific event measures

(i) r_{mod}.
(ii) Volume error.
(iii) Peak time error.
(iv) Peak flow error.

It is worth noting that it was very difficult to suggest only three statistical measures to be used for model calibration, either lumped or distributed. A number of other measures such as thresholds, false alarms, flow duration curves, probability of detection (POD) and critical success index (CSI) were also proposed and briefly discussed. The reason for this difficulty is that the NWS (as do other institutions and agencies) uses different measures at different times during the process of parameter calibration (e.g., Turcotte et al., 2003; Anderson, 2002; Smith et al., 2003; Hogue et al., 2003). Measures used in the calibration process are often designed to help make decisions about changes to specific parameters but are not necessarily a reflection of the overall model performance. For example, in the early stages of calibrating the NWS models, overall and seasonal biases are examined. Flow interval statistics are examined to calibrate base flows, while for peak flows, a different set of statistics and adjustments is used. Consequently, the OHD team directed DMIP 2 participants to the paper by Anderson (2002) and Smith et al. (2003) for a description of the process and statistics used. For our analysis (reported here), these recalibrated simulations were used instead of those submitted for the September, 2007 workshop.

2. Results and discussion

We present our results in a cohesive progression (temporally and spatially) from 'general' to 'specific' in order to understand model performance and if models achieved good answers for the right reason (Kirchner, 2006). Section 2.1 presents a water balance analysis to get a general view of model behavior. Overall streamflow simulation results at basin outlets for the entire calibration/verification period are analyzed in Section 2.2. We take a closer look at streamflow simulations by examining specific events in Section 2.3, including distributed model improvement over lumped results in Section 2.4. Interior hydrograph simulations are discussed along the way to assess model performance at ungauged sites. Following these sections, calibration impacts are examined in Section 2.5 by first looking at the effect of precipitation consistency on calibration, and then examining the improvement in simulation accuracy gained by calibrating *a priori* model parameters. Interior processes are further examined via soil moisture simulations and routing for a subset of participants in Sections 2.6 and 2.7, respectively. The format of Reed et al. (2004) is followed as much as possible to provide consistency with the DMIP 1 results. As noted in DMIP 1 (Reed et al., 2004) it is impossible to present and discuss all of the analyses that were performed.

Table 3
Streamflow simulations submitted by DMIP 2 participants. The parent basins are listed in normal text, while the interior points within each parent basin are listed in italics. Values in the table are 'u' for uncalibrated and 'c' for calibrated simulations as called for in the DMIP 2 modeling instructions.

Model	BLUO2	*CONNR*	ELDO2	*DUTCH*	TIFM7	*LANAG*	*POWEL*	TALO2	*KNSO2*	*SPRIN*	*WSILO*	*CAVES*	*SLOA4*	*SAVOY*	*ELMSP*	SLOA4	*CAVES*	*ELMSP*	*SAVOY*
ARS	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
AZ1	u c	u c																	
AZ2	u c	u c	u c	u c															
CEM	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
DH1	c																		
DH2	c																		
EMC	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
ICL								u c	u c	u c	u c	u c	u c			u c	u c	u c	u c
ILL	u c	u c						u c	u c	u c	u c	u c	u c	u c	u c				
LMP	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
NEB	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
OHD	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
UAE	u c	u c																	
UCI																u c	u c	u c	u c
UOK	u c	u c						u c	u c	u c	u c	u c	u c	u c	u c				
VUB	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c	u c
WHU	u c	u c	u c	u c															

2.1. Mean annual water balance comparisons

A general assessment of model behavior can be gained by examining the water balance over a multi-year period. Similar to the evaluation of land surface models in recent experiments (e.g., Mitchell et al., 2004; Lohmann et al., 2004, 1998; Wood et al., 1998; Duan et al., 1996; Timbal and Henderson-Sellers, 1998; Shao and Henderson-Sellers, 1996), we investigated the ability of each model to correctly partition precipitation into runoff, evaporation, and losses. This is a new analysis compared to DMIP 1, and was requested by participants at the September 2007 DMIP 2 workshop.

In this evaluation, we computed the water balance quantities for each model using the general continuity equation:

$$\frac{dS}{dt} = P_{obs} - E - R_{model} - L \quad (1)$$

where S represents all the various water storages on the land surface including soil moisture, canopy storage, and storage in rivers, P_{obs} is observed mean annual basin-average precipitation in mm, E is evaporation in mm, L represents the intercatchment groundwater transfer (losses or gains) and R is runoff in mm depth over the basin. We computed these quantities on an annual basis over a multi-year period and assume that the change in storage over that period is equal to zero. Observed mean annual precipitation over the basin and computed runoff from each of the models was used to compute a budget-based estimate of evaporation E and losses L:

$$E + L = P_{obs} - R_{\mathrm{model}} \quad (2)$$

This analysis was conducted for three parent basins (ELDO2, TIFM7, and BLUO2) and one interior point (DUTCH, within ELDO2) as these had the largest complement of submissions. For each of seven calibrated models and each basin, we plot the value of $E + L$ computed using Eq. (2) versus the model computed value of runoff (Fig. 1): an observations-based estimate of $E + L$ is also shown. For clarity, Fig. 2a shows an enlargement of the Blue River and Fig. 2b shows an enlargement of the ELDO2, TIFM7, and DUTCH results for runoff/evaporation values of 350 and 800 mm/year, respectively. Each diagonal on the figures represents the partitioning of observed precipitation into computed runoff and evaporation (plus losses) for a basin, with the x and y intercepts equal to the value of the mean annual areal observed precipitation. From each diagonal, a model's plotting symbol can be projected to the x or y axis to yield that model's basin-averaged mean annual runoff or evaporation and losses, respectively. All models should plot on a single line with a -1 slope and x and y intercepts equal to the observed mean areal precipitation if they have the correct water budget. All models that have the same partitioning of water should plot at the same point.

The results indicate that all models partition precipitation reasonably well for the calibrated parent basins. Given that the BLUO2 basin has the largest spring in Oklahoma (which flows out of the basin; Osborn, 2009), it is perhaps surprising that all the models performed so well in terms of precipitation partitioning for this basin. Note, however, that three out of seven models performed poorly for the DUTCH basin inside ELDO2 (NEB, ARS, EMC). Not surprisingly, these three models had low values of r_{mod} and large %Bias statistics for the DUTCH basin (see subsequent discussion for Figs. 13a and 13b in Section 2.5.2).

2.2. Comparison of distributed and lumped model performance

One of the foremost science questions in DMIP 2 is "*Can distributed hydrologic models provide increased simulation accuracy compared to lumped models?*" This section provides a general assessment of the participant simulations. Following Reed et al. (2004), Duan et al. (2006), and others, this analysis presents results

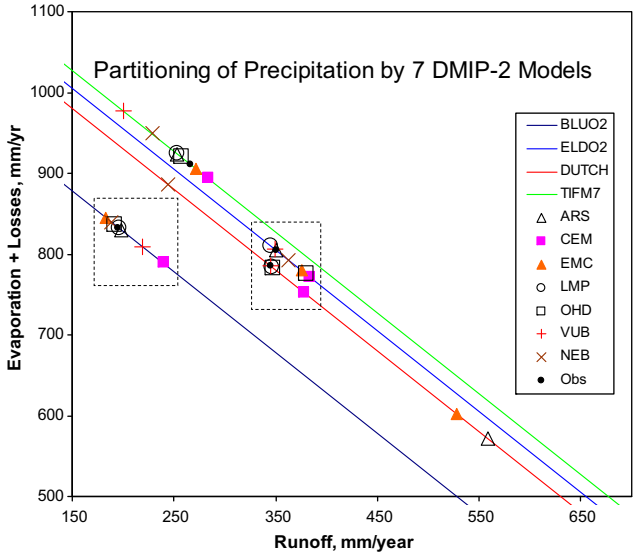


Fig. 1. Partitioning of observed precipitation into runoff and evaporation + losses for seven DMIP 2 models. The areas in the two boxes are enlarged in Figs. 2a and 2b.

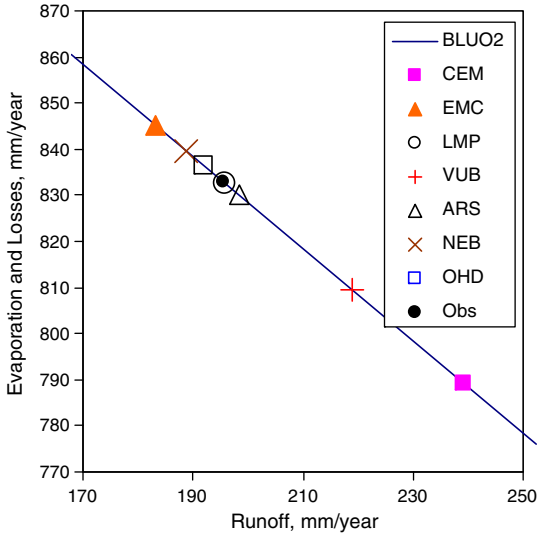


Fig. 2a. Same as Fig. 1 except the scale has been expanded for clarity around the observed points for BLUO2.

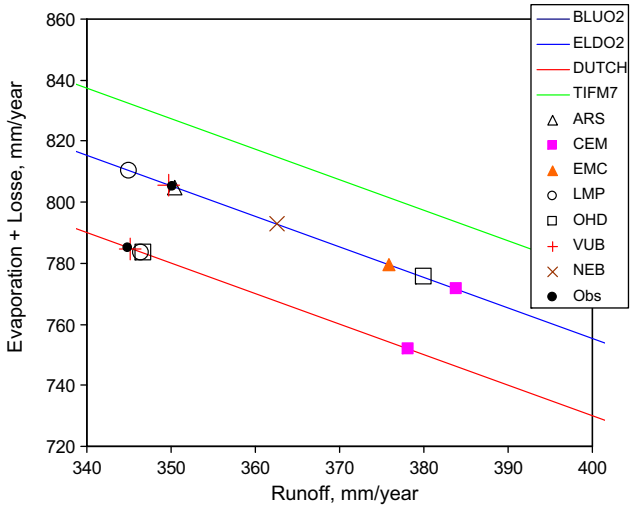


Fig. 2b. Expansion of Fig. 1 showing ELDO2, DUTCH, and TIFM7.

from the combined calibration and verification periods so as to provide a broad evaluation of model performance. Unless otherwise noted, hereafter the term 'overall statistics' is defined as those statistics computed for the entire calibration and verification period (see Section 1). Overall statistics are computed for calibrated and uncalibrated streamflow simulations at parent basin outlets and interior points.

Fig. 3 provides a general view of the distributed and lumped model results for each parent basin and the interior points. Interior point results are also plotted to address the questions: "*What is the potential for distributed models set up for basin outlet simulations to generate meaningful hydrographs at interior locations for flash flood forecasting?*" We plot the hourly overall modified correlation coefficient r_{mod} for calibrated simulations over the combined calibration and verification periods for the entire flow range. This measure was selected to correspond to the DMIP 1 results and is good for comparisons across basins. Following Reed et al. (2004), the parent basins and interior points are organized in order of increasing drainage area. The basins that are calibrated are usually the larger parent basins, and are plotted as positions 11–15. The interior basins (points where calibration was not explicitly performed) are plotted in positions one through ten. The median of the r_{mod} statistic for each calibrated model for each basin and similarly for each uncalibrated model is also shown; these were computed from the results provided by the seven groups that submitted a complete set of streamflow simulations.

Clearly, model performance tends to improve with basin size, in agreement with the findings of DMIP 2 (Reed et al., 2004), for both calibrated and uncalibrated models, and likely reflects the fact that there is greater uncertainty in the spatially averaged rainfall estimates for smaller basins. For example, Reed et al. (2007) found that peak errors and percent standard deviation of peak errors vary approximately linearly with the logarithm of the drainage area. Similarly, Carpenter and Georgakakos (2004) found the uncertainty in flow simulations from a distributed model increases in a well-defined manner as drainage area decreases. However, note also that explicit calibration at the interior points was not allowed, so that the basins in positions 11–15 should be expected to have better performance.

Note also the good overall calibrated performance of the first benchmark lumped model LMP. In 10 out of 15 cases, the LMP model ranks within the top two models. In three of the remaining four cases, the LMP model ranks above the median of the calibrated r_{mod}. The second benchmark lumped model (CEM) also performed well. In 13 out of 15 cases the CEM model ranks at or above the median of the calibrated r_{mod} statistic and amongst the top models. Taken together, one of the two benchmark (lumped) models ranked either highest or 2nd highest in all cases, reinforcing previous findings (e.g., Reed et al., 2004) that a calibrated lumped model can typically outperform a calibrated distributed model in terms of overall r_{mod} for these study basins.

In addition, we also show results for the *uncalibrated* OHD and LMP models. It is interesting that in many of the parent and interior basins, these uncalibrated models provide higher overall r_{mod} values than many of the calibrated distributed models. Other examples of this are presented in Section 2.5.

Basin-to-basin differences in performance are also revealed by Fig. 3. All models gave relatively poor performance at basin eight (Blue River at interior uncalibrated point CONNR); this is not surprising given that the basin contains several complications including sinkholes and gaining and losing sections, and that this area contains the largest spring in Oklahoma, (flowing northeast out of the Blue River basin) supplying water for the city of Ada, Oklahoma. Smith et al. (this issue) discuss these hydrogeologic complexities in more detail. On the other hand, remember that BLUO2 behaved well in terms of the multi-year-average partitioning of precipitation (Figs. 1 and 2).

Fig. 4 shows the overall %Bias statistic for each model for each basin, computed over the entire span of the calibration and verification periods. Parent basins and interior points are shown and are organized in order of increasing drainage area. Except for basin eight (CONNR), the median of the calibrated values is closer to zero than for the uncalibrated models. Notably, this holds even for the calibrated interior points. Again, hydrogeologic complexities in

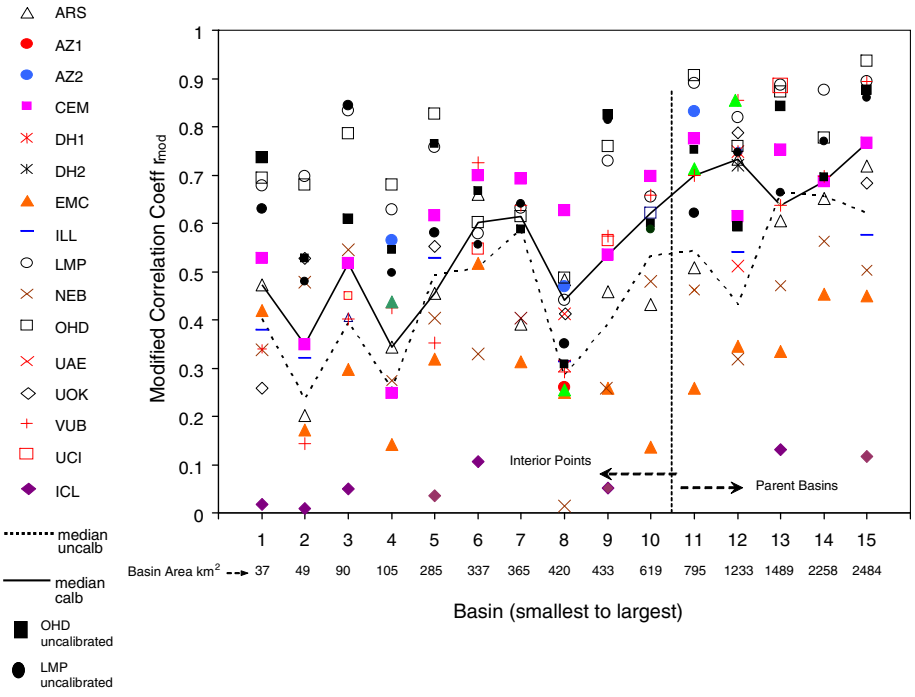


Fig. 3. Overall r_{mod} for all calibrated models and all basins. Results are for both calibration and verification periods. The solid line is the median of the calibrated models, while the dashed line is the median r_{mod} for the uncalibrated models. The uncalibrated r_{mod} for OHD and LMP for each basin are also shown for reference.

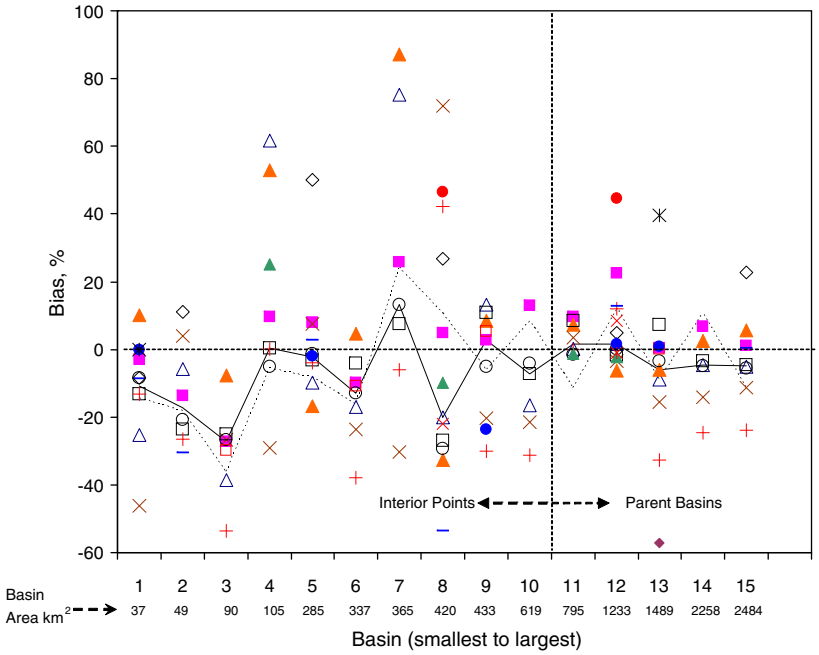


Fig. 4. Overall simulation %Bias for all calibrated models and all basins. Results are for both calibration and verification periods. The solid line is the median of the calibrated models, while the dashed line is the median %Bias for the uncalibrated model.

parent basin BLUO2 and its interior point CONNR appear to complicate the modeling process, with only two models (CEM and WHU) achieving an overall bias within ±10%. On the other hand, at basin 11 (ELDO2), all of the models were able to achieve a bias less than ±10%. Of the seven groups that generated simulations for all basins, no single model performed best in all cases in terms of overall %Bias.

Based on the overall r_{mod} and %Bias statistics computed for the interior points, the distributed models calibrated at the basin outlet exhibited a wider range of performance than did the benchmarks (the lumped models). In terms of overall r_{mod}, the calibrated models that performed well at parent basins also tended to perform well at interior points. At the smallest interior point basins (basins 1–5, and 9), only the OHD model gave equivalent or better overall r_{mod} values than the LMP model (this is not surprising in that the LMP and OHD models use similar parameterization schemes). For the larger interior point basins (6–10), several of the calibrated distributed models (UCI, VUB, and CEM) gave better r_{mod} performance than the LMP model. Further, several distributed models gave better (i.e., smaller) %Bias performance than the LMP model at outlets and interior points.

Table 4
Number of events used to for event statistics.

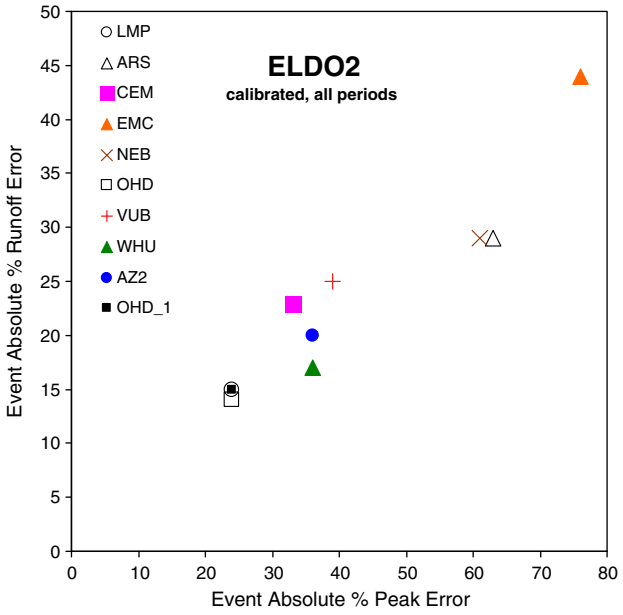
	Number of events used (calibration and verification periods)	
	DMIP 2	DMIP 1
BLUO2	41	24
CONNR	41	–
ELDO2	28	24
DUTCH	31	–
SLOA4	40	21 (WTTO2)
CAVES	23	
ELMSP	34	
SAVOY	40	
TALO2	25	21
KNSO2	29	20
CAVES	23	–
ELMSP	34	–
SAVOY	40	–
SPRIN	28	–
WSILO	30	–
TIFM7	42	24
LANAG	29	–
POWEL	35	–

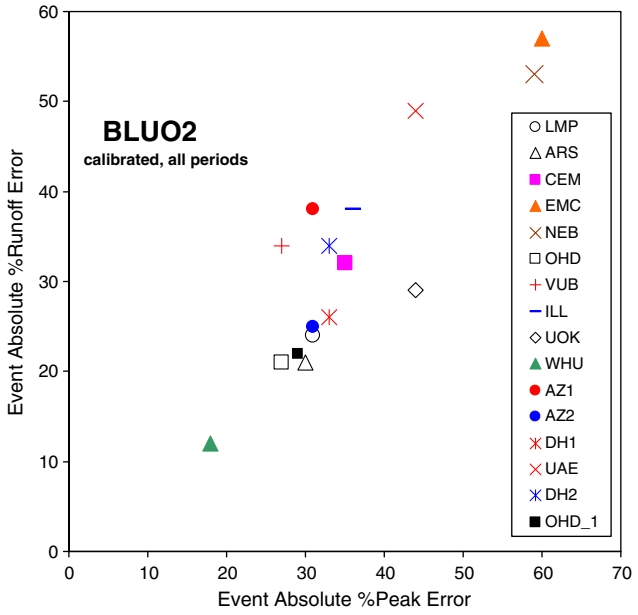
2.3. Event statistics

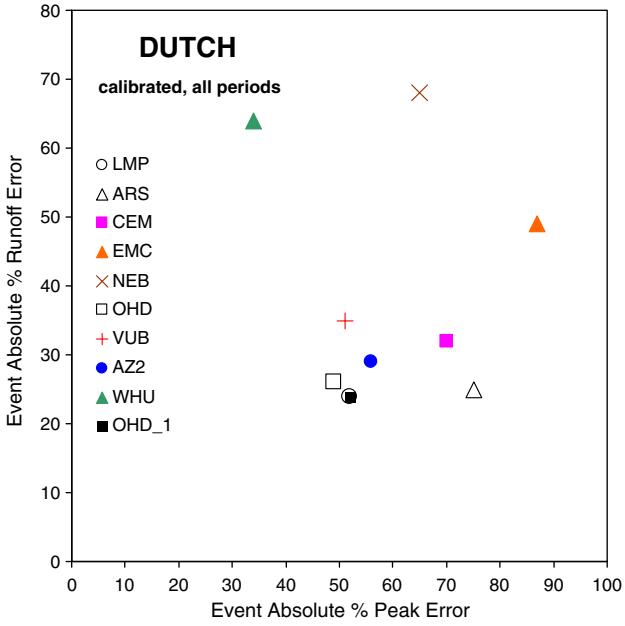
For a more in-depth view of model performance, goodness-of-fit statistics were computed for a number of specific rainfall/runoff events (see Table 4) as was done in DMIP 1 (Reed et al., 2004). In general, many more events were available for analysis than in DMIP 1. Events were selected from both the calibration and verification periods. As in the other DMIP 2 experiments, no state updating was allowed.

For this analysis, we computed values of two of the four metrics mentioned in the introduction and used in DMIP 1 (Reed et al., 2004): the event absolute percent peak error and event absolute percent runoff error (defined in Smith et al., 2004a). Figs. 5 and 6 present the averaged event absolute peak error plotted against averaged event absolute runoff error (averaging is across all events examined – see Table 4). The calibrated results for ELDO2 and BLUO2 (including their respective interior points) are presented as typical for all the study basins. Small values for both errors are desirable, plotting towards the lower left of each graph. Note that not every participating group submitted results for each basin. Appendix C presents the event results for the remainder of the study basins.

For the ELDO2 basin and its interior point at Dutch Mills (DUTCH), some models performed well in both cases and some models did not (Fig. 5). For ELDO2, the LMP, OHD, WHU, AZ2, CEM and VUB simulations plotted as a group within a range of 20–40% absolute peak error and between 10% and 25% absolute percent runoff error. The remaining models were less accurate in terms of both error statistics. At the DUTCH interior point (which







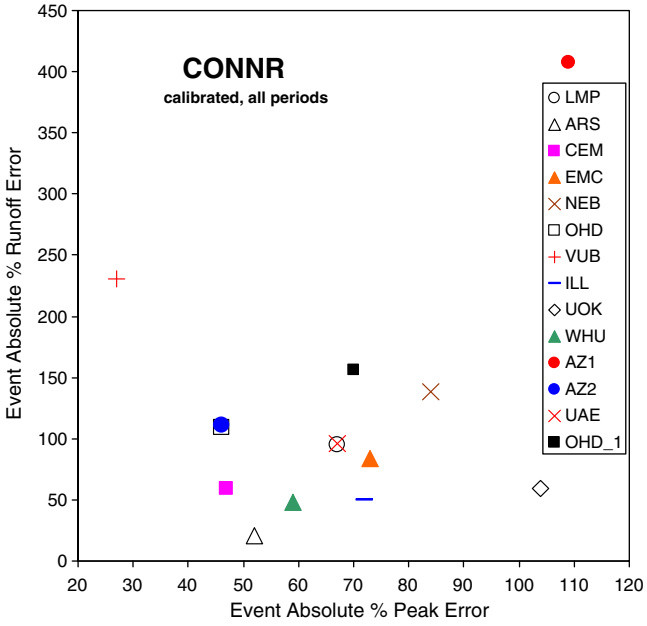


Fig. 5. Event-based statistics for calibrated models, ELDO2 (top) and the interior point DUTCH (bottom).

Fig. 6. Event-based statistics for calibrated models, BLUO2 (top) and the interior point CONNR (bottom).

streamflow data was not used for calibration), there is more spread in the model performance. The LMP, OHD, AZ2, and WUB results again cluster, but with slightly worse results compared to ELDO2. These models are joined by ARS, which improved in a relative sense compared to the best group. WHU did not perform as well due to the large runoff error of around 64%, even though peak error reduced from 36% to 34%. The NEB and EMC model results plotted the furthest from the optimal values. The OHD_1 simulation was derived by running the OHD DMIP 1 calibrated parameters with the DMIP 2 precipitation data (see Section 2.5.1). Comparing OHD and OHD_1, calibration using the biased 1993–1999 precipitation data did not affect combined peak error at ELDO2 or DUTCH, but resulted in a slight 1–2 point worsening in the combined runoff volume error.

Not surprisingly, the Blue River basin (BLUO2 and its interior point CONNR) shows less consistent results (Fig. 6) than do ELDO2 and DUTCH. WHU gives the best event statistics and stands alone among all the participating models. At CONNR, the model results are worse than at BLUO2 as seen by the large *x*- and *y*-plotting scales. Models that performed relatively well for BLUO2 did not necessarily do so at CONNR, underscoring the difficulty of modeling the Blue River basin. ARS and CEM results are only slightly worse at CONNR compared to BLUO2.

Looking collectively at the results, several calibrated lumped and distributed models that performed well at the parent basin outlet were also generally able to perform well at the interior locations for specific rainfall/runoff events, albeit at a slightly lower level of accuracy. This result suggests that these models effectively capture the fast responding portions of the hydrograph at both the parent basin outlet and interior points. The exception is the BLUO2, where the relative ranking of model performance is quite different at the basin outlet compared to the CONNR interior point.

2.4. Improvement of distributed over lumped models for specific events

Here, we address the question of the ability of distributed models to provide improvements over a lumped model, by investigating performance (in terms of volume, peak discharge, and peak timing statistics) on a number of rainfall/runoff events. The benchmark used here is the aggregated lumped model simulation derived by taking the average of the LMP and CEM values (this way, the results are more rigorous than if based on simulations from only one model as in DMIP 1). Hereafter, this standard is referred to as the LMP–CEM.

Figs. 7–9 present the event-average improvement statistics for volume, peak discharge, and peak timing, respectively, for calibrated models computed for the reference (combined calibration and verification) period using Eqs. (10–12) of Appendix A in Smith et al. (2004a). The number of events ranged from 23 (CAVES) to 42 (TIFM7) as shown in Table 4. In all three figures, the boxes on the abscissa labels denote parent basins and the interior points for each parent basin are plotted to the right of each box. Each plotting symbol represents an aggregate measure of the performance (for many events) of a specific calibrated model for a specific basin. The median value of distributed model performance for each basin is also plotted. The basins SLOA4, SAVOY, ELMSP, and CAVES are plotted twice, due to the modeling instructions that called for SLOA4 to be modeled as an independent headwater basin and then with SAVOY, ELMSP, and CAVES as interior points for TALO2. There are 148 model-basin pairs; this is nearly three times the number of pairs (51) for the same analyses in DMIP 1.

The average event volume improvement plot (Fig. 7) shows both improved and degraded performance by the distributed models. For all locations, the median improvement of the calibrated distributed models is negative. For clarity, we have limited the *y*-axis plotting to −80%, and so the values of −154% for CONNR by VUB and the value of −104% for LANAG by EMC cannot be seen. Distributed models realize positive gains over the LMP–CEM benchmark in 18% of the model-basin aggregate cases. These improvements are generally less than 10% except at BLUO2 and CONNR, where the improvements are better (15–30%). Corresponding to Reed et al. (2004), the improvement nearly doubles to ≃32% over the LMP–CEM benchmark when models having poor

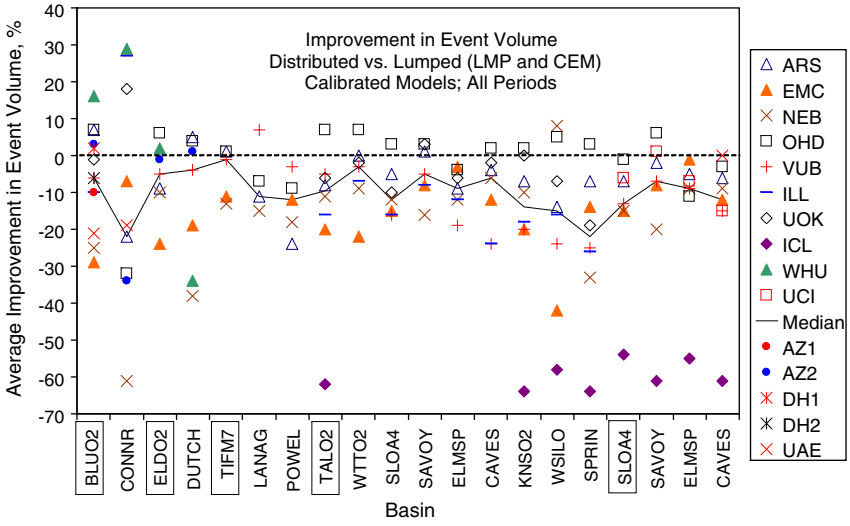


Fig. 7. Improvement in event volume: distributed versus lumped (LMP–CEM). Calibrated results for both the calibration and verification periods are shown. Parent basins names are denoted by boxes, while interior points in each parent basin are listed to the right of each box.

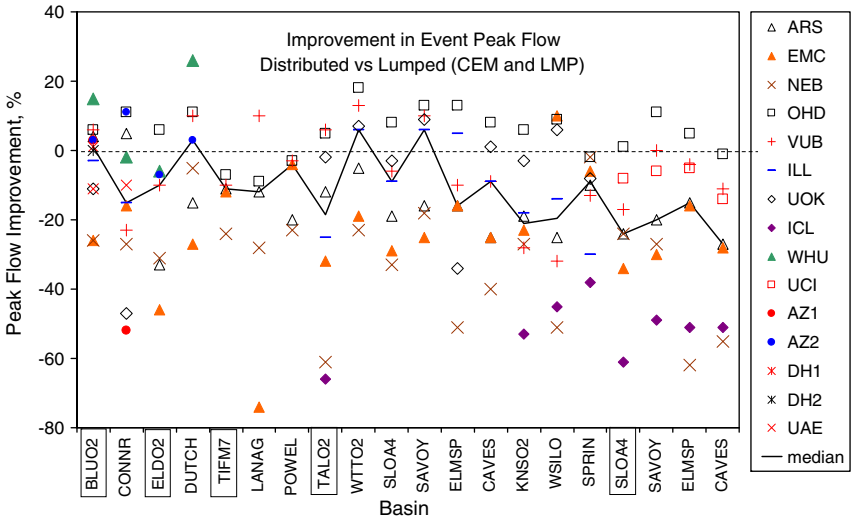


Fig. 8. Improvement in event peak flow: distributed versus lumped (LMP–CEM). Calibrated results for both the calibration and verification periods are shown. Parent basins names are denoted by boxes, while interior points in each parent basin are listed to the right of each box.

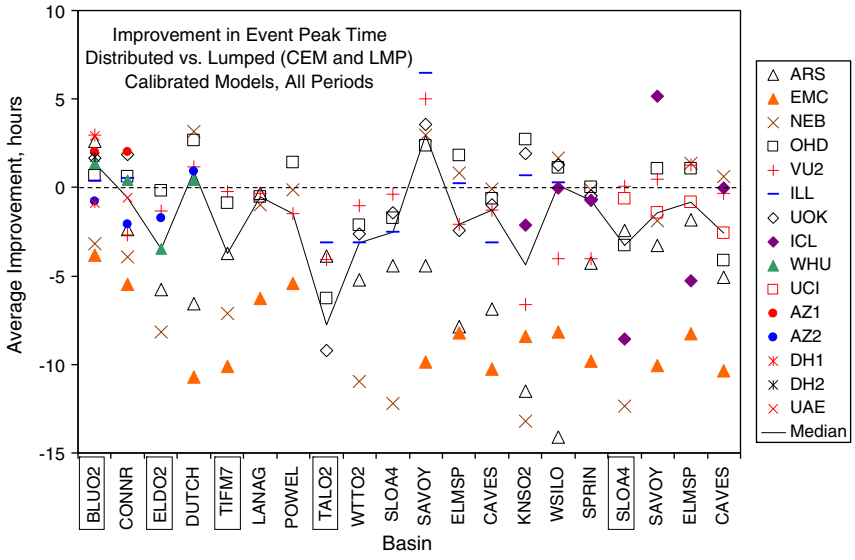


Fig. 9. Improvement in event peak time: distributed versus lumped (LMP–CEM). Calibrated results for both the calibration and verification period are shown. Parent basins names are denoted by boxes, while interior points in each parent basin are listed to the right box.

performance (performance values smaller than −5%) are excluded. In general, no difference in performance spread is noted between parent basins and their constituent interior basins.

An interesting comparison can be made between Fig. 7 and the equivalent (Fig. 15a of Reed et al., 2004) from DMIP 1. In DMIP 1, only the OHD model was able to provide any improvement over the lumped model in terms of runoff volume at the parent basins. Further, the improvement was less than 5%, and only for three parent basins. It is encouraging that in DMIP 2 a greater number of models were able to realize improvement over lumped models, and that this improvement extends to interior points as well.

Peak flow performance results (Fig. 8) are similar to event volume results, although larger improvements are seen here with 24% of the cases showing positive flood peak improvements greater than zero, and 36% of the cases show improvement greater than or equal to the −5% value used in Reed et al. (2004). The median performance for this statistic is positive in four cases. Basin BLUO2 received the greatest improvement, with eight models providing improved performance over the LMP–CEM benchmark. This result is not surprising given the long narrow basin shape and orientation.

Fig. 8 corresponds to Fig. 15b in the DMIP 1 results paper by Reed et al. (2004). Only the OHD model was able to provide improvement over the lumped model in all the headwater basins in DMIP 1, with two other models also providing improvements in the BLUO2 basin. Four models in DMIP 1 were able to provide noticeable improvements for one of the three interior points. Fig. 8 of the current paper shows that while there were many cases in which distributed models did worse that the LMP–CEM benchmark, there were also many more cases in DMIP 2 in which distributed models generated improved peak flow statistics compared to lumped models.

Fig. 9 examines the improvement in event-average peak timing. Other than a marginal improvement of 0.09 h for VUB in the SLOA4 headwater basin, noticeable peak time improvement due to distributed modeling is seen in only one of the parent basins (BLUO2); this result was also seen in DMIP 1. At least one distributed model achieved a peak timing improvement at each interior point (except at the TALO2 sub-basins WTTO2 and SLOA4) even though explicit calibration was not performed at those points. Considering parent and sub-basins, 28% of the model-basin pairs showed positive event-average peak timing improvement greater than zero. In 42% of the cases, even-average peak time improvements are greater than −1 h.

As in DMIP 1, the calibrated distributed models were able to provide improvements in peak hydrograph timing for only one (BLUO2) of the five parent basins. However, while only two DMIP 1 models were able to generate improved peak timing for BLUO2, nine models were able to do so in DMIP 2. Given that DMIP 2 included two lumped models, both using unit graphs, our results seem to support (for the basins other than BLUO2) the hypothesis of Reed et al. (2004) that physically-based routing schemes are more sensitive than unit hydrograph methods to errors in runoff depths given that velocities are dependent on flow rate. It is possible that in the elongated BLUO2 basin, the strength of the interaction of the spatial variability of rainfall with the flow network is greater than any errors in timing reported by Reed et al. (2004). Whereas in the other basins, the spatial variability of precipitation in the DMIP 2 forcing data was not great enough to cause large variability in peak hydrograph timing, or perhaps the stream network in these other basins is simply not very sensitive to spatial variation of precipitation (Smith et al., 2004b). These issues will require further investigation.

Among calibrated models, no single model gave improvements in all three event improvement statistics for all basins. Of the five groups that submitted a full set of simulations (resulting in 100 model-basin pairs in these analyses), the OHD model ranked the highest in terms of event volume improvement (by generating 13 out of the 27 instances of positive values) followed by the ARS model (4 of the positive values). Similarly, OHD ranked highest in peak flow improvement (with 15 of 36 cases of positive improvement) followed by VUB (6 cases). Finally, OHD also ranked highest in terms of peak time improvement (10 of 42 cases of positive improvement), followed by VUB (6 cases) and NEB (5 cases).

In general, lumped and distributed models that performed well at basin outlet points also performed well at interior points (albeit at a slightly degraded level of performance). This is especially true in terms of both overall calibrated r_{mod} statistic (Fig. 3) and the event statistics (Figs. 7–9). This finding has a stronger basis than DMIP 1, given that DMIP 2 had 9 gauged interior points with a broad range of drainage areas, compared to only three points in DMIP 1. The smallest sub-basin in DMIP 1 (Peacheater Creek at

Christie, OK) had a drainage area of 65 km^2) whereas two of the DMIP 2 interior points (SPRIN and WSILO) had even smaller drainage areas (37 km^2 and 49 km^2, respectively). Note that Reed et al. (2004) cautioned that some of the models in DMIP 1 had a coarse computational resolution compared to the size of the smallest basin and called for more studies on smaller, nested basins. Our results suggest that distributed models can provide reasonable simulations at interior locations over a relatively wide range of drainage areas. These results show promise for flash flood forecasting at interior locations where specific calibration cannot be performed.

2.5. Impact of parameter calibration

2.5.1. Calibration period results

Common problems associated with hydrologic model calibration include questions concerning the length and quality of the period of available data (e.g., Ghizzoni et al., 2007; Oudin et al., 2006; Brath et al., 2004; Andreassian et al., 2001, 2004; Gupta et al., 2003; Young et al., 2000; Bradley and Kruger, 1998; Gan et al., 1997; Xu and Vandewiele, 1994; Sorooshian et al., 1983). The specific DMIP 2 science question related to this issue is: "*What simulation improvements can be realized through the use of a more recent (i.e., higher quality) period of radar-based (i.e., multisensor) precipitation data than was used in DMIP 1? What is the impact of calibrating a distributed model with temporally inconsistent multisensor precipitation observations?*"

DMIP 1 participants identified the data period spanning 1993–1996 as being problematic (i.e., low bias; Reed et al., 2004; Guo et al., 2004; Young et al., 2000), and so DMIP 2 made available multi-sensor precipitation data for the subsequent period starting in 1996. Figs. 10a and 10b show the cumulative simulation errors for BLUO2 and SLOA4, respectively, which can be compared with the DMIP 1 results (Fig. 11 reproduced from Figs. 2a and 2b of Reed et al. (2004). Clearly, as in DMIP 1, not all participants placed a priority on minimizing simulation bias during calibration, with some of the simulations showing large positive or negative cumulative errors. However, in contrast to DMIP 1, the accumulated error plots (Figs. 10a and 10b) appear to be more homogeneous (relatively constant rate of error accumulation), suggesting both that the precipitation forcing is more consistent over time and that the accumulated errors could be reduced by parameter adjustments (as is standard practice for NWS calibration).

For the period April 2001–September 2002, all of the BLUO2 simulations indicate under-prediction, resulting mainly from a sequence of five under-simulated medium and large size flood events (December 18, 2001; February 1, 2002; February 20, 2002, March 20, 2002; and April 8, 2002). Possible causes may be anomalies in the precipitation or PE forcings, or errors in streamflow measurement. Illston and Basara (2002) and Illston et al. (2004) identified a severe drought in the summer of 2000 for the southwestern portion of Oklahoma, and commented that the precipitation during the following winter did not fully recharge the soil at the deepest layers. Further, Schneider et al. (2003) commented that in the western portion of the domain, soil moisture has a long 'memory' of precipitation deficits that can last one or more seasons. Nonetheless, it seems unlikely that all of the DMIP 2 models poorly simulated the drought conditions so as to generate poor storm hydrographs 1½ years later. Note that DMIP 2 participants were free to use any type of PE forcing in DMIP 2. To derive the

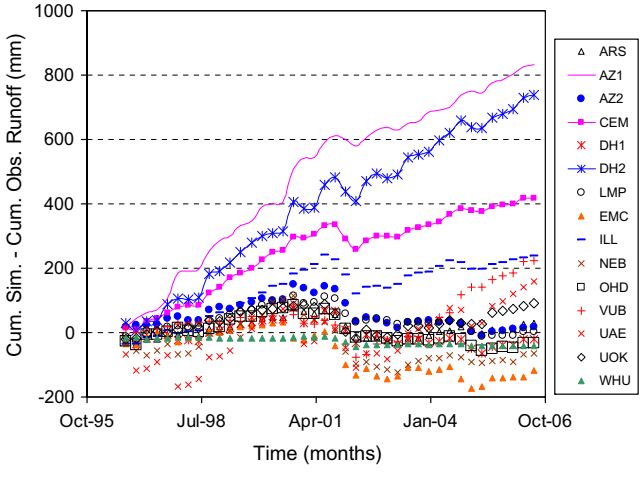


Fig. 10a. Accumulated simulation error in mm for BLUO2.

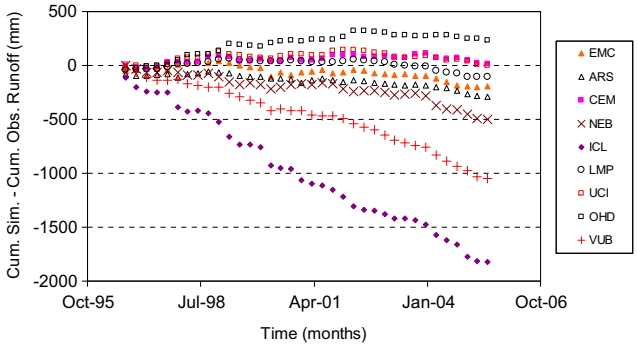


Fig. 10b. Accumulated simulation error in mm for and SLOA4.

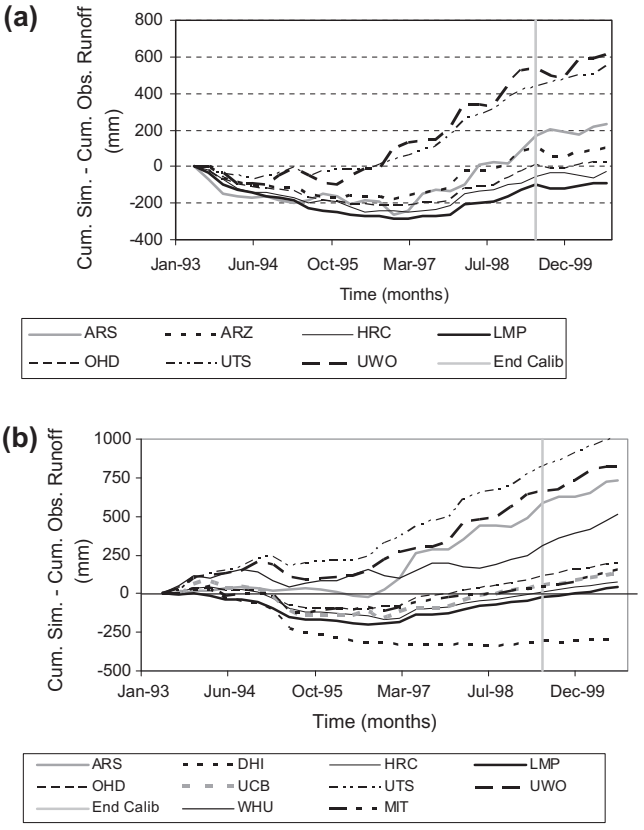


Fig. 11. DMIP 1 accumulated streamflow simulation error: BLUO2 (a, top) and WTTO2 (b, bottom).

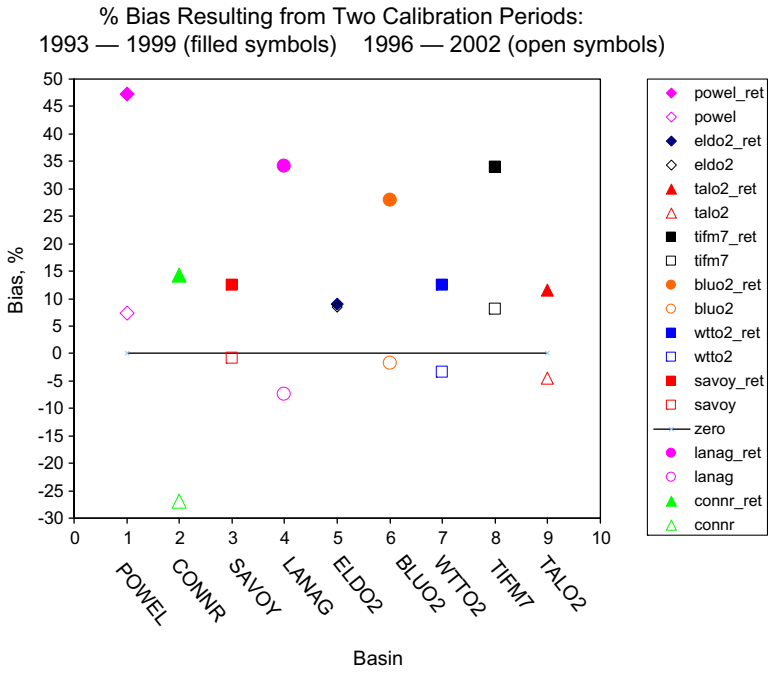


Fig. 12. % Simulation bias resulting from calibration using the 1993–1999 period versus the 1996–2002 period of NEXRAD-based precipitation. Results are for the OHD and OHD_1 models. The filled symbols represent the OHD_1 results calibrated using the 1996–2002 DMIP 1 data.

time series of daily PE forcing, some used the climatic monthly means (e.g., OHD, LMP), while others (AZ1, AZ2, ILL, and NEB) used the North American Regional Reanalysis data set (NARR, Mesinger et al., 2006). ARS estimated PE using NCDC daily temperature data with a Priestly–Taylor formulation and UOK estimated PE using observations from the Oklahoma Mesonet. Given that all the simulations show a similar trend (while using different PE forcing data sets), it is likely that the cause of the under-simulation is a period of anomalous precipitation or problems with streamflow measurement.

DMIP 1 specified the period 1993–1999 (hereafter called 'biased' period) for model calibration whereas DMIP 2 used the period 1996–2002 (hereafter called 'unbiased' period); note that there is a 3-year overlap between the two calibration periods. Fig. 12 shows the results of an experiment designed to highlight differences caused by these two periods on the calibration results. The models were calibrated using each period and then used to simulate the reference period from 1996 to 2006. This experiment fits into the category of 'dynamic sensitivity studies' as defined by Andreassian et al. (2004) and later by Oudin et al. (2006), in which reference calibration and corresponding reference simulation periods are specified, using a reference precipitation data set (in our case, the unbiased data) and then model recalibration is performed using the biased data for comparison with the reference simulation.

Only OHD provided simulations for this experiment (UCI provided simulations only for WTTO2, which was not a calibrated basin for DMIP 2) so the results are not as comprehensive as we would like. In the following, we use the term OHD to refer to the reference simulation and OHD_1 to refer to the simulation using biased data. For all cases where explicit outlet calibration was allowed (i.e., parent basins ELDO2, BLUO2, TALO2, and TIFM7), calibration using the biased data period generally resulted in larger runoff biases for the reference simulation period 1996–2006 than when using the unbiased data period (Fig. 12). Calibration (at the parent basin outlets) on the biased period also generally resulted in larger runoff biases for the interior points (with the exception of CONNR). While the difference in runoff bias at ELDO2 is barely noticeable compared to BLUO2, TALO2, and TIFM7, visual examination of the OHD and OHD_1 hydrographs showed that the majority of the small and intermediate events were over-predicted while the large events underpredicted, so that the ELDO2 values presented in Fig. 12 do not represent the entire picture.

Table 5
Number of instances in which OHD_1 %Bias is greater than calibrated DMIP 2 model %Bias for the overall 1996–2006 period. Interior points are shown indented below the parent basin.

Basin	OHD_1 %Bias Worse than
ELDO2	8 of 9
DUTCH	3 of 9
BLUO2	13 of 15
CONNR	5 of 13
TALO2	7 of 10
SAVOY	5 of 8
KNSO2	7 of 10
TIFM7	7 of 7
LANAG	6 of 7
POWEL	5 of 7

To put these %Bias values in context, we count the number of instances that the %Bias for the OHD_1 model is greater in absolute value than the calibrated %Bias statistics from the other models (Table 5). The OHD_1 simulation is seen to be worse than the majority of the participants' simulations from the calibrated parent basins.

The choice of calibration period had similar impacts on the r_{mod} statistic. In eight out of ten cases, calibration on the unbiased period improved the r_{mod} value for the entire reference simulation period (see the italic values in Table 6; for reference, we present the r_{mod} value for the LMP model). In four cases (ELDO2, DUTCH, CONNR, and SAVOY), the improvement in r_{mod} gained by the OHD model over LMP was lost when calibrating using the biased data period. We also examined the impact of the calibration period on flood events; mixed impacts on hydrograph peak and volume errors were found.

Table 6
Overall r_{mod} statistic for the 1996–2006 period from the OHD model calibrated on two periods: OHD_1 (1993–1999) and OHD (1996–2002). LMP results shown for reference.

	r_{mod}		
	OHD_1 1993–1999	OHD 1996–2002	LMP 1996–2002
ELDO2	.881	*.907*	.891
DUTCH	.621	*.680*	.628
BLUO2	.748	*.760*	.819
CONNR	.268	*.486*	.440
TALO2	.904	*.936*	.894
SAVOY	.695	*.759*	.730
KNSO2	.795	*.827*	.758
TIFM7	.803	.777	.876
LANAG	.644	.448	.466
POWEL	.571	*.615*	.632

Faced with an increasing number of data sets available to drive distributed models (e.g., Di Luzio et al., 2008; Nelson et al., 2010; Mesinger et al., 2006; Hamlet and Lettenmaier, 2005), we suggest that more rather than less care is needed to evaluate the quality of data used for both modeling studies and operational forecasting. In particular, we must counter the temptation to accept new sources of data as inherently 'good' or 'better' simply because they have a higher spatial and temporal resolution. This caution seems reasonable in light of growing list of available approaches to identify and correct biased precipitation estimates (e.g., Looper et al., this issue; Zhang et al., 2011; Guentchev et al., 2010). Care must be taken to correct only man-induced errors in the data and not inconsistencies due to real climate change that are now beginning to appear in multi-decade hydroclimatic records (Milly et al., 2008).

2.5.2. Improvement provided by calibration

An important science question in DMIP 2 concerns potential improvements provided by calibration of model parameter estimates: "What combination of parameterization schemes and calibration strategies seem to be most effective and what is the level of effort required?" Addressing this question in DMIP 2 was constrained by the limits of user/model familiarity and expertise.

For the purpose of DMIP 2, parameter estimation is defined as the derivation of *a priori* estimates of model parameters from physical basin characteristics such as soils data. Calibration is the subsequent process of refining the *a priori* (or other initial) parameter values so that an acceptable level of error is achieved between simulated and observed hydrologic variables (Smith et al., this issue). Central to this discussion is the value of having initial or *a priori* values of the model parameters, the development of which is an active area of research; e.g., see the recent Model Parameter Estimation Experiment (MOPEX) project (Andreassian et al., 2006) and especially the MOPEX special issue (J. Hydrology, vol. 320, 2006) as well as Zhang et al. (2011, in review), Moriasi and Starks (2010), Mizukami and Koren (2008), Peschel et al. (2006), Wang and Melesse (2006), and Koren et al. (2000, 2003a) for additional work on the estimation of *a priori* parameters. Appendix B describes the range of parameterization and calibration strategies used by DMIP 2 participants and should be referred to in the following discussion. We also refer to participants' papers in this Special Issue for additional details on parameter estimation and calibration.

For this experiment, calibration was allowed only at the outlets of the parent basins. During the runs to generate uncalibrated and calibrated simulations at the parent basin outlets, participants were instructed to also generate simulations at interior points.

Figs. 13 and 14 illustrate changes (due to calibration) in r_{mod} and %Bias statistics for ELDO2 and TALO2, respectively (and their interior points). We present these results as typical of the study basins. Appendix D presents the remaining results. The plotted points are connected by arrows, indicating the direction from uncalibrated to calibrated value. For points that plot outside the %Bias plotting scale, we show their coordinates. Values of r_{mod} closer to 1.0 and %Bias closer to zero are desired. Results are also shown for the interior points to evaluate the effectiveness of the parameterization/calibration strategies. The plotting schemes of Viney et al. (2006) are used here to analyze calibration impacts.

For ELDO2, Fig. 13a shows that calibration resulted in improved %Bias measure for all models except NEB (which showed a very slight degradation; uncalibrated −2.3% versus calibrated 3.5%). Calibration also improved r_{mod} in all but one case (CEM). At the interior point Dutch Mills (Dutch, Fig. 13b) both uncalibrated and calibrated performance is worse than at the outlet point (as one would expect from not performing explicit calibration). Note however that several models (LMP, AZ2, VUB, and OHD) gave reasonable performance (at interior point DUTCH), using their *a priori* parameter estimates, and that calibration at the outlet point

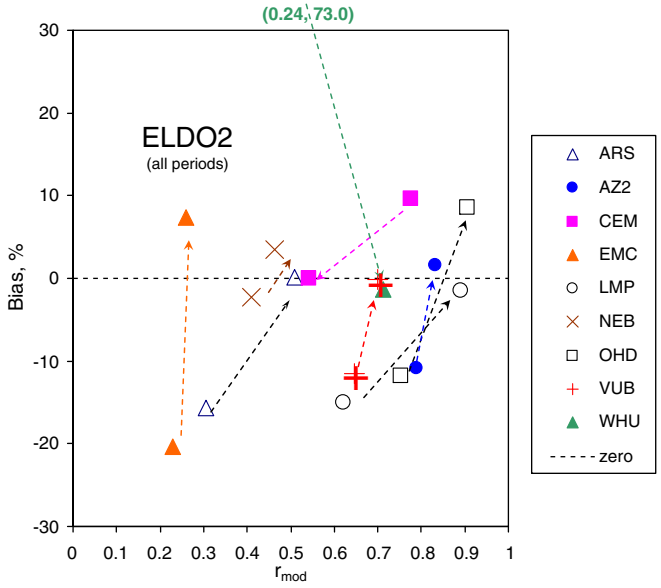


Fig. 13a. Improvement in %Bias and r_{mod} via calibration for ELDO2.

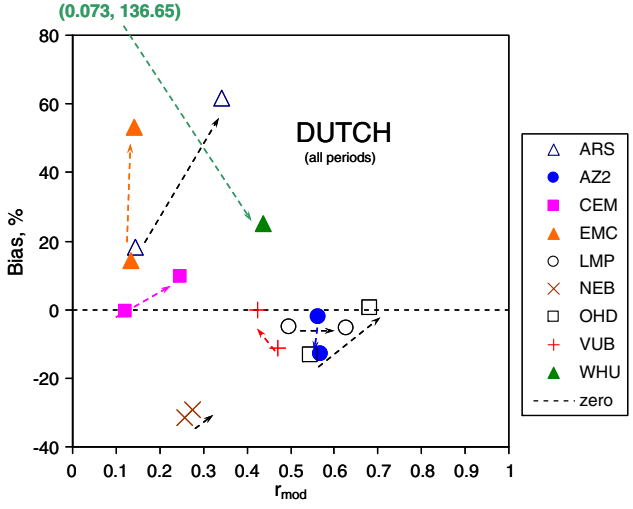


Fig. 13b. Improvement in %Bias and r_{mod} via calibration for ELDO2 interior point DUTCH.

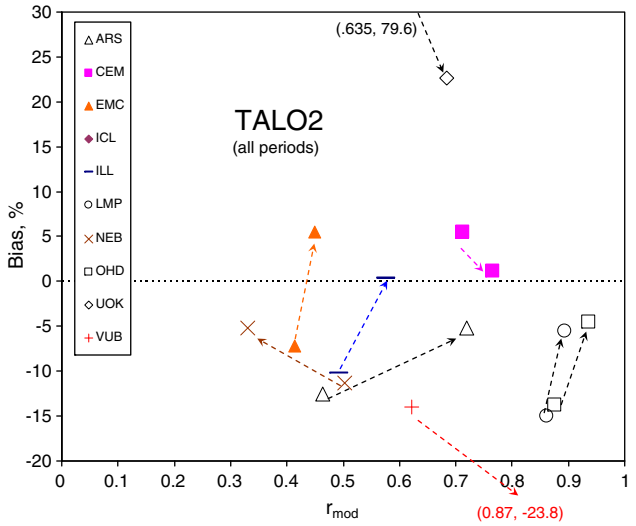


Fig. 14a. Improvement in %Bias and r_{mod} via calibration for TALO2.

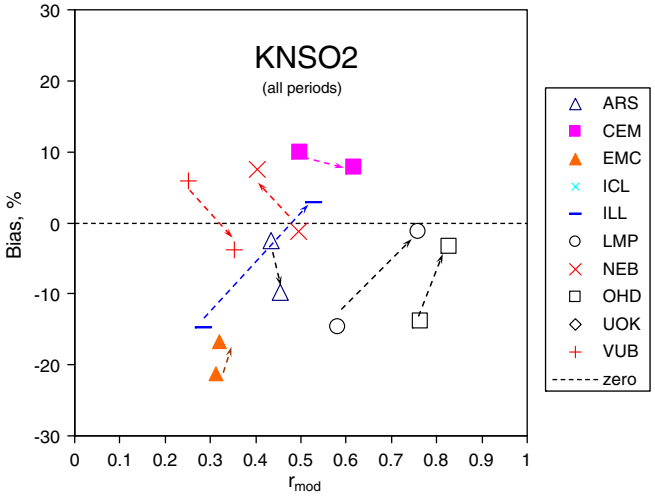


Fig. 14b. Improvement in %Bias and r_{mod} via calibration for TALO2 interior point KNSO2.

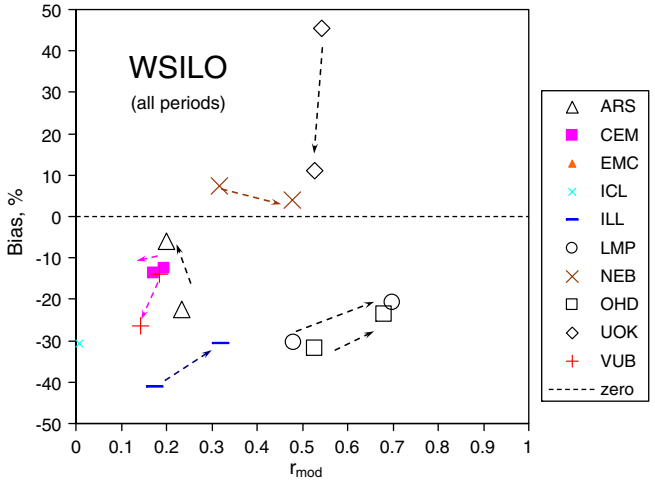


Fig. 14c. Improvement in %Bias and r_{mod} via calibration for TALO2 interior point WSILO.

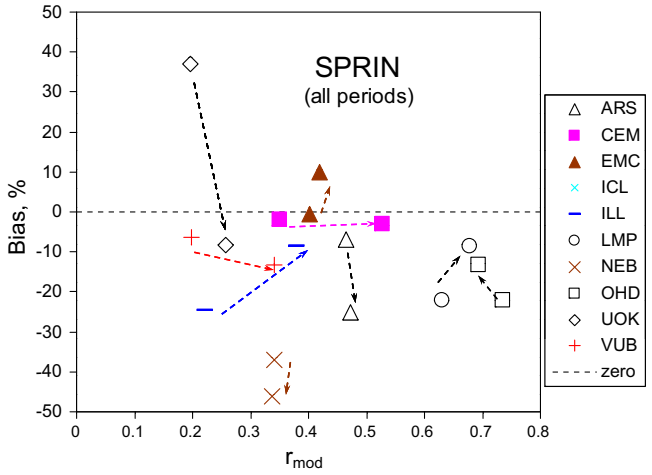


Fig. 14d. Improvement in %Bias and r_{mod} via calibration for TALO2 interior point SPRING.

(ELDO2) did not seem to significantly improve their performance (in terms of these statistics) at the interior point.

Fig. 14a presents the results for TALO2. As in ELDO2, somewhat mixed results can be seen. Calibration improves both the r_{mod} and %Bias for several models: ARS, OHD, LMP, UOK, CEM, EMC, and ILL. However, calibration of the VUB and NEB models improved either r_{mod} or %Bias, but not both.

TALO2 contains seven interior points; here we discuss the three that are unique to TALO2: KNSO2, SPRIN, and WSILO. Figs. 14b–d shows that calibration at TALO2 provided improvements in the r_{mod} and %Bias statistics for six models at KNSO2: CEM, EMC, UOK, ILL, LMP, and OHD. Model performance at the KNSO2 interior point is not as good as at the TALO2 calibration point. At WSILO, three models were improved by calibration: LMP, OHD, and ILL. Calibration at TALO2 improved the r_{mod} and %Bias for three models (LMP, UOK, and ILL) at the SPRIN interior location (see Fig. 14d). Model performance across all the interior points for TALO2 is fairly consistent.

Fig. 15 shows an overview of the impacts of parameterization and calibration. Each subplot in the left column shows r_{mod} for the parent basins arranged in order of decreasing r_{mod} performance. The right side of each parent basin shows the r_{mod} value for a corresponding selected interior point (i.e., the group plotting order is the same). The purpose of this plot is not to rank models but rather to show that while most models benefited from parameter calibration, calibration alone did not result in large improvements in model performance over the level achieved using *a priori* parameters. It is interesting that calibration of the model parameters was unable to compensate for differences in model structures.

DMIP 2 participants used a variety of calibration strategies. Some leveraged the information in lumped model parameter sets to constrain or inform the calibration of *a priori* distributed parameters (e.g., OHD, ICL, UCI, WHU, CEM). The use of such lumped information appears valid given the good performance of the lumped models (LMP and CEM) in the DMIP 2 basins. Alternatively, one strategy (ARS) independently optimized the parameters in each computational element without regard to spatial pattern, letting the parameter values float between defined limits. As in DMIP 1, calibration resulted in performance gains for most models. The calibration strategies for two models (OHD, LMP) improved both the r_{mod} and %Bias statistics in all cases of the five parent basins. The strategies used by ARS and EMC improved both statistics in three of the parent basins, by NEB and VUB in two basins and by CEM in one. Calibration of the models based on the Sacramento

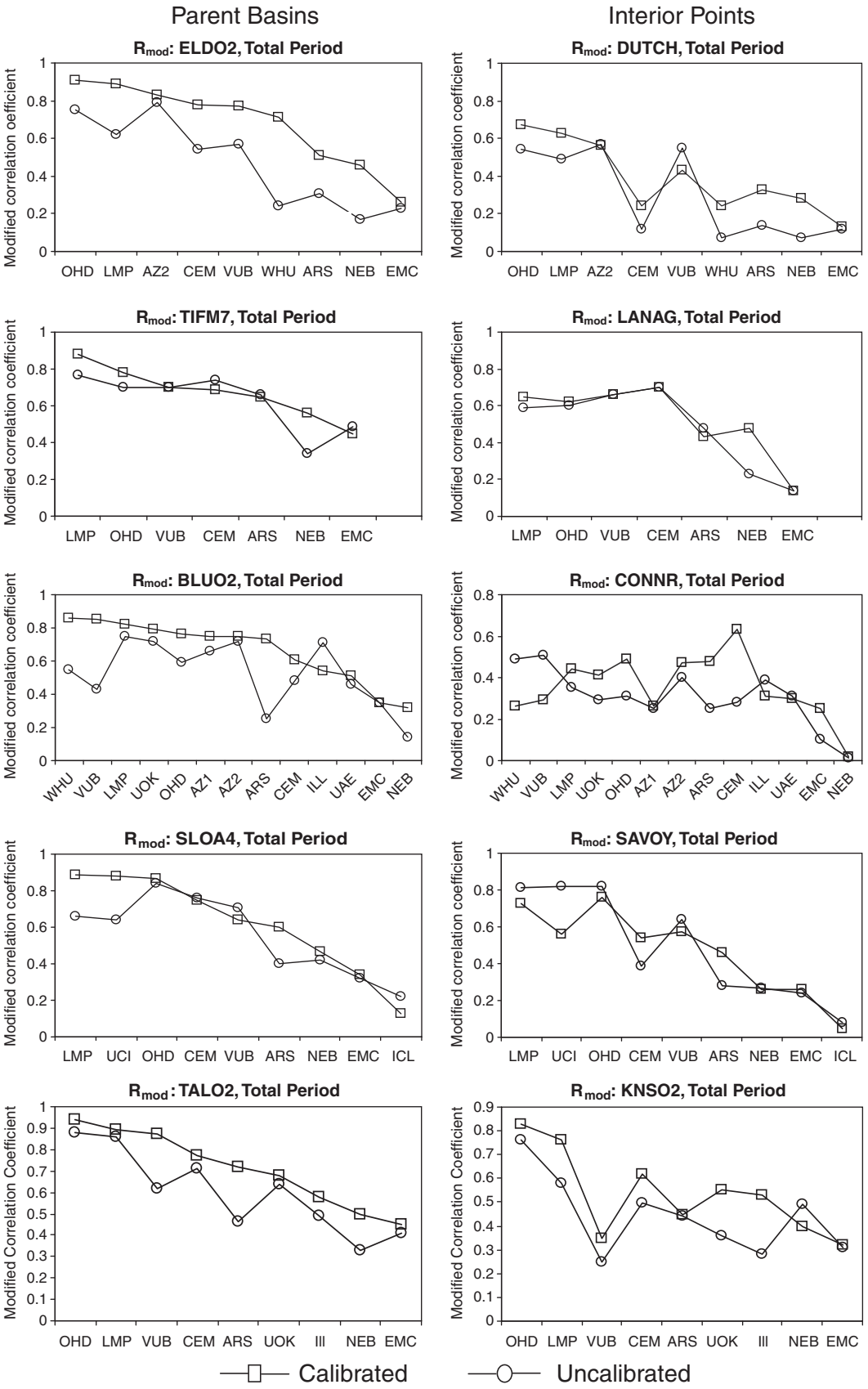


Fig. 15. Values of r_{mod} for parent basins (left) and selected interior locations (right).

model structure (AZ1, AZ2), and UCI) consistently provided improvements at BLUO2, ELDO2, and SLOA4. The strategy used by WHU improved r_{mod} and %Bias at the two parent basins for which it was applied (BLUO2 and ELDO2) as did the strategy used

by UOK, although the final %Bias value for TALO2 remained quite high. The strategy used by ILL produced mixed results for TALO2 and BLUO2. Our results suggest that a strategy using a well-defined *a priori* parameter set and informed by the results of lumped calibration provides better accuracy and more consistent results.

In the majority of cases, calibration at the basin outlet improved simulation statistics at the outlet itself, but these improvements did not consistently translate to improvements at the interior points. In some cases, the 'good' results achieved at interior points using *a priori* parameters were made *worse* when the models were calibrated to basin outlet hydrographs. This suggests that currently used methods are not consistently/unambiguously able to extract accurate information about parameter field distribution (and hence interior point behaviors) from basin outlet hydrographs and the available rainfall data (Pokhrel and Gupta, in press; van Werkhoven et al., 2008). Although not a part of DMIP 2, calibration is sometimes performed at interior points in order to leverage all available data (e.g., Ivanov et al., 2004; Vivoni et al., 2006; Khakbaz et al., this issue). While of considerable interest, our results do not provide insights regarding the use of this kind of data.

Model simulations with *a priori* parameters showed a range of performance. In several parent and interior basins, a few uncalibrated distributed models performed better than some calibrated distributed models, at least in terms of overall and event-based values of r_{mod} and %Bias (Figs. 3–6). This highlights the combined strength of these models their *a priori* parameters for these basins. The OHD model was most consistent in this regard, followed by LMP.

Consistent with Bastidas et al. (2003), our findings reinforce the notion that improvements in calibration techniques, while useful, may currently be less effective than ongoing community efforts to develop advanced *a priori* parameter estimation techniques (e.g., Williamson and Odom, 2007; Zhang et al., 2011, in review), in conjunction with new modeling approaches. These efforts are complemented by the recently proposed focus on developing strategies for using data to help in diagnosing and correcting model structural inadequacies (Gupta et al., 2008; Clark et al., 2011).

2.6. Soil moisture experiment results

A major addition in DMIP 2 was the experiment to evaluate soil moisture simulations by distributed models for the Oklahoma basins. The science questions here are: "*Can distributed models predict processes such as runoff generation and soil moisture re-distribution at interior locations? At what scale can we evaluate soil moisture models given current models and sensor networks?*"

For this evaluation, participants provided 4 km gridded daily runoff and soil moisture estimates over an area covering the entire Oklahoma Mesonet instrumentation domain in Fig. 3 of Smith et al. (this issue). Simulations were generated using *a priori* parameters without calibration, and with no hillslope/channel routing being performed. This domain exhibits a strong gradient in the climate index, (herein defined as the ratio of annual precipitation to potential evaporation, P/PE), ranging from 0.57 in the western portion to 1.18 in the eastern portion. Lower values of P/PE indicate a dry climate, while larger values imply a wetter climate. The strong P/PE gradient, combined with the variety of soil types and landcover, facilitates the evaluation of soil moisture simulations throughout a climatically diverse region and over a wide range of conditions (Gu et al., 2008; Wood et al., 1998). Others have used the P/PE index (e.g., Koren et al., 2006, 2008; Duan et al., 2006; Schaake et al., 2004; Dooge, 1992) or its reciprocal, the dryness index (e.g., Milly, 1994; Wollock and McCabe, 1999; Zhang et al., 2001; Sankarasubramanian and Vogel, 2002; Wagener et al., 2007) in hydrometeorological studies.

Illston et al. (2008) and Koren et al. (2006) noted that several issues accompany the use of the volumetric soil moisture data from the Mesonet sites. First, the upper bound of the soil moisture estimates is limited by the accuracy of the Campbell Scientific 229L sensor (Campbell Scientific 229L User Manual, 2010). Temperature observations from this sensor are converted to soil water matric potential using empirical relationships (Schneider et al., 2003). The units of soil matric potential are kilopascals, (kPa), negative by convention. Given that the lower limit of the observed values of the temperature reference is approximately 1.4 °C, the equation for computing soil water matric potential does not return values near saturation between 0 and −10 kPa. Thus, the upper limit of the soil moisture observations corresponds to a matric potential of −10 kPa. Second, the instantaneous volumetric soil moisture measurement at a station is related to the soil type and the physiographic properties of the location as well as to the availability of moisture supply (i.e., precipitation) in the area. This complicates comparisons of stations located in different areas even during similar weather conditions. Third, hydrologic model states and volumetric soil moisture measurements may not have a one-to-one correspondence, and hence, a completely objective comparison of these two quantities may not be possible. Moreover, the Oklahoma Mesonet soil moisture measurements were designed for drought monitoring over a large area (average coverage is one site per 3000 km^2) and as a result, these observations do not represent soil moisture variability at the hillslope-type scale, but may be used as indicators of soil moisture variability over mid- to large-size watersheds.

To reduce the impacts of these issues, participants were asked to compute estimates of a soil moisture saturation ratio defined as (Koren et al., 2006, 2008):

$$SR = \frac{\theta - \theta_r}{\theta_s - \theta_r} \tag{3}$$

where θ is the computed or observed volumetric water content, θ_r is the residual volumetric water content (or wilting point), and θ_s is the saturation volumetric water content (or porosity). The instructions called for SR to be computed for three layers: 0–5 cm, 0–25 cm, and 25–75 cm depth. The SR index attempts to reduce the effects of the individual soil property variation on intercomparison. Others have used similar relative measures. For example, Sridhar et al. (2008) used a form of Eq. (3) in order to define a soil moisture index (SMI) that had the same numeric range as the US Drought Monitor. Schneider et al. (2003), Illston et al. (2004, 2008), and Gu et al. (2008) used a fractional water index (FWI) to avoid the issues mentioned above.

Only EMC and OHD submitted a full set of soil moisture simulations. These consisted of time series of 4 km gridded, daily average runoff and SR values for a large modeling domain encompassing the state of Oklahoma (Smith et al., this issue). Models were run with *a priori* (uncalibrated) parameters. After the DMIP 2 conference, VUB submitted a simulation of SR for the "West" Oklahoma Mesonet site (see Appendix E). The results shown in Appendix E indicate that the VUB model is able to fairly well reproduce water content variations in wet and dry periods at a point. The plot of the observed soil moisture in Appendix E also shows how the upper limit of SR is affected by the sensor (as discussed above).

For the analysis of the soil moisture simulations at interior locations, we computed basin averages of the gridded runoff and SR values for 75 basins within the modeling domain used previously by Koren et al. (2006) and ranging in size from 20 km^2 to 15,000 km^2; please refer to that paper for more information. Observed streamflow data for each basin is available from the USGS. Our evaluation compared simulated to observed runoff at basin outlets, and compared basin averages of simulated SR to values of SR derived from the Oklahoma Mesonet observations.

To compute "observed" SR values, the Oklahoma Mesonet soil moisture measurements, which are recorded automatically every 30 min (Illston et al., 2003), were aggregated to obtain daily average

values. For each layer, point SR values were interpolated to a 4 km grid for the entire State of Oklahoma using inverse distance weighting, with weights computed on a daily basis from stations having available data on a given day. We assumed that this weighting scheme is appropriate given that spatial correlation of the point soil moisture observations degrades quickly with distance. The gridded daily maps of SR were then used to generate daily time series of basin average soil moisture for the period (Koren et al., 2006).

Table 7 presents several statistics of 10-day averaged daily runoff and daily SR values at three depths (0–5 cm, 0–25 cm, and 25–75 cm) for the 6-year simulation period from January 1, 1997 to December 31, 2002. In computing the runoff statistics, a 10-day interval was used to reduce the impact of omitting hillslope/channel routing from the experiment. This period is notable in that it includes severe droughts that occurred in 1998 and 2000 (Illston et al., 2003, 2004; Illston and Basara, 2002; Hong and Kalnay, 2002). The simulations from both models correlate well with observed data, in spite of the fact that severe droughts occurred over large parts of the region in both 1998 and 2000 while above average soil moisture conditions were experienced in 1997 and other years. For the OHD model, the correlation coefficients for soil moisture in the two upper layers are high and degrade with depth. For the EMC model, the soil moisture correlations display the opposite behavior, and increase with depth. For both runoff volumes and soil moisture, the correlation coefficient and Nash–Sutcliffe (NS) efficiency for the OHD model are consistently higher than for the EMC model. Some degradation of soil moisture simulation accuracy for both models can be observed for the deeper soil layer with the NS efficiency becoming negative.

For both models and for all basins, the average soil moisture correlation coefficients in Table 7 are higher than those between the NDVI/NDWI and soil moisture averaged for 17 sites in the Oklahoma Mesonet domain in the study by Gu et al. (2008). This suggests that the EMC and OHD models can simulate soil moisture reasonably well.

Figs. 16 and 17 show the relationship of SR versus P/PE for the upper (0–25 cm) and lower (25–75 cm) soil depth layers respectively, while Fig. 18 shows runoff versus the P/PE. Here, P for each of the 75 basins was derived from the DMIP 2 radar-based precipitation grids, while PE was taken from the NOAA Evaporation Atlas (Farnsworth et al., 1982). Values from this atlas were supplied by DMIP 1 and 2 and are still used by NWS RFCs. The plots in these figures clearly show that the models do a good job of reproducing the patterns suggested by the measurements regarding the dependency of soil moisture and runoff on P/PE. Consequently, efforts have been initiated to assimilate soil moisture observations of the type provided by the Oklahoma Mesonet into distributed models (e.g., Lee et al., in press).

Table 7
Overall soil moisture and runoff statistics from 75 basins.

Model	Statistics				
	RMSE	Bias	Abs. error	Correlation coefficient *r*	NS
10-day averaged daily runoff (mm/day)					
EMC	0.745	0.142	0.435	0.715	0.350
OHD	0.580	0.091	0.322	0.811	0.606
Soil saturation index (0–05 cm layer)					
EMC	0.123	0.044	0.100	0.733	0.039
OHD	0.109	−0.032	0.089	0.803	0.241
Soil saturation index (0–25 cm layer)					
EMC	0.117	−0.005	0.098	0.738	0.148
OHD	0.111	−0.031	0.092	0.794	0.238
Soil saturation index (25–75 cm layer)					
EMC	0.138	−0.092	0.121	0.827	−0.421
OHD	0.128	−0.077	0.110	0.746	−0.221

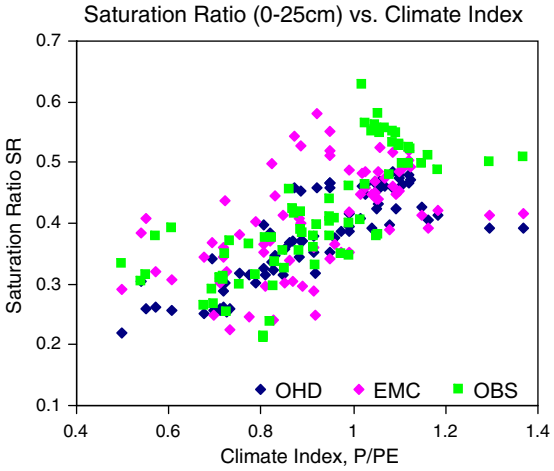


Fig. 16. Soil moisture saturation ratio versus climate index for top 0–25 cm soil layer.

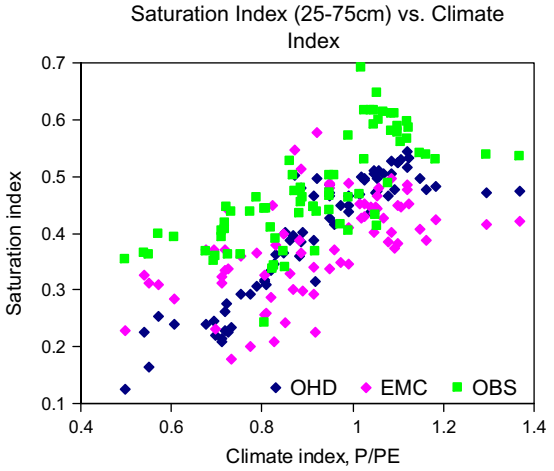


Fig. 17. Soil moisture saturation ratio versus climate index for 25–75 cm soil layer.

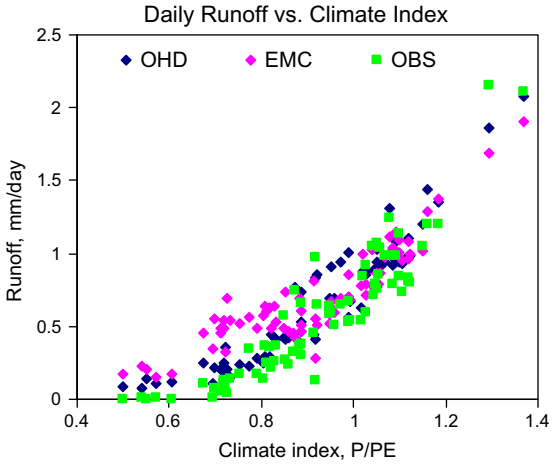


Fig. 18. Daily runoff versus climate index.

To address the question of soil moisture predictability versus scale, we analyzed the relationship of SR RMSE versus basin scale as shown in Fig. 19 for two soil layers: 0–25 cm and 25–75 cm. The RMSE measure is computed over the January 1, 1997 to December 31, 2002 for each of the 75 basins. The results for the two largest basins (11,700 and 15,200 km^2) are omitted as basin average soil moisture estimates over such large areas have little

meaning. For both EMC and OHD models, the RMSE is greater in the lower soil layer compared to the upper layer. There is a very slight reduction in RMSE as drainage area increases (left column), although the number of data points at large basin scales is small. To counter this imbalance, the SR RMSE values are averaged within drainage area intervals, with the intervals defined so as to contain nearly equal numbers of data points (right column). There is a slightly greater reduction in SR RMSE with drainage area in the upper soil layer compared to the lower soil layer (right side).

Taken together, the EMC and OHD soil moisture and streamflow results highlight the similarities and differences between models that have traditionally been classified as either hydrologic simulation models or land surface models (LSMs). Our results show that for these basins and time scales, a hydrologic model modified to compute physically-based soil moisture and temperature (OHD) was able to produce better simulations of runoff and soil moisture than was a traditional LSM (EMC).

2.7. Routing experiment results

This final section examines the questions: In what ways do routing schemes contribute to the simulation success of distributed models? Can the differences in the rainfall–runoff transformation process be better understood by running computed runoff volumes from a variety of distributed models through a common routing scheme? Our intent was to address the DMIP 1 recommendation to separate the analysis of routing and rainfall runoff techniques (Reed et al., 2004; Lohmann et al., 2004). We chose the OHD kinematic hillslope and channel routing scheme to represent a widely-used approach in distributed modeling, with the hope that our results would have broad applicability. For this experiment, ARS, EMC, and ARZ-2 provided calibrated hourly time series of gridded runoff volumes for the BLUO2 (October 10, 2000–March 20, 2001), while ARS also provided simulations for TALO2 (April 25, 2000–July 31, 2000). These periods contained 14 and 4 flood events of various magnitudes for BLUO2 and TALO2, respectively. These runoff volume time series were routed through a common (OHD) calibrated kinematic hillslope and channel routing scheme; hereafter, we refer to these as the OHD_routed simulations. The analysis assumed that the parameters of the participants' calibrated models were within reasonable ranges.

For brevity, results are shown only for BLUO2 (Fig. 20a–d). The graphs show the average values of the %Bias, RMS, %peak flow, and %peak time errors for 14 BLUO2 events. Each sub-plot shows the statistics for the participants' original simulation along side the statistic from the corresponding OHD-routed simulation.

Looking collectively at the plots, it can be seen that there is not much difference between the statistics for the original and OHD_routed simulations. In some cases the statistics from the OHD_routed simulations are slightly improved and in other cases they are slightly degraded. The results suggest that for this basin and these events, errors in modeling the rainfall–runoff process will not necessarily be reduced by the routing component in a distributed model.

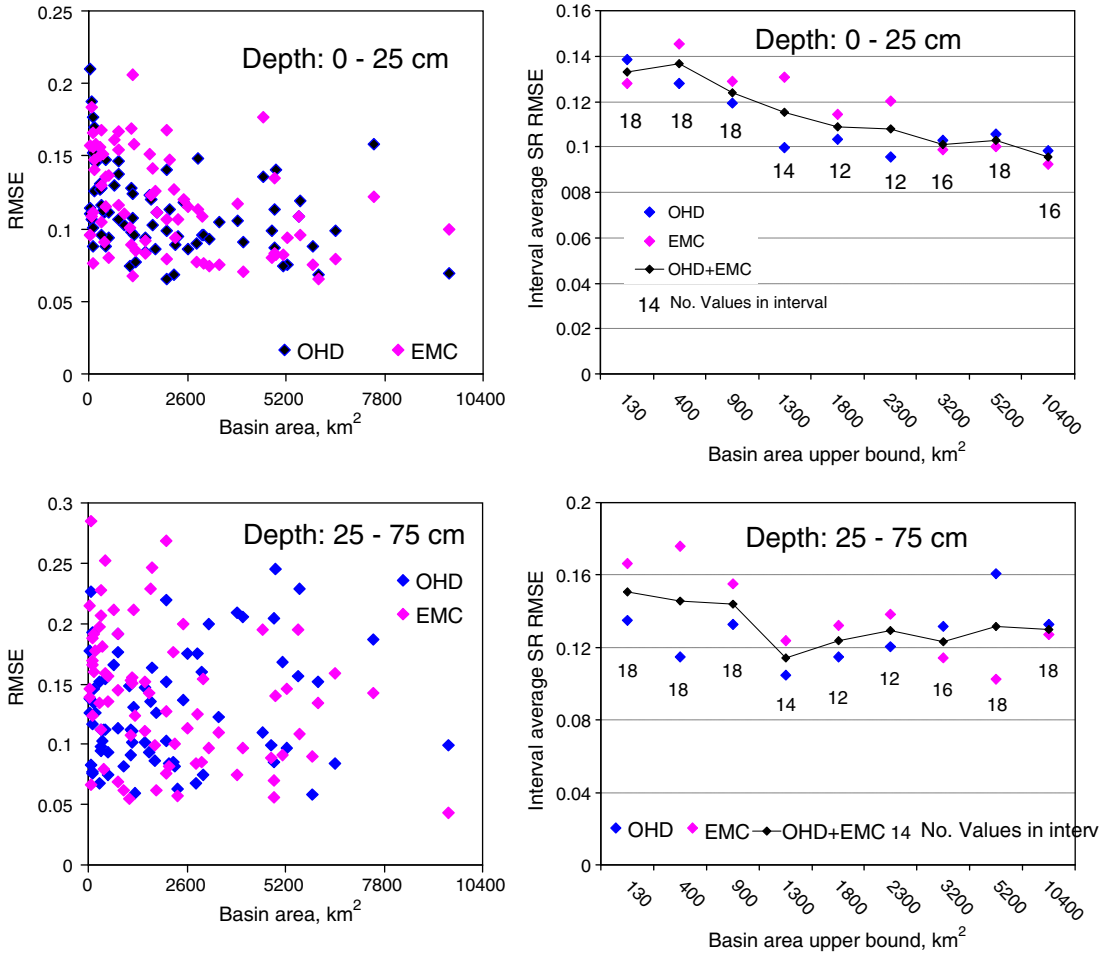


Fig. 19. Relationship of RMSE of soil moisture saturation ratio (SR) versus drainage area. The plots on the left are the RMSE values for 75 basins computed over the period from January 1, 1997 to December 31, 2002. Two soil layers are shown. On the right side the RMSE values are averaged within drainage area intervals. The line in the graphs on the right side is the average of the OHD and EMC errors within each interval.

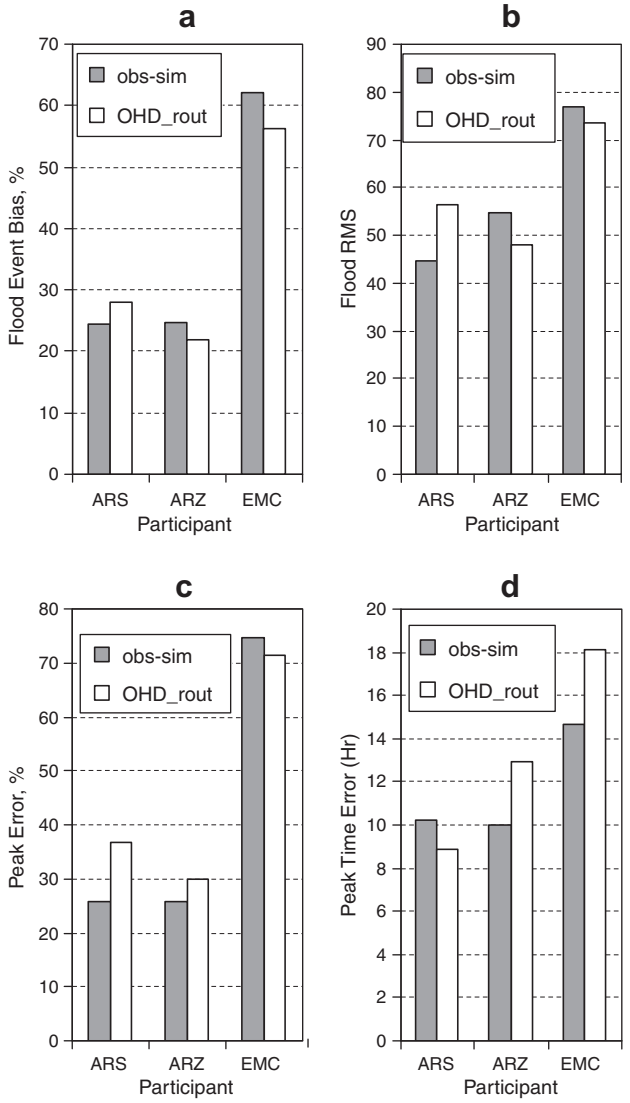


Fig. 20. Results of the routing experiment for BLUO2. The plots show the average values of the (a) bias, (b) RMS, (c) peak flow, and (d), peak time errors for 14 events between October10, 2000 to March 20, 2001. Each plot shows the statistics for the participant original simulation (obs–sim) along side the statistic from the corresponding OHD-routed simulation.

3. Conclusions

The key findings of the Oklahoma DMIP 2 experiments are as follows:

(1) Distributed models, calibrated using basin outlet hydrographs, do a reasonably good job of partitioning precipitation into runoff, evaporation, and losses at basin outlets. In the cases shown here, this finding is supported by the ability of the distributed models to generate good streamflow simulations at interior points. These two findings provide confidence that distributed models can account for spatial variability in basin features and precipitation while successfully preserving the water balance.
(2) The data used in calibrating models must be stationary and unbiased.
(3) Two distributed models were able to provide reasonably good soil moisture simulations; however the streamflow simulation performance of one model was markedly better than the other.
(4) Many of the distributed models participating in this study were able to provide improved hydrograph simulations compared to lumped models, when both types were calibrated. We used two calibrated lumped models to form a combined benchmark, thereby establishing a more rigorous basis for drawing conclusions.
(5) The calibration strategies tested in DMIP 2 provided gains in performance. However, the improvements from calibration did not greatly impact relative model performance established by using *a priori* parameters.
(6) In several parent and interior basins, some uncalibrated distributed and lumped models performed better than other calibrated distributed models. This result highlights the strength of several model/parameter combinations.
(7) Those lumped and distributed models that performed well at basin outlet points also, in general, performed well at interior points having a wide range of drainage areas.
(8) In a limited experiment, errors in modeling the rainfall/runoff process were not considerably impacted by the type of routing component used by a distributed model.

Our experiences during this study, and the results reported here, lead us to conclude that distributed models are indeed making gains toward realizing their full, hoped-for potential. While distributed models provided improved outlet simulations on a limited basis compared to lumped models, this result should not be considered as a cause for major concern. It is quite possible that there are scientifically sound reasons why distributed models may not be able to outperform lumped models for basin outlet simulations in all cases (e.g., Pokhrel and Gupta, in press; Smith et al., 2004b). More important than being able to outperform a calibrated lumped model, we should probably be placing greater emphasis on the finding that distributed models are able to produce reasonable answers at interior points and to do so for the right reasons (e.g., interior soil moisture and runoff). The results also imply that at least for the foreseeable future, operational agencies should consider maintaining currently-used lumped models as they transition to distributed models for river and water resources forecasting. Forecasters would then able to look at lumped and distributed results together and use situational awareness to choose the better model result for a given event.

Clarke (2008) presents an interesting critique of model comparison experiments, and offers many sound suggestions on the types of statistical tests that should be performed. He suggests that in this era of many watershed models, studies need to help practitioners identify models which are "in some senses better than others." Several times he mentions the DMIP experiments in his discussion. While we value his interest in DMIP, we would like to urge caution in as much as the DMIP studies were designed to provide a venue for evaluating and improving a class of models rather than to derive a list of recommended models. Unlike the MOPEX experiment, which tried to avoid such a ranking by removing the specific names of models from the discussion of results (Duan et al., 2006), we have identified the models so that an ongoing examination of modeling strategies can help to inform model improvements.

Nonetheless, we do agree with Clarke's (2008) suggestion that what (unavoidably) is being tested is the model/developer combination (and not the model itself). It would be enormously complicated if the model testers were (in general) also required to learn how to use alternative models (and it would not remove the fact that the results would then be unavoidably associated with the strategies employed by the people doing the testing). Additionally, it would generally be very costly to do so. As a result, DMIP 1 and 2 were designed as cost-effective venues to pursue scientific and research-to-operations questions. The

experiments were open to any and all participants, whether they developed the model they used or not: some participants in DMIP 2 were not the model developers (e.g., NEB). Of course, from a 'usability' standpoint there is merit to having models being tested by those who don't develop them. However, one must protect against the fact that, the testers may, due to inexperience or lack of knowledge, be unable to garner the most representative results possible from a model and thereby provide a sound, and perhaps more importantly *fair*, basis for performance evaluation. It may be helpful to mention here that several of the DMIP 1 participants used their participation not as an opportunity to 'recommend' their models but instead (as intended by the DMIP organizers) as an opportunity to test and subsequently improve their models (ARS – Gassman et al., 2007; MIT – Ivanov et al., 2008; Mascaro et al., 2010; OHD – Koren, 2006). Moreover, DMIP 2 participants in the western basin experiments are viewing those tests as an opportunity to evaluate new model components (V. Andreassian, personal communication).

Viewed collectively, the results of DMIP 1 and 2 provide a robust intercomparison of over 20 models. We believe that the results herein represent a rigorous assessment of distributed modeling with operational (versus research) quality data.

4. Recommendations

DMIP 2 was a successful experiment, leading to the enhanced understanding of general model performance and the impacts of forcing data errors and calibration. Detailed experiments within a diagnostic framework such as that proposed by Clark et al. (2008) are necessary to diagnose specific strengths and to uncover weaknesses in modeling approaches. Such experiments should be conducted over many basins covering a broad range of climatic indices (Andreassian et al., 2009). Along these lines, we suggest that more comprehensive analyses be performed that use a common hillslope/channel routing scheme to isolate differences in rainfall/runoff models.

Distributed models should be viewed as complements rather than replacements of lumped models in operational forecasting environments, at least for the foreseeable future. Lumped models provide a valuable integrated view of the basin outlet response.

DMIP 2 results highlight the importance of *a priori* parameters for distributed model simulations. Given this importance, more efficient approaches need to be developed for *a priori* parameter estimation that account for regional properties as well as general physical laws. Efforts should also include more robust techniques of parameter estimation with the use of new data sources.

We recommend that additional experiments be conducted to complement our limited routing tests. These experiments should consider the recalibration of participants' rainfall/runoff models joined with the common routing scheme.

Distributed model calibration is still largely based on approaches in which the variable *a priori* parameter field is multiplied by a basin-wide scalar. Such approaches limit potential improvement gained by calibration of internal basins when climate/physical properties and their uncertainties vary significantly (e.g., comparing CONNR to BLUO2 or applying scalars to sub-basins in mountainous watersheds). In the mean time, there is a need for more general calibration approaches that apply spatially-variable scalars to account for the spatial uncertainty in model parameters and input data. We support the call for continued cooperation between the parameter estimation and model development communities (Rosero et al., 2009).

Yet another major need is the testing of models in a 'pseudo-forecast environment' with forecast-quality forcing data and state updating. Such tests are a logical complement to the process simulation experiments in DMIP 1. While much work has been done to evaluate the improvements realized by distributed models in simulation mode, the NWS also needs to investigate the potential gains when used for hydrologic forecasting.

Lastly, we repeat the recommendation of Reed et al. (2004) that experiments are needed to understand the impacts of hydrologic, model structure, and parametric uncertainty on distributed model performance.

Acknowledgements

We are very grateful to the many participants and their sponsoring institutions and agencies for their collaboration on this project. In addition, Billy Olsen, Eric Jones, and Bill Lawrence of the NWS ABRFC provided a great deal of help with data and reviews of manuscripts. Brad Illston of the Oklahoma Climatological Survey provided help in interpreting the Oklahoma Mesonet soil moisture observations. Todd Hallihan of Oklahoma State University provided his insights into the hydrogeology of the Blue River basin. Lastly, we are grateful to the reviewers of this manuscript for their beneficial comments.

Appendix A

Non-OHD DMIP 2 coauthors and affiliations (see Table A-1).

Appendix B

Parameterization and calibration strategies (see Table B-1).

Table A-1
Non-OHD DMIP 2 coauthors and their affiliations.

CEMAGREF, France (CEM)	Vazken Andreassian, Julien Lerat, Cecile Loumagne, Charles Perrin, Pierre Ribstein
U. Arizona, Tucson, Arizona (AZ1, AZ2)	Hoshin V. Gupta, Koray K. Yilmaz, Prafulla Pokhrel, Thorsten Wagener
DHI Water and Environment, Horsholm, Denmark (DH1, DH2)	Michael Butts, Keiko Yamagata
U. California at Irvine, Irvine, California (UCI)	Soroosh Sorooshian, Behnaz Khakbaz, Alireza Behrangi, Kuolin Hsu, Bisher Imam
Vrije U. of Brussels, Belgium (VUB)	Florimond De Smedt, Ali Safari, Mohsen Tavakoli
Wuhan University, Wuhan China (WHU)	Lan Li, Xin Wang, Jian Wu, Chao Yang Mengfei Yang, Zhongbo Yu
U. Alberta Edmonton, Canada (UAE)	Thian Gan, Zahidul Islam
U. Oklahoma, Norman, Oklahoma (UOK)	Baxter Vieux, Jonathan Looper
I. M. System Group and NOAA/NCEP/EMC (EMC)	Youlong Xia, Kenneth Mitchell, Michael Ek
Imperial College of London (ICL)	Neil McIntyre, Barbara Orellana
U. Illinois, Urbana-Champaign, Illinois (ILL)	Murugesu Sivapalan, Hongyi Li, Fuqiang Tian
U. Nebraska at Lincoln, Nebraska (NEB)	Jae Ryu (Now at University of Idaho)
USDA Agricultural Research Service, Temple, Texas (ARS)	Jeff Arnold
USDA Agricultural Research Service, Corvalis, Oregon (ARS)	Gerald Whittaker, Remegio Confesor
Blackland Research Center, Temple, Texas (ARS)	Mauro Di Luzio

Table B-1
Parameterization and calibration strategies for the DMIP 2 participants.

Participant	Parameterization	Calibration
OHD	Gridded *a priori* SAC-SMA parameters derived from soil texture (Anderson et al., 2006; Koren et al., 2000). Gridded routing parameters derived from observed USGS data and geomorphologic relationships (Koren et al., 2003b; 2004)	Start with *a priori* parameters. Revise *a priori* parameters using lumped calibrated parameters (derived using procedures in Smith et al., 2003): scale gridded *a priori* values by ratio of the SAC-SMA parameter value from the lumped calibration to the average parameter value from the *a priori* grid. Then use scalar multipliers to uniformly adjust each parameter while maintaining spatial variability. Scalars are calibrated manually and/or automatically. Automatic calibration uses a multi-time scale objective function (Kuzmin et al., 2008)
UCI	Our semi-distributed model has three main components: (1) SAC-SMA as the water balance component for each sub-basin; (2) Overland flow routing; and (3) River channel routing 13 Major parameters of SAC-SMA were defined via calibration of the model while the parameters of overland flow and channel routing components were obtained without calibration. The overland flow routing parameters were defined through GIS processing of the selected basin. The parameters of channel routing component (e.g. Manning roughness and cross section properties) were obtained from previous study on the Illinois River basin at Watts (Ajami et al., 2004)	Semi-Lumped calibration (SL) strategy (Ajami et al., 2004) used to simulate streamflow at the outlet and interior points. Distributed precipitation forcing was applied at each sub-basin. Identical SAC-SMA model parameters used at all sub-basins and model calibration optimized a single parameter set in the distributed model structure. Only the observed streamflow at the basin outlet during calibration period was available to assess the goodness of fit for the tested calibration strategies, SL calibration scenario, which gave the best performance during calibration period among the other calibration strategies, was selected. The best performance of our model in terms of streamflow simulation at the outlet as well as interior points is obtained when the optimal parameter set is estimated, through calibration of the lumped SAC-SMA model (entire watershed) and then applied identically to all sub-basins in the distributed model configuration. The results of this calibration scenario is discussed in Khakbaz et al. (2011, this issue)
NEB	Hydrologic parameters of HSPF are determined by watershed delineation processes (e.g. hypsometric curve) built in BASINS; Soil and land use parameters are derived from posterior parameterization based on spatial analysis to facilitate partitioning of the watershed into land segments and stream network (EPA BASINS, 2001; EPA HSPF, 1970)	Annual water balance was manually adjusted, if needed, and then automatic calibration software, the Parameter Estimation (PEST), was utilized to calibrate hydrologic simulation. PEST is a model-independent parameter estimator. The search algorithm built in PEST implements a robust variant of the Gauss–Marquardt–Leyenberg method (GML) of parameter estimation by maintaining a continuous relationship between model parameters and model outputs (Doherty and Johnston, 2003; Marquardt, 1963)
EMC	Parameterizations of Noah hydrologic and soil physics are described in Chen et al. (1996) and Ek et al. (2003). Noah uses parameterization of snowpack and frozen ground from Koren et al. (1999) and runoff parameterization from Schaake et al. (1996). Saturated hydraulic conductivity Ksat and hydrologic conductivity Kdt are taken from MOPEX (Duan et al., 2006). Routing model from Lohmann et al. (1998) is used	For the calibration period, manually adjust two soil parameters (Ksat and Kdt) values to minimize root mean square error between observed and simulated streamflow for each basin. Noah default values are Kdt = 3.0/Ksat; Ksat is a table used in the Noah and it depends on soil types. The Ksat and Kdt for MOPEX experiment (Duan et al., 2006) are expressed as: Ksat = 5.5Ksat (default), Kdt = 2.59–0.044Ksat. In DMIP2 experiment, use Ksat = bKsat (default) and Kdt = 2.59–0.044Ksat. Here b is a calibrating parameter. The value of b is calibrated for each basin using a manual method to achieve the minimum of root mean square error between observed and simulated annual mean streamflow
ARS	Segmentation and parameterization of the sub-watersheds using digital terrain data and geomorphologic relationships. Sub-watershed partitioned in Hydrologic Response Units (HRUs). HRU's parameters derived from soil and land use data information (Di Luzio et al., 2004) Precipitation input records defined as area-weighted average over each sub-watershed (Di Luzio and Arnold, 2004)	From *a priori* parameters and SWAT documentation, set reasonable limits on the range of each variable to be calibrated. Initialize the calibration with a random draw from the range of each variable. The calibration procedure is an application of the non-dominated sorted genetic algorithm (NSGAII, Deb et al., 2002) to two objectives. The objectives are defined as the root mean square error (RMSE) of event driven flow predictions and the RMSE of base flow predictions for the watershed. The procedure is implemented to run on a parallel computer (Confesor and Whittaker, 2007). The parameters are free to vary within the bounds without regard to spatial patterns
CEM	Lumped parameterization Model GR4J uses four parameters (Perrin et al., 2003): – Soil moisture accounting reservoir capacity (mm) – Intensity of interwatershed groundwater flows (gain/losses function) (mm)	*Calibrated simulations* On gauged points, calibration is performed in three steps combining global and local optimization (Perrin et al., 2008): 1. Parameter space is regularly sampled by selecting 3 values for each of the four GR4J parameters and producing all possible combinations of the different parameters values. These values are given by the quantiles 30%, 50% and 70% of the distribution of GR4J

(*continued on next page*)

Table B-1 (*continued*)

Partici-pant	Parameterization	Calibration
	– Routing store capacity (mm) – Time base of the unit hydrograph (h)	parameters after a calibration on 1054 catchments in France 2. The parameter set having the highest Nash–Sutcliffe efficiency is selected among the 181 sets as a starting point for step 3 3. Local optimization based on a steepest descent algorithm is then applied and leads to the final parameter set On interior points, model parameters are derived from downstream gaged catchment parameters by the following procedure: The first three parameters are set identical to those of the downstream catchment (SMA store capacity, intensity of gain/losses and routing store). The fourth one (time base of the hydrograph) is calculated by the following regression formula obtained on 1054 catchments in France (Oudin et al., 2008): $TB = 0.25 + 63 \times \left(\frac{Surf}{Ra^2 \cdot Std(Rh>0)}\right)^{0.313}$ where TB is the time base of the unit hydrograph (h), Surf is the drainage area in km^2, Ra is the mean annual rainfall in mm/year, and Std($Rh > 0$) is the standard deviation of strictly positive hourly rainfall in mm/h *Uncalibrated simulations* The following set of parameters is applied (mean values of the five parameters sets obtained after calibration on the five DMIP gaged points): SMA store capacity of 300 mm, Inter-watershed Groundwater Flows intensity of −0.3 mm, Routing store capacity of 60 mm, Time base of the hydrograph of 20 h
UAE	Parameters can be grouped as Vegetation, Soil and Channel (Biftu and Gan, 2001; Biftu and Gan, 2004) Vegetation: LAI derived from monthly greenness fraction data, Initial canopy storage derived from percent of canopy capacity, other parameter values as attenuation coefficient taken from literature (Kalinga and Gan, 2006; Biftu and Gan, 2001) Soil: Soil types derived from the DMIP soil data, Soil hydraulic properties derived from Rawls and Brakensiek (1985), roughness values initially estimated and then calibrated Channel: The mean cross sectional top width and Manning's roughness coefficient are calibrated	Four model calibration parameters viz. as exponential decay parameter of saturated hydraulic conductivity, Manning's roughness coefficient for soil and vegetation, mean cross sectional top width and Manning's roughness coefficient for channel At first the exponential decay parameter (f) was set in order to provide sufficient base flows as well as to properly model the seasonal variation of local ground water table at sub-basin scale. Starting with some initial values following Beven's (1982) suggestions the value of f was set by observing the ground water table of sub-basins as well as the observed and simulated discharge. Then Manning's roughness coefficient for soil and vegetation was calibrated to refine the response function for different sub-basins Finally starting with some estimated value from field observation and previous study, the Manning's roughness parameter for all channels was refined by matching the lag time and magnitude of the simulated and observed peak discharge. However as the Muskingum–Cunge method for channel routing is relatively insensitive to the mean top width of the water surfaces at channel reaches (Biftu and Gan, 2001); in calibration they remain equal to the cross-sectional measurement database of DMIP 2
UOK	The Green and Ampt soil parameter maps are determined from the Rawls and Brakenseik relationships (Rawls et al. 1983a, b). Manning coefficient maps derived by relating Land Use and Land Cover (LULC) maps to the corresponding roughness coefficients (Vieux, 2004) Rating curves from USGS gauging stations are used where available. Trapezoidal channel cross-sections are used where cross-sections are surveyed or rating curves exist. Floodplain storage is modeled in the lower reaches of the Blue and Illinois using cross sections extracted from 10 meter USGS DEM. For trapezoidal channel cells, geomorphic relationships between channel width and drainage area are estimated from National Agriculture Imagery Program (NAIP) orthophotos. The geomorphic relationship is used to assign base width to the channel network based on drainage area Baseflow separation is performed on the observed hydrographs using the PART software by the USGS. The PART program searches for days that meet a daily antecedent streamflow recession of 0.1 log cycles. Days that meet this requirement are then used to estimate the baseflow during the storm flow periods	First, separate events by saturation excess versus infiltration excess. Plot observed and simulated peaks and volumes for events to identify saturation excess (typically under predicted) or infiltration excess (typically over predicted). Identify events after extended moisture deficits (e.g. drought) when initial runoff is driven by infiltration excess. The scatter plots provide an insight due to the appearance of two trends between observed and simulated runoff Next, objective functions for event volume and event peak were created. The objective function for volume was the root mean square error between simulated and observed runoff volume. The objective function for peak was the root mean square error between the simulated and observed peak flow rate. Using the objective function for volume, first the soil depth and hydraulic conductivity were calibrated. Finally the objective function for peak was used to calibrate between channel slope and roughness Finally, evaluate Nash–Sutcliffe efficiency for each set of parameters. The Nash–Sutcliffe efficiency provides a quantification for matching the distributions of observed to simulated streamflow. The Nash–Sutcliffe efficiency metric is very sensitive to differences in shape between the observed and simulated hydrographs. Ultimately the Nash –Sutcliffe efficiency metric was used to assess the calibrated results

VUB	Gridded parameters are derived from soil texture, land use and topography, using GIS. Global model parameters are time and space invariant and are either adjustment coefficients or empirical constants that need to be preset, or calibrated if observed streamflow data are available	Start with *a priori* global model parameters, possibly adjusted by manual calibration. Optimize global model parameters using PEST program (Doherty and Johnston, 2003) minimizing the sum of squared differences between observed and calculated flow values
ILL	Geomorphologic parameters derived from DEM, soil parameters derived from STATSGO, vegetation parameters derived from MODIS/Terra production, Manning roughness coefficients are picked up from literature according to landscape properties, closure parameters representing spatial heterogeneity are *a priori* set to 1.0 which means there is no significant spatial variability. All parameters are areal averaged values at the REW scale	The same set of parameters describing key processes are applied to all REWs. We start with *a priori* parameters, geomorphologic parameters, vegetation parameters and most soil parameters are fixed during calibration, while hydraulic conductivity, Manning roughness and closure parameters are calibrated manually within restricted range based on a set of objective functions including Nash–Sutcliffe coefficient, IVF(Index of Volumetric Fit), regime curve, etc. (Tian et al., 2008). The strategy is to first calibrate the routing parameters, then baseflow parameters, and finally calibrate overland runoff parameters
Univ. AZ-1	Gridded SAC-SMA parameters and lumped routing parameters for Muskingum scheme. Initial *a priori* SAC-SMA parameters derived from soil texture (Koren et al., 2003a). Both *a priori fields* (11 pars) and routing parameters are then adjusted via Multiple-Criteria calibration using spatial regularization	Step (1) Start with a priori parameters Step (2) Devise 11 regularization equations to reduce parameter dimensionality using parameter-to-watershed physical properties and parameter-to-parameter relationships. This results in 35 super-parameters to be calibrated (11 par_fields $\times$ 3 reg_pars/field + 2 routing_pars), which when changed act to adjust overall properties of the parameter fields while maintaining the spatial patterns of variability given by the a priori parameter fields. See Pokhrel et al. (2008) for details Step 3) Adjust the super-parameters using a multi-criteria approach that simultaneously minimizes MSE and log-MSE criteria (Pokhrel et al., 2010). This gives a Pareto-optimal (PO) solution set Step 4) Select the Pareto-solution that overall gives best monthly flow volume bias
Univ. AZ-2	Gridded *a priori* SAC-SMA parameters derived from soil texture (Koren et al., 2000). Gridded routing parameters derived from observed USGS data and geomorphologic relationships (Koren et al., 2003b; Koren et al., 2004).	Step (1) Start with a priori parameter fields Step (2) Consider three primary behavioral functions of any watershed system (overall water balance, vertical redistribution, and temporal redistribution), and identify "signature patterns" of behavior that are related to the primary watershed functions and detectable using observed precipitation–runoff data Step (3) Formulate quantitative representations of these patterns in the form of "signature measures" that summarize the relevant and useful diagnostic information present in the data. The signature measures can be easily extracted from the flow duration curve and a simplified watershed lag-time calculation. See Yilmaz et al. (2008) for details Step (4) Perform sensitivity analysis to detect and group together model parameters demonstrably related to each signature measure Step (5) Use Monte Carlo parameter sampling and a two-step, semi-automated constraining approach to adjust the 11 SAC-SMA parameter fields and 1 hydraulic routing parameter field to improve the signature measures. For each parameter field a non-linear transformation having a single parameter is used to adjust the parameter values away from their a priori estimates while preserving the spatial patterns. See Yilmaz et al. (2008) Step (6) Select the parameter set that gives the best overall signature measure performance improvement Step (7) Evaluate performance on an independent evaluation period
WHU	Gridded *a priori* infiltration parameters derived from soil texture (Chow et al., 1988). Gridded routing parameters derived from observed USGS data and inverse problem (Li, 2001a,b)	The partly parameters using scale gridded spatial values by geomorphologic relationships of the NDVI, landuse and soil texture data from USGS The second parameters start with *a priori* parameters. Revise *a priori* parameters using partly lumped calibrated parameters (derived using procedures in Li and Zhong, 2003): Automatic calibration uses objective function (Li and Zhong, 2003) Last parameters value of all flow from derived from observed USGS data and inverse problem (Li, 2001a,b)

(*continued on next page*)

Table B-1 (*continued*)

Partici-pant	Parameterization	Calibration
ICL	Uncalibrated simulations: parameter values per subcatchments (9) derived from relationships similar to Atkinson et al. (2002); Dooge (1974) and Fernandez et al. (2000)	Start with calibrated lumped parameter values Test of scalar calibration multipliers over lumped parameter estimates and a priori parameter estimates, using uniform random sampling and MOSCEM Calibration of parameter values using uniform sampling and (1) the same model parameters in all subcatchments, i.e., distributed inputs (2) different model parameters among subcatchments
DHI 1 MIKE 11	The channel component of MIKE 11 uses different levels of approximation to the St Venant equations and uses physically-based parameters; either routing parameters or resistance coefficients and cross section geometry (Havnø et al., 1995). The rainfall–runoff component of MIKE 11 is conceptual and generally requires calibration. Madsen et al. (2002), demonstrate an expert system approach to minimize the number of calibration parameters	Multi-objective automatic calibration as applied in DMIP 1 (Butts et al., 2004). The methodology is based on the SCE methods (Madsen, 2000) and is provided as a generic tool for MIKE software (Madsen, 2000, 2003). The user selects the multiple objectives, starting point and parameter bounds and the tool includes sensitivity and uncertainty analyses. The same tool is (AUTOCAL) used in both MIKE 11 and MIKE SHE
DHI 2 MIKE SHE	The rainfall–runoff process representations in MIKE SHE can be either conceptual or physics-based (Butts et al., 2004; Graham and Butts, 2006). In DMIP 2, conceptual representations were used which requires calibration. The channel component of MIKE SHE is (identical to) the MIKE 11 river component – see above	Multi-objective automatic calibration as applied in DMIP 1 (Butts et al. 2004). The methodology is based on the SCE methods (Madsen, 2000) and is provided as a generic tool for MIKE software (Madsen, 2000, 2003). The user selects the multiple objectives, starting point and parameter bounds and the tool includes sensitivity and uncertainty analyses. The same tool is (AUTOCAL) used in both MIKE 11 and MIKE SHE

Appendix C

Event-based statistics for parent basins and interior points (see Figs. C-1–C-4).

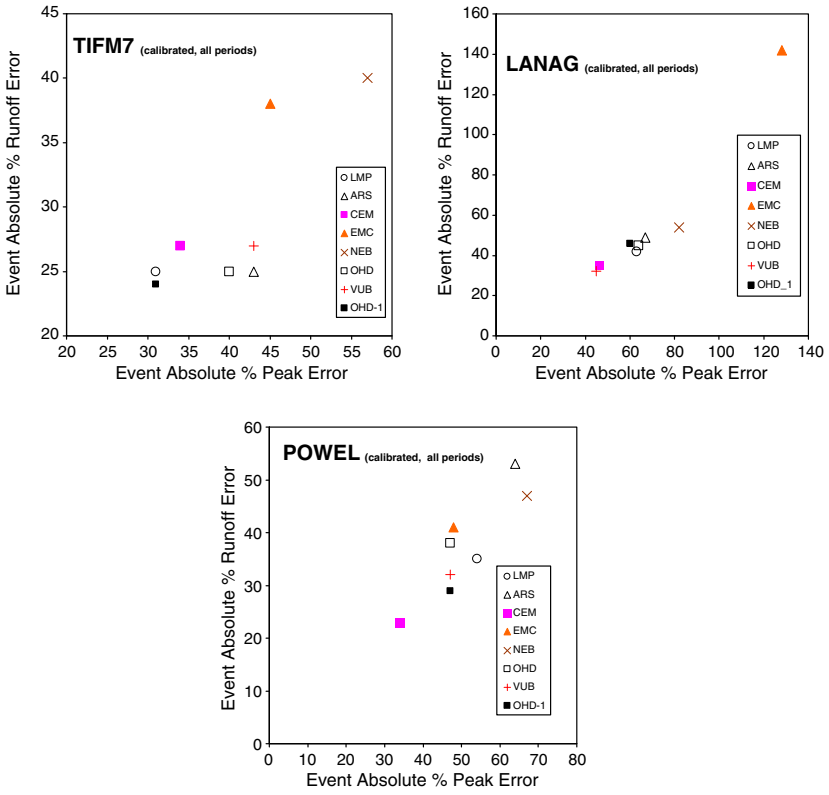


Fig. C-1. Event-based statistics for calibrated models at TIFM7 and the interior points LANAG and POWEL.

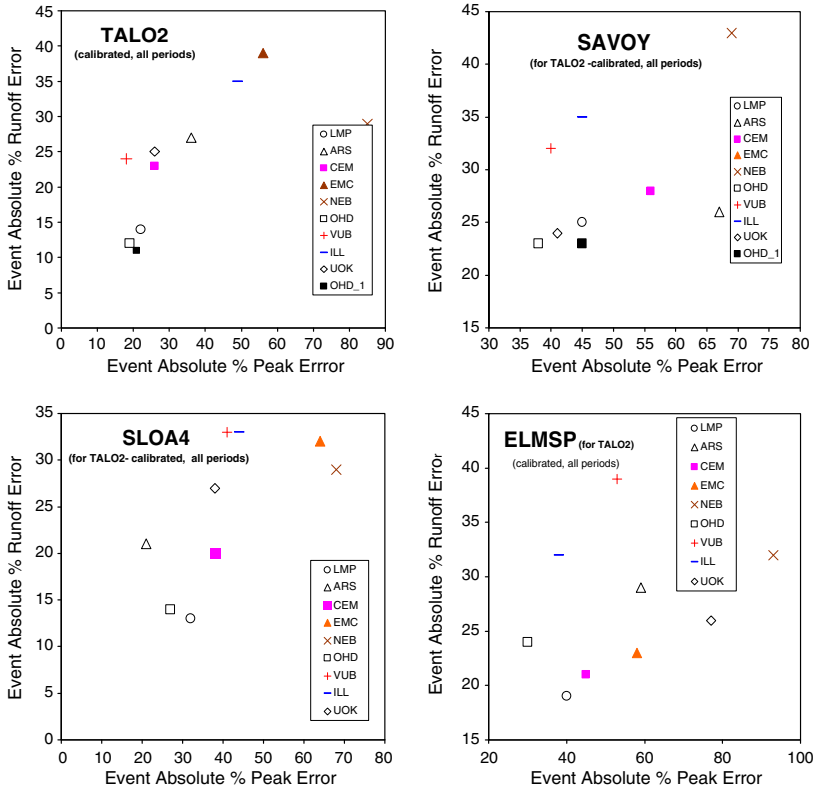


Fig. C-2. Event-based statistics for calibrated models at TALO2 and the interior points SAVOY, SLOA4, and ELMSP.

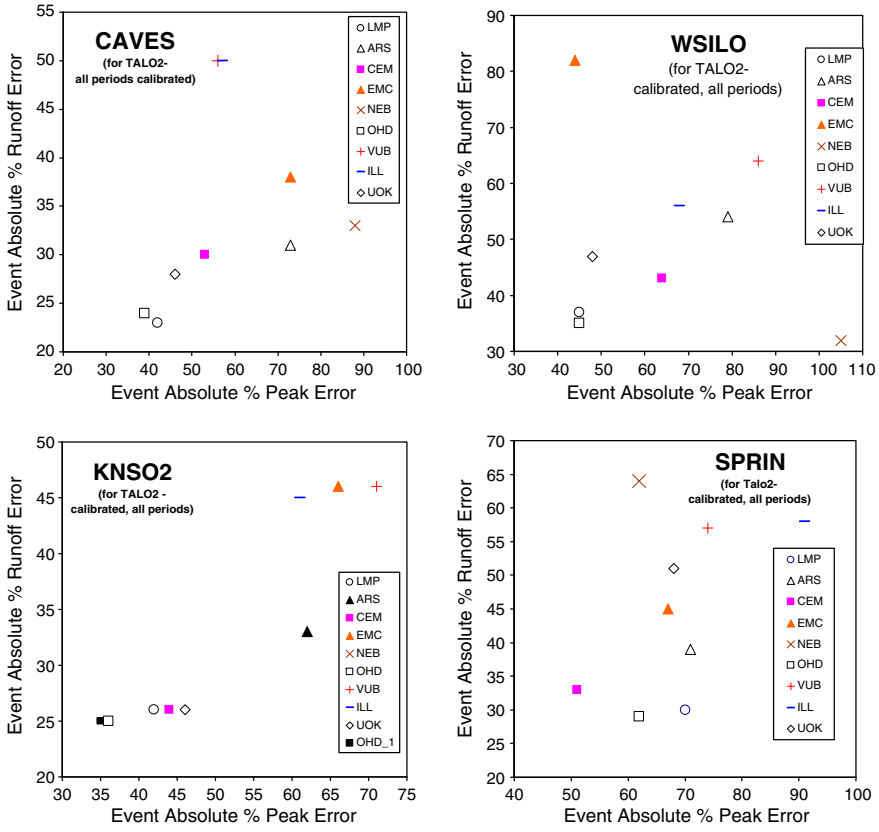


Fig. C-3. Event-based statistics for calibrated models at TALO2 and the interior points CAVES, WSILO, KNSO2, and SPRIN.

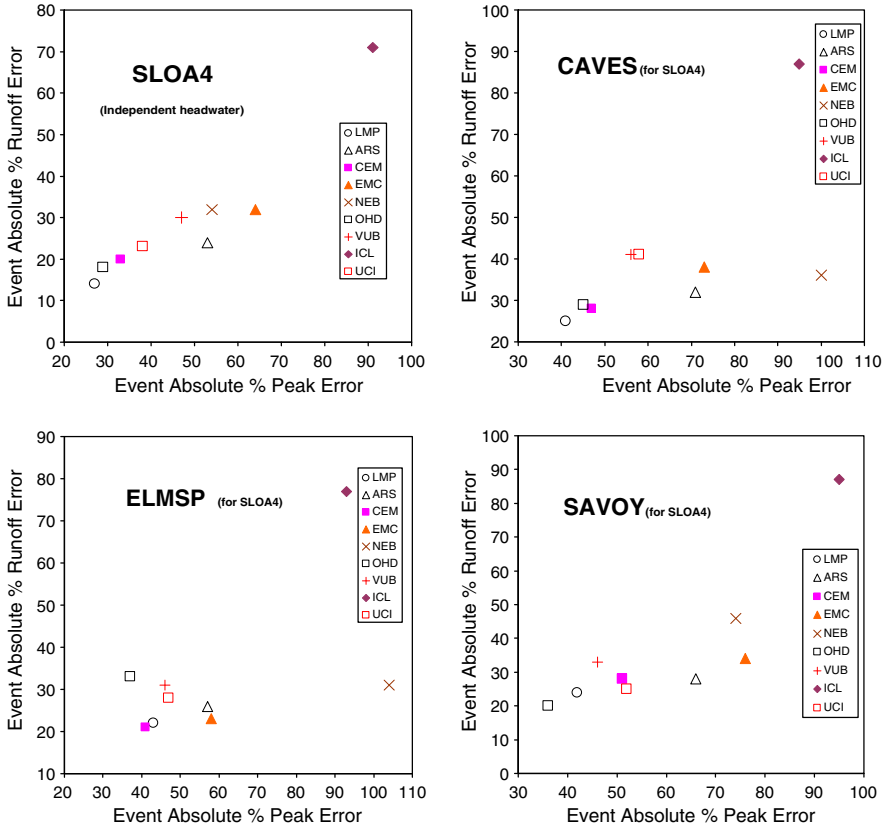


Fig. C-4. Event-based statistics for calibrated models at SLOA4 and the interior points CAVES, SAVOY, and ELMSP.

Appendix D

Change in %Bias and r_{mod} due to calibration at basin outlet (see Figs. D-1–D-4).

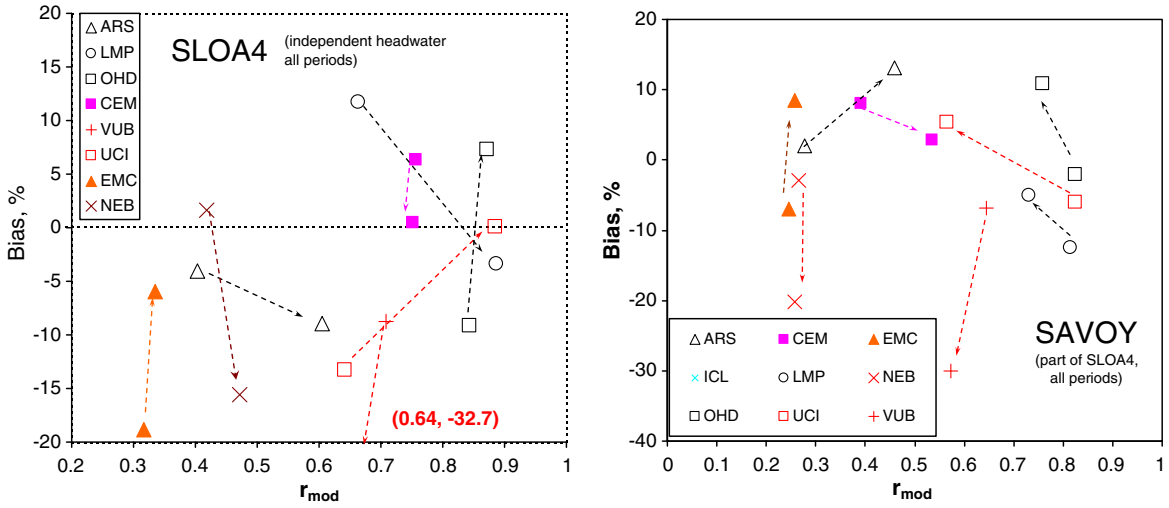


Fig. D-1. Change in %Bias and r_{mod} via calibration for SLOA4 and interior point SAVOY.

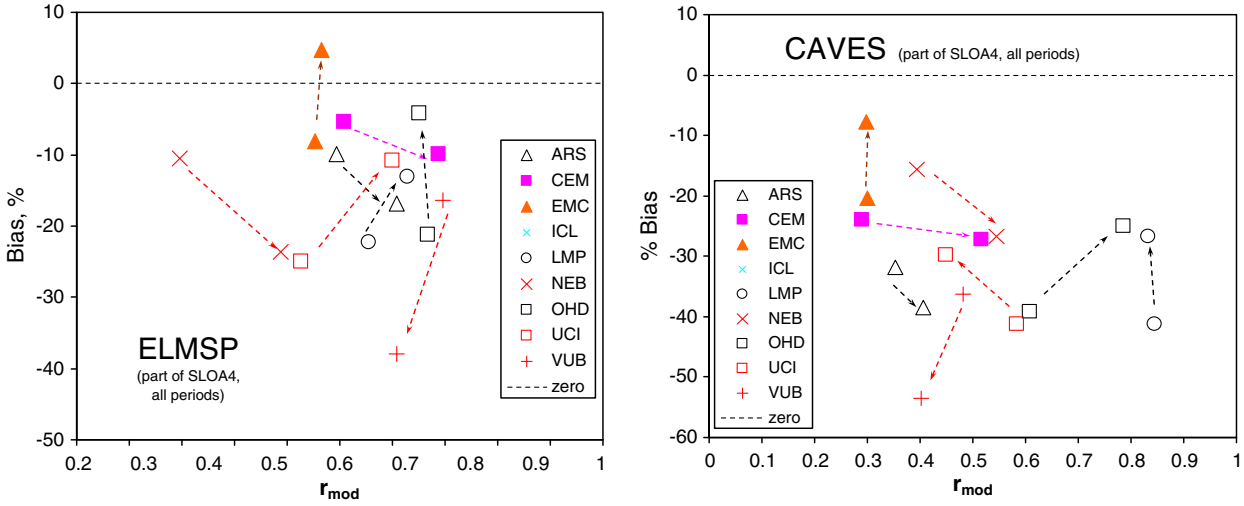


Fig. D-2. Change in %Bias and r_{mod} via calibration for SLOA4 interior points ELMSP and CAVES.

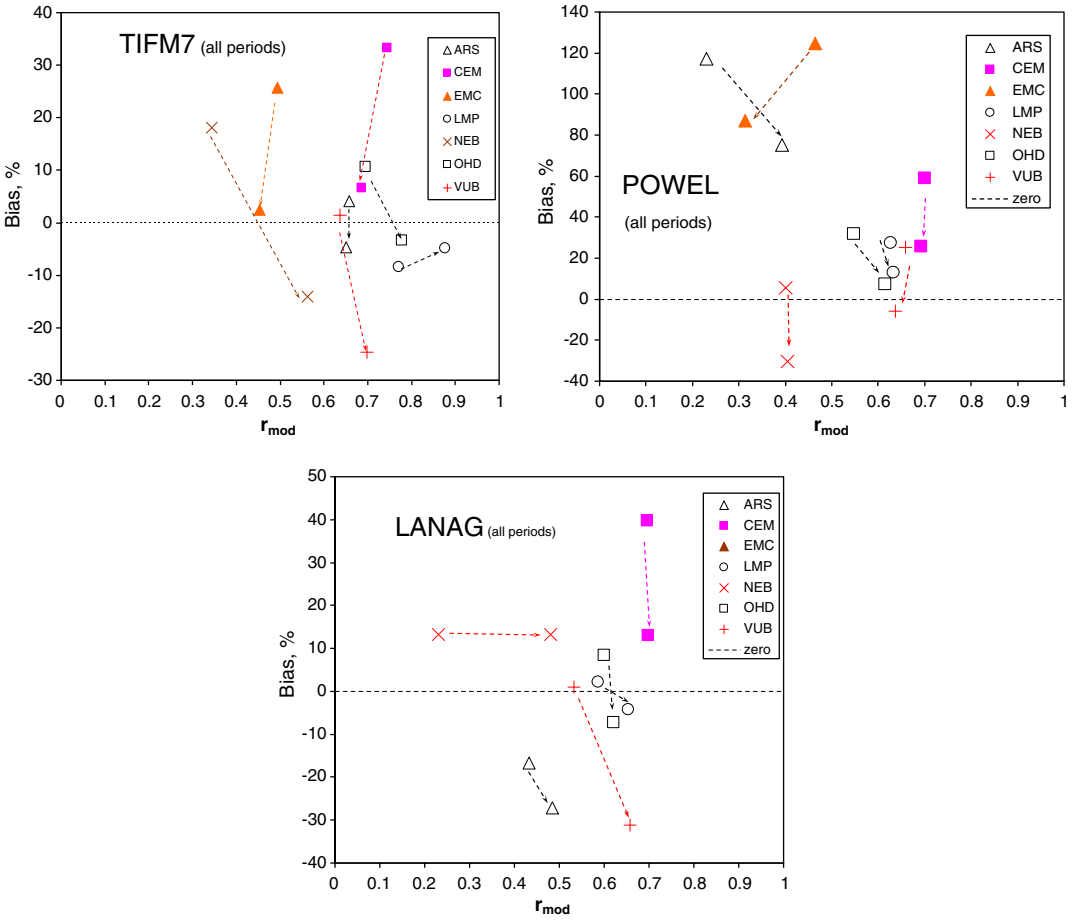


Fig. D-3. Change in %Bias and r_{mod} via calibration for TIFM7 and interior points LANAG and POWEL.

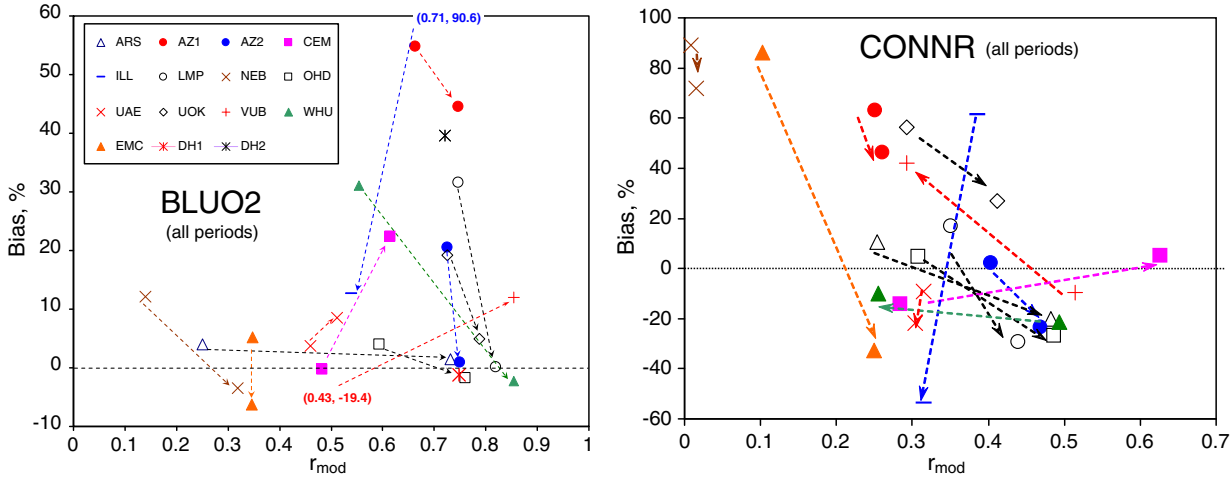


Fig. D-4. Change in %Bias and r_{mod} via calibration for BLUO2 and interior point CONNR.

Appendix E

VUB soil moisture simulation at the Westville, Oklahoma Mesonet site (see Fig. E-1).

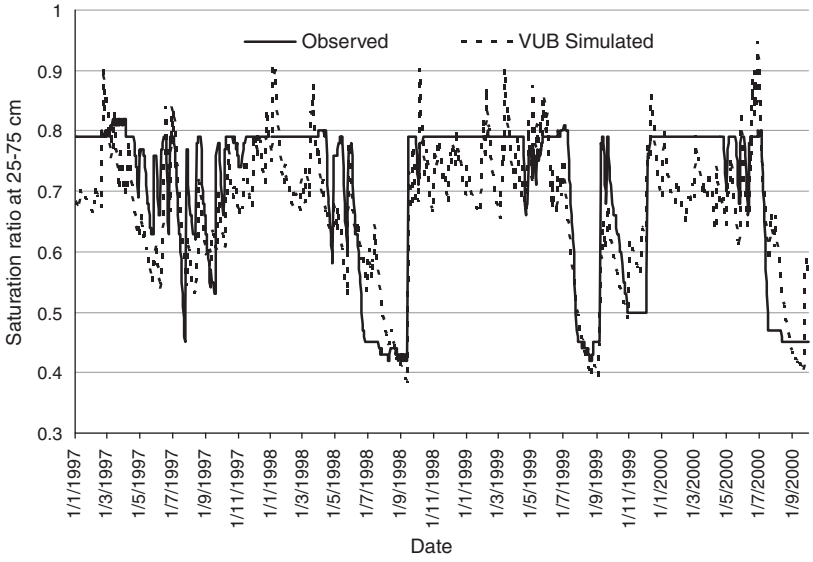


Fig. E-1. VUB Soil Moisture simulation at the Westville, Oklahoma Mesonet Site. In the VUB model, the root zone water balance is modeled in each grid cell yielding temporal and spatial variations of soil water content over the basin. The simulated hourly values at the "West" Oklahoma Mesonet site were converted using Eq. (3) and compared to observations made at 25–75 cm depth from January 1, 1997 to December 31, 2000.

References

Ajami, N.K., Gupta, H., Wagener, T., Sorooshian, S., 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. J. Hydrol. 298, 112–135.

Anderson, E.A., 2002. Calibration of conceptual hydrologic models for use in river forecasting. <http://www.weather.gov/oh/hrl/hsmb/hydrology/calibration/index.html>.

Anderson, R.M., Koren, V.I., Reed, S.M., 2006. Using SSURGO data to improve Sacramento model *a priori* parameter estimates. J. Hydrol. 320, 103–116.

Andreassian, V., Perrin, C., Michel, C., Usart-Sanchez, I., Lavabre, J., 2001. Impact of imperfect rainfall knowledge on the efficiency and the parameters of watershed models. J. Hydrol. 250 (1–4), 206–223.

Andreassian, V., Perrin, C., Michel, C., 2004. Impact of imperfect potential evapotranspiration knowledge on the efficiency and parameters of watershed models. J. Hydrol. 286, 19–35.

Andreassian, V., Hall, A., Chahinian, N., Schaake, J., (Eds.), 2006. Large Sample Basin Experiments for Hydrological Model Parameterization: Results of the Model Parameter Estimation Experiment – MOPEX. IAHS Publication 307.

Andreassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, M.-H., Valery, A., 2009. HESS Opinions: Crash test for a standardized evaluation of hydrological models. Hydrol. Earth Syst. Sci. 13, 1757–1764.

Arnold, J.G., Fohrer, N., 2005. SWAT2000: current capabilities and research opportunities in applied watershed modeling. Hydrol. Process. 19 (3), 563–572.

Atkinson, S., Woods, R.A., Sivapalan, M., 2002. Climate and landscape controls on water balance model complexity over changing time scales. Water Resour. Res. 38 (12), 1314. doi:10.1029/2002WR001487,50.1-50.17.

Bastidas, L.A., Gupta, H.V., Hsu, K.-L., Sorooshian, S., 2003. Parameter, structure, and model performance for land-surface schemes. In: Duan, Q. et al. (Eds.), Calibration of Watershed Models: Water Science and Applications, vol. 6. AGU Press, pp. 229–238.

Beven, K.J., 1982. On subsurface stormflow: an analysis of response times. Hydrol. Sci. J. 27, 505–521.

Bicknell, B.R., Imhoff, J.C., Kittle Jr., J.L., Donigian, A.S., Johanson, R.C., 1997. Hydrological Simulation Program—FORTRAN. User's Manual for Release 11. EPA-600/R-97-080, USEPA, Athens, GA, 755 p.

Biftu, G.F., Gan, T.Y., 2001. Semi-distributed, physically based, hydrologic modeling of the paddle river basin, Alberta using remotely sensed data. J. Hydrol. 244, 137–156.

Biftu, G.F., Gan, T.Y., 2004. Semi-distributed, hydrologic modeling of dry catchment with remotely sensed and digital terrain elevation data. Int. J. Remote Sens. 25 (20), 4351–4379.

Bradley, A.A., Kruger, A., 1998. Recalibration of hydrologic models for use with WSR-88D precipitation estimates. Paper 2.16, Special Symposium on Hydrology, Annual Meeting of the American Meteorology Society, Phoenix, Arizona, 11–16 January.

Brath, M., Montanari, A., Toth, E., 2004. Analysis of the effects of different scenarios of historical data availability on the calibration of a spatially-distributed hydrologic model. J. Hydrol. 291, 232–253.

Burnash, R.J.C., 1995. The NWS river forecast system – catchment modeling. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Littleton, Colorado, pp. 311–366.

Burnash, R.J.C., Ferral, R.L., McGuire, R.A., 1973. A Generalized Streamflow Simulation System Conceptual Model for Digital Computers. US Department of Commerce National Weather Service and State of California Department of Water.

Butts, M.B., Payne, J.T., Kristensen, M., Madsen, H., 2004. An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow prediction. J. Hydrol. 298 (1–4), 242–266.

Campbell Scientific 229L Water Matric Potential Sensor User Manual, 2010. On-line reference: <http://www.campbellsci.com/229-l> (accessed 14.07.10).

Carpenter, T.M., Georgakakos, K.P., 2004. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. J. Hydrol. 298, 202–221.

Chen, F., Mitchell, K.E., Schaake, J., Xue, Y., Pan, H.-L., Koren, V., Duan, Q.Y., Ek, M., Betts, A., 1996. Modeling of land-surface evaporation by four schemes and comparison with FIFE observations. J. Geophys. Res. 101, 7251–7268.

Chow, V.T., Maidment, D.R., Mays, L.W., 1988. Applied Hydrology. McGraw-Hill, Book Co., New York, p. 572.

Clark, M.P., Slater, A.G., Rupp, D.E., Woods, R.A., Vrugt, J.A., Gupta, H.V., Wagener, T., Hay, L.E., 2008. Framework for understanding structural errors (FUSE): a modular framework to diagnose differences between hydrological models. Water Resour. Res. 44, W00B02. doi:10.1029/2007WR006735.

Clark, M.P., McMillan, H.K., Collins, D.B.G., Kavetski, D., Woods, R., 2011. Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses. Hydrol. Process. 25, 523–543 (Published online 23 November 2010 in Wiley Online Library. doi:10.1002/hyp.7902).

Clarke, R.T., 2008. A critique of present procedures used to compare performance of rainfall–runoff models. J. Hydrol. 352, 379–387.

Confesor Jr., R.B., Whittaker, G.W., 2007. Automatic calibration of hydrologic models with multi-objective evolutionary algorithm and pareto optimization. J. Am. Water Resour. Assoc. (JAWRA) 43, 981–989.

Deb, K., Pratap, A., Agarwal, S., Meyarivank, T., 2002. A fast and elitist multiobjective genetic algorithm. NSGA-II. IEEE Trans. Evol. Comput. 6, 182–197.

Di Luzio, M., Arnold, J.G., 2004. Formulation of a hybrid calibration approach for a physically based distributed model with NEXRAD data input. J. Hydrol. 298, 136–154.

Di Luzio, M., Srinivasan, R., Arnold, J.G., 2004. A GIS-coupled hydrological model system for the watershed assessment of agricultural nonpoint and point sources of pollution. Trans. GIS 8 (1), 113–136.

Di Luzio, M., Johnson, G.L., Daly, C., Eischeid, J.K., Arnold, J.G., 2008. Constructing retrospective gridded daily precipitation and temperature datasets for the conterminous United States. J. Appl. Meteorol. Climatol. 47 (2), 475–497.
Doherty, J., Johnston, J.M., 2003. Methods for calibration and predictive analysis of a watershed model. J. Am.Water Resour. Assoc. 39 (2), 251–265.
Dooge, J.C.I., 1974. Linear Theory of Hydrologic Systems. Technical Bull. 1468, US Department of Agriculture, 231 pp.
Dooge, J.C.I., 1992. Sensitivity of runoff to climate change: a Hortonian approach. Bull. Am. Meteorol. Soc. 73 (12), 2013–2024.
Duan, Q.Y., Schaake, J.C., Koren, V.I., 1996. FIFE 1987 water budget analysis. J. Geophys. Res. 101 (D3), 7197–7207.
Duan, Q., Schaake, J., Andréassian, V., Franks, S., Goteti, G., Gupta, H.V., Gusev, Y.M., Habets, F., Hall, A., Hay, L., Hogue, T., Huang, M., Leavesley, G., Liang, X., Nasonova, O.N., Noilhan, J., Oudin, L., Sorooshian, S., Wagener, T., Wood, E.F., 2006. Model parameter estimation experiment (MOPEX): an overview of science strategy and major results from the second and third workshops. J. Hydrol. 320, 3–17.
Ek, M., Mitchell, K.E., Lin, Y., Rogers, E., Grunmann, P., Koren, V., Gayno, G., Tarpley, G.D., 2003. Implementation of Noah land-surface model advances in the NCEP operational mesoscale Eta model. J. Geophys. Res. 108 (D22), 8851. doi:10.1029/2002JD003296.
EPA BASINS, 2001. Better assessment Science Integrating Point and Nonpoint Sources. BASINS V. 3.0 User's Manual.
EPA HSPF, 1970. Hydrologic Simulation Program-Fortran. HSPF's User Manual.
Farnsworth, R.K., Thompson, E.S., Peck, E.L., 1982. Evaporation Atlas for the Contiguous 48 United States. NOAA Technical Report NWS 33, 27 pp.
Fernandez, W., Vogel, R.M., Sankarasubramanian, A., 2000. Regional calibration of a watershed model. J. Hydrol. Sci. 45 (5), 670–689.
Gan, T.Y., Dlamini, E.M., Biftu, G.F., 1997. Effects of model complexity and structure, data quality, and objective functions on hydrologic modeling. J. Hydrol. 192, 81–103.
Gassman, P.W., Reyes, M.R., Green, C.H., Arnold, J.G., 2007. The soil and water assessment tool: historical development, applications, and future research directions. Trans. Am. Soc. Agr. Biol. Eng. 50 (4), 1211–1250, ISSN 0001-2351.
Ghizzoni, T., Giannoni, F., Roth, G., Rudari, R., 2007. The role of observation uncertainty in the calibration of hydrologic rainfall–runoff models. Adv. Geosci. 12, 33–38.
Graham, D.N., Butts, M.B., 2006. Flexible, integrated watershed modelling with MIKE SHE. In: Singh, V.P., Frevert, D.K. (Eds.), Watershed Models. CRC Press, pp. 245–272, ISBN: 0849336090.
Gu, Y., Hunt, E., Wardlow, B., Basara, J.B., Brown, J.F., Verdin, J.P., 2008. Evaluation of MODIS NDVI and NDWI for vegetation drought monitoring using Oklahoma Mesonet soil moisture data. Geophys. Res. Lett. 35, L222401. doi:10.1029/2008GL035772.
Guentchev, G., Barsugli, J.J., Eischeid, J., 2010. Homogeneity of gridded precipitation datasets for the Colorado River basin. J. Appl. Meteorol. Climatol. 49, 2404–2415. doi:10.1175/2010JAMC2484.1.
Guo, J., Liang, X., Leung, L.R., 2004. Impacts of different precipitation data sources on water budgets. J. Hydrol. 298, 311–334.
Gupta, H.V., Sorooshian, S., Hogue, T.S., Boyle, D.P., 2003. Advances in automatic calibration of watershed models. In: Duan, Q., Sorooshian, S., Gupta, H., Rosseau, H., Turcotte, R. (Eds.), Advances in Calibration of Watershed Models. Calibration of Watershed Models, Water Science and Applications, vol. 6. AGU, pp. 197–211.
Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process. 22, 3802–3813.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez-Baquero, G.F., 2009. Decomposition of the mean squared error & NSE performance criteria: implications for improving hydrological modeling. J. Hydrol. 377, 80–91. doi:10.1016/j.jhydrol.2009.08.003.
Hamlet, A.F., Lettenmaier, D.P., 2005. Production of temporally consistent gridded precipitation and temperature fields for the continental United States. J. Hydrometeorol. 6 (3), 330–336.
Havnø, K., Madsen, M.N., Dørge, J., 1995. MIKE 11—a generalized river modelling package. In: Singh, V.P. (Ed.), Computer Models of Watershed Hydrology. Water Resources Publications, Colorado, USA, pp. 733–782.
Hogue, T.S., Gupta, H.V., Sorooshian, S., Tomkins, C.D., 2003. A multi-step automatic calibration scheme for watershed models. In: Duan, Q. et al. (Eds.), Calibration of Watershed Models: Water Science and Applications, vol. 6. AGU Press, pp. 165–174.
Hong, S.-Y., Kalnay, E., 2002. The 1998 Oklahoma–Texas drought: mechanistic experiments with NCEP global and regional models. J. Clim. 15, 945–963.
Illston, B.G., Basara, J.B., 2002. A Soil Moisture Analysis of drought conditions using the Oklahoma Mesonet. In: Preprints, 13th Conference on Applied Climatology; American Meteorological Society, Portland, Oregon, May 13–16.
Illston, B.G., Basara, J.B., Crawford, K.C., 2003. Soil moisture observations from the Oklahoma Mesonet. GEWEX News 13 (3), 13–14.
Illston, B.G., Basara, J.B., Crawford, K.C., 2004. Seasonal to interannual variations of soil moisture measured in Oklahoma. Int. J. Climatol. 24, 1883–1896.
Illston, B.G., Basara, J.B., Fisher, D.K., Elliott, R., Fiebrich, C.A., Crawford, K.C., Humes, K., Hunt, E., 2008. Mesoscale monitoring of soil moisture across a statewide network. J. Atmos. Oceanic Technol. 25, 167–182.
Ivanov, Y.V., Vivoni, E.R., Bras, R.L., Entekhabi, D., 2004. Preserving high-resolution surface and rainfall data in operational-scale basin hydrology: a fully-distributed physically-based approach. J. Hydrol. 298, 80–111.
Ivanov, V.Y., Bras, R.L., Vivoni, E.R., 2008. Vegetation-hydrology dynamics in complex terrain of semiarid areas: 1. A mechanistic approach to modeling dynamic feedback. Water Resour. Res. 44, W03429. doi:10:1029/2006WR005588.
Jain, S.K., Sudheer, K.P., 2008. Fitting of hydrologic models: a close look at the Nash–Sutcliffe index. ASCE J. Hydrol. Eng. 13 (10), 981–986.
Jones, E.T., Roth, K., Costanza, K., 2009. A comparison of the NWS Distributed versus lumped hydrologic model. In: AMS 23 Conference on Hydrology, Paper 4.3, Phoenix Arizona, January 10–15.
Kalinga, O.A., Gan, T.Y., 2006. Semi-distributed modeling of basin hydrology with radar and gauged precipitation. Hydrol. Process. 20, 3725–3746.
Kampf, S.K., Burges, S.J., 2007. A framework for classifying and comparing distributed hillslope and catchment hydrologic models. Water Resour. Res. 43, W05423. doi:10.1029/2006WR005370.
Khakbaz, B., Imam, B., Hsu, K., Sorooshian, S., this issue. From lumped to distributed via semi-distributed: calibration strategies for semi-distributed hydrologic models. J. Hydrol. doi:10.1016/j.jhydrol.2009.02.021.
Kirchner, J.W., 2006. Getting the right answers for the right reasons: linking measurements, analyses, and models to advance the science of hydrology. Water Resour. Res. 42, W03S04. doi:10.1029/2005WR004362.
Koren, V., 2006. Parameterization of frozen ground effects: sensitivity to soil properties predictions in ungauged basins: promises and progress. In: Proceedings of Symposium S7 held during the Seventh IAHS Scientific Assembly at Foz do Iguaçu, Brazil, April 2005. IAHS Publ. 303, pp. 125–133.
Koren, V., Barrett, C.B., 1994. Satellite based distributed monitoring, forecasting, and simulation (MFS) system for the Nile River. Application of remote sensing in hydrology. Proceedings of the Second International Workshop, NHRI Symposium No. 14, October, 1994, Saskatoon, Canada.
Koren, V., Schaake, J., Mitchell, K.E., Duan, Q.-Y., Chen, F., Baker, J., 1999. A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. J. Geophys. Res. 104, 19569–19585.
Koren, V., Smith, M., Wang, D., Zhang, Z., 2000. Use of soil property data in the derivation of conceptual rainfall–runoff model parameters. In: Proceedings of the 15th Conference on Hydrology, Long Beach, CA, Amer. Meteor. Soc. 10–14 January, 2000, pp. 103–106.
Koren, V., Smith, M., Duan, Q., 2003. Use of *a priori* parameter estimates in the derivation of spatially consistent parameter sets of rainfall–runoff models. In: Duan, Q. et al. (Eds.), Calibration of Watershed Models: Water Science and Applications, vol. 6. AGU Press, pp. 239–254.
Koren, V., Reed, S., Smith, M., Zhang, Z., 2003b. Combining physically-based and conceptual approaches in the development and parameterization of a distributed system. Weather radar information and distributed hydrological modeling. In: Proceedings of Symposium HSO_3 held during IUGG2003 at Sapporo, July 2003, Publication No. 282, pp. 101–108.
Koren, V., Reed, S., Smith, M., Zhang, Z., Seo, D.-J., 2004. Hydrology laboratory research modeling system (HL-RMS) of the US National Weather Service. J. Hydrol. 291, 297–318.
Koren, V., Moreda, F., Reed, S., Smith, M., Zhang, Z., 2006. Evaluation of a grid-based distributed hydrological model over a large area. Predictions in ungauged basins: promise and progress. In: Proceedings of Symposium S7 held during the Seventh IAHS Scientific Assembly at Foz do Iguacu, Brazil, April 2005. IAHS Publication 303, pp. 47–56.
Koren, V., Smith, M., Cui, Z., and Cosgrove, B., 2007. Physically-Based Modifications to the Sacramento Soil Moisture Accounting Model: Modeling the Effects of Frozen Ground on the Rainfall–Runoff Process. NOAA Technical Report NWS 52, <http://www.weather.gov/oh/hrl/hsmb/docs/hydrology/PBE_SAC-SMA/NOAA_Technical_Report_NWS_52.pdf>.
Koren, V., Moreda, M., Smith, M., 2008. Use of soil moisture observations to improve parameter consistency in watershed calibration. Phys. Chem. Earth 33, 1068–1080.
Kuzmin, V., Seo, D.-J., Koren, V., 2008. Fast and efficient optimization of hydrologic model parameters using *a priori* estimates and stepwise line search. J. Hydrol. 353, 109–128.
Lee, H., Seo, D.J., Koren,V., in press. Assimilation of streamflow and soil moisture data into operational distributed hydrologic models: Effects of uncertainties in the data and initial model soil moisture states. Adv. Water Resour. doi.org/10.1016/j.advwatres.2011.08.012
Li, L., 2001a. A distributed dynamic parameters inverse model for rainfall–runoff. Soil-vegetation-atmosphere transfer schemes and large-scale hydrological models. In: Proceedings of a Symposium Held during the Sixth IAHS Scientific Assembly at Maastricht, The Netherlands, July, 2001. IAHS Pub. No. 270, pp. 103–112.
Li, L., 2001b. A physically-based rainfall–runoff model and distributed dynamic hybrid control inverse technique. Soil-vegetation-atmosphere transfer schemes and large-scale hydrological models. In: Proceedings of a Symposium Held during the Sixth IAHS Scientific Assembly at Maastricht, The Netherlands, July, 2001. IAHS Pub. No. 270, pp. 135–142.
Li, L., Zhong, M., 2003. Structure of the LL-II distributed rainfall–runoff model based on GIS. Water Resour. Power 21 (4), 35–38 (in Chinese).
Liu, Y.B., De Smedt, F., 2004. WetSpa Extension: Documentation and User Manual. Department of Hydrology and Hydraulic Engineering, Vrije Universiteit Brussel, Belgium, 108 pp.
Lohmann, D., Lettenmaier, D.P., Liang, X., Wood, E.F., Boone, A., Chang, S., Chen, F., Dai, Y., Desborough, C., Dickinson, R.E., Duan, Q., Ek, M., Gusev, Y.M., Habets, F., Irannejad, P., Koster, R., Mitchell, K.E., Nasonova, O.N., Noilhan, J., Schaake, J., Schlosser, A., Shao, Y., Shmakin, A.B., Verseghy, D., Warrach, K., Wetzel, P., Xue,

Y., Yang, Z.-L., Zeng, Q.C., 1998. The project for intercomparison of land-surface parameterization schemes (PILPS) phase 2(c) Red–Arkansas River basin experiment: 3. Spatial and temporal analysis of water fluxes. Global Planet. Change 19, 161–179.
Lohmann, D. et al., 2004. Streamflow and water balance intercomparisons of four land surface models in the North American land data assimilation system project. J. Geophys. Res. 109, D07S91. doi:10.1029/2003JD003517.
Looper, J.P., Vieux, B.E., Morena, M.A., this issue. Assessing the impacts of precipitation bias on distributed hydrologic model calibration and prediction accuracy. J. Hydrol. doi:10.1016/j.jhydrol.2009.09.048.
Madsen, H., 2000. Automatic calibration of a conceptual rainfall–runoff model using multiple objectives. J. Hydrol. 235, 276–288.
Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv. Water Resour. 26, 205–216.
Madsen, H., Wilson, G., Ammentorp, H.C., 2002. Comparison of different automated strategies for calibration of rainfall–runoff models. J. Hydrol. 261 (1–4), 48–59.
Marquardt, D.W., 1963. An algorithm for least-squares estimation of nonlinear parameters. J. Soc. Ind. Appl. Math. 11 (2), 431–441.
Martinez, G.F., Gupta, H.V., 2010. Towards improved identification of hydrologic models: continental scale diagnostic evaluation of the 'abcd' monthly water balance model for the conterminous US. Water Resour. Res. 46 (W08507), 21. doi:10.1029/2009WR008294.
Mascaro, G., Vivoni, E.R., Deidda, R., 2010. Implications of ensemble quantitative precipitation forecast errors on distributed streamflow forecasting. J. Hydrometeorol. 11, 69–86.
McCuen, R.H., Snyder, W.M., 1975. A proposed index for comparing hydrographs. Water Resour. Res. 11 (6), 1021–1024.
Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., et al., 2006. North american regional reanalysis (NARR). Bull. Am. Meteorol. Soc. 87 (3), 343–360.
Michel, C., Perrin, C., Andreassian, V., Oudin, L., Methevet, T., 2006. Has basin- scale modeling advanced beyond empiricism? In: Large Sample Basin Experiments of Hydrological Model Parameterization: Results of the Model Parameter Estimation Experiment – MOPEX; IAHS Publication 307, pp. 108–114.
Milly, P.C.D., 1994. Climate, soil water storage, and the average annual water balance. Water Resour. Res. 30 (7), 2143–2156.
Milly, P.C.D., Betancourt, J., Falkenmark, M., Hirsch, R.M., Kundzewicz, Z.W., Lettenmaire, D.P., Stouffer, R.J., 2008. Stationarity is dead: whither water management? Science 319, 573–574.
Mitchell, K.E. et al., 2004. The multi-institutional North American land data assimilation system (NLDAS): utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. J. Geophys. Res. 109, D07S90. doi:10.1029/2003JD003823.
Mizukami, N., Koren, V., 2008. Methodology and evaluation for melt factor parameterization for distributed Snow-17. American Geophysical Union, Fall Meeting 2008, abstract H31J-08, San Francisco.
Moore, R.J., 1985. The probability-distributed principle and runoff production at point and basin scales. J. Hydrol. Sci. 30 (2), 273–297.
Moriasi, D.N., Starks, P.J., 2010. Effects of the resolution of soil dataset and precipitation dataset on SWAT2005 streamflow calibration parameters and simulation accuracy. J. Soil Water Conserv. 65 (2), 63–78.
Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models. Part I, a discussion of principles. J. Hydrol. 10, 282–290.
Nelson, B.R., Seo, D.-J., Kim, D., 2010. Multi-sensor precipitation reanalysis. J. Hydrometeorol. 11, 666–682. doi:10.1175/2010JHM1210.1.
Osborn, N. I., 2009. Arubckle-Simpson Hydrology Study – Final Report to the US Bureau of Reclamation, prepared by the Oklahoma Water Resources Board, December, Cooperative Agreement No. 03FC601814, 42 pp.
Oudin, L., Perrin, C., Mathevet, T., Andreassian, V., Michel, C., 2006. Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. J. Hydrol. 320, 62–83.
Oudin, L., Andréassian, V., Perrin, C., Michel, C., Le Moine, N., 2008. Spatial proximity, physical similarity and ungauged catchments: confrontation on 913 French catchments. Water Resour. Res. 44, W03413. doi:10.1029/2007WR006240.
Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279 (1–4), 275–289.
Perrin, C., Andréassian, V., Serna, C.R., Mathevet, T., Le Moine, N., 2008. Discrete parameterization of hydrological models: evaluating the use of parameter sets libraries over 900 catchments. Water Resour. Res. 44, W08447. doi:10.1029/2007WR006579.
Peschel, J.M., Haan, P.K., Lacey, R.E., 2006. Influences of soil dataset resolution on hydrologic modeling. J. Am.Water Resour. Assoc., 1371–1389.
Pokhrel, P., Gupta, H.V., in press. On the ability to infer spatial catchment variability using streamflow hydrographs. Water Resour. Res. doi:10.1029/2010WR009873.
Pokhrel, P., Gupta, H.V., Wagener, T., 2008. A spatial regularization approach to parameter estimation for a distributed watershed model. Water Resour. Res. 44, W12419. doi:10.1029/2007WR006615.
Pokhrel, P., Yilmaz, K.K., Gupta, H.V., 2010. Multiple-criteria calibration of a distributed watershed model using spatial regularization and response signatures. J. Hydrol. doi:10.1016/j.jhydrol.2008.12.004.
Rabuffetti, D., Ravazzani, G., Barbero, S., Mancini, M., 2009. Operationl flood-forecasting in the Piemonte region – development and verification of a fully distributed physically-oriented hydrological model. Adv. Geosci. 17, 111–117.
Rawls, W.J., Brakensiek, D.L., 1985. Prediction of Soil Water Properties for Hydrologic Modeling. Watershed Management in the Eighties. ASCE, pp. 293–299.
Rawls, W.J., Brakensiek, D.L., Miller, N., 1983a. Predicting Green and Ampt infiltration parameters from soils data. J. Hydraul. Eng. 109 (1), 62–70.
Rawls, W.J., Brakensiek, D.L., Soni, B., 1983b. Agricultural management effects on soil water processes, Part I: soil water retention and Green and Ampt infiltration parameters. Trans. Am. Soc. Agric. Eng. 26 (6), 1747–1752.
Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., DMIP Participants, 2004. Overall distributed model intercomparison project results. J. Hydrol. 298 (1–4), 27–60.
Reed, S., Schaake, J., Zhang, Z., 2007. A distributed hydrologic model and threshold frequency-based method for flash flood forecasting at ungauged locations. J. Hydrol. 337, 402–420.
Rosero, E., Yang, Z.-L., Gulden, L.E., Niu, G.Y., Gochis, D.J., 2009. Evaluating enhanced hydrologica representation in Noah LSM over transition zones: implications for model development. J. Hydrometeorol. 10, 600–622. doi:10:1175/2009JHM1029.1.
Ryu, J.H., 2009. Application of HSPF to the distributed model intercomparison project: case study. J. Hydrol. Eng. 14 (8), 847–857.
Sankarasubramanian, A., Vogel, R.M., 2002. Annual hydroclimatology of the United States. Water Resour. Res. 38 (6). doi:10.1029/2001WR000619.
Schaake, J.C., Koren, V., Duan, Q.-Y., Mitchell, K.E., Chen, F., 1996. Simple water balance model for estimating runoff at different spatial and temporal scales. J. Geophys. Res. 101, 7461–7475.
Schaake, J.C., Duan, Q., Koren, V., Mitchell, K.E., Houser, P.R., Wood, E.F., Robock, A., Lettenmaier, D.P., Lohmann, D., Cosgrove, B., Sheffield, J., Luo, L., Higgins, R.W., Pinker, R.T., Tarpley, J.D., 2004. An intercomparison of soil moisture fields in the North American land data assimilation system (NLDAS). J. Geophys. Res. 109, D01S90. doi:10.1029/2002JD003309.
Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2208 (simultaneously published online as Invited Commentary in Hydrologic Processes (HP Today), Wiley InterScience, doi:10.1002/hyp.6825).
Schmidt, J.A., Anderson, A.J., Paul, J.H., 2007. Spatially-variable, physically-derived flash flood guidance. In: Paper 6B.2, Proceedings of the 21st Conference on Hydrology, 87th Meeting of the AMS, San Antonio, Texas, January 15–18.
Schneider, J.M., Fisher, D.K., Elliott, R.L., Brown, G.O., Bahrmann, C.P., 2003. Spatiotemporal variations in soil water: first results from the ARM SGP CART network. J. Hydrometeorol. 4, 106–120.
Seibert, J., 2001. On the need for benchmarks in hydrological modeling. Hydrol. Process. 15 (6), 1063–1064.
Shao, Y., Henderson-Sellers, A., 1996. Modeling soil moisture: a project for intercomparison of land surface parameterization schemes phase 2(b). J. Geophys. Res. 101 (D23), 7227–7250.
Smith, M.B., Laurine, D.P., Koren, V.I., Reed, S.M., Zhang, Z., 2003. Hydrologic model calibration in the National Weather Service. In: Duan, Q. et al. (Eds.), Calibration of Watershed Models: Water Science and Applications, vol. 6. AGU Press, pp. 133–152.
Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S., Zhang, Z., Duan, Q.-Y., Moreda, F., Cong, S., 2004a. The distributed model intercomparison project (DMIP): motivation and experiment design. J. Hydrol. 298 (1–4), 4–26.
Smith, M.B., Koren, V.I., Zhang, Z., Reed, S.M., Pan, J.-J., Moreda, F., 2004b. Runoff response to spatial variability in precipitation: an analysis of observed data. J. Hydrol. 298, 267–286.
Smith, M., Koren, V., Reed, S., Zhang, Z., Zhang, Y., Moreda, F., Cui, Z., Mizukami, N. Anderson, E. and Cosgrove, B.A., this issue. The distributed model intercomparison project – Phase 2: Motivation and design of the Oklahoma experiments. J. Hydrol. doi:10.1016/j.jhydrol.2011.08.055.
Sorooshian, S., Gupta, V.K., Fulton, J.L., 1983. Evaluation of maximum likelihood parameter estimation techniques for conceptual rainfall–runoff models: influence of calibration data variability and length on model credibility. Water Resour. Res. 19, 251–259.
Sridhar, V., Hubbard, K.G., You, J., Hunt, E.D., 2008. Development of the soil moisture index to quantify agricultural drought and its "user friendliness" in severity-area-duration assessment. J. Hydrometeorol. 9, 660–676.
Tian, F., Hu, H., Lei, Z., Sivapalan, M., 2006. Extension of the representative elementary watershed approach for cold regions via explicit treatment of energy related processes. Hydrol. Earth Syst. Sci. 10, 619–644.
Tian, R., Hu, H., Lei, Z., 2008. Thermodynamic watershed hydrological model: constitutive relationship. Sci. China Ser. E: Technol. Sci. 51 (9), 1353–1369. doi:10.1007/s11431-008-0147-0.
Timbal, B., Henderson-Sellers, A., 1998. Intercomparisons of land-surface parameterizations coupled to a limited area forecast model. Global Planet. Change 19, 247–260.
Turcotte, R., Rousseau, A.N., Fortin, J.-P., Villeneuve, J.-P., 2003. A process-oriented, multi-objective calibration strategy accounting for model structure. In: Duan, Q. et al. (Eds.), Calibration of Watershed Models: Water Science and Applications, vol. 6. AGU Press, pp. 153–165.
Van Werkhoven, K., Wagener, T., Reed, P., Tang, Y., 2008. Rainfall characteristics define the value of streamflow observations for distributed watershed model identification. Geophys. Res. Lett. 35, L11403. doi:10.1029/2008GL034162.
Vieux, B.E., 2004. Distributed Hydrologic Modeling Using GIS. second ed. Water Science Technology Series, vol. 48. Kluwer Academic Publishers, Norwell, Massachusetts, 289 pp.

Vieux, B.E., Vieux, J.E., Chen, C., Howard, K.W., 2003. Operational deployment of a physics-based distributed rainfall–runoff model for flood forecasting in Taiwan. IASH General Assembly at Sapporo, Japan July 3–11.

Viney, N.R., Croke, B.F.W, Breuer, L., Bormann, H., Bronstert, A., Frede, H., Graff, T., Hubrechts, L, Huisman, J.A., Jakeman, A.J., Kite, G.W., Lanini, J., Leavesley, G., Lettenmaier, D.P., Lindstrom, G., Seibert, J., Sivapalan, M., Willems, P., 2006. Ensemble modelling of the hydrological impacts of land use change. In: Zerger, A., Argent, R.M. (Eds.), MODSIM 2005 International Congress on Modelling and Simulation. Modelling and Simulation Society of Australia and New Zealand, December 2005, pp. 2967–2973. ISBN: 0-9758400-2-9.

Vivoni, E.R., Entekhabi, D., Bras, R.L., Ivanov, V.Y., Van Horne, P., Grassotti, C., Hoffman, R.N., 2006. Extending the predictability of hydrometeorolgical flood events using radar rainfall nowcasting. J. Hydrometeorol. 7, 660–677.

Wagener, T., Sivapalan, M., Troch, P., Woods, R., 2007. Catchment classification and hydrologic similarity. Geogr. Compass 1/4, 901–931.

Wang, X., Melesse, A.M., 2006. Effects of STATSGO and SSURGO as inputs on SWAT model's snowmelt simulation. J. Am. Water Resour. Assoc., 1217–1236.

Williamson, T., Odom, K.R., 2007. Implications of SSURGO vs. STATSGO data for modeling daily streamflow in Kentucky. In: ASA-CSSA-SSSA 2007 International Annual Meetings, November 4–8, 2007, New Orleans, Louisiana.

Wollock, D.M., McCabe, G.M., 1999. Explaining spatial variability in mean annual runoff in the conterminous United States. Clim. Res. 11, 149–159.

Wood, E., Lettenmaier, D.P., Liang, X., Lohmann, D., Boone, A., Chang, S., Chen, F., Dai, Y., Dickinson, R.E., Duan, Q., Ek, M., Susev, Y.M., Habets, F., Irannejad, P., Koster, R., Mitchel, K.E., Nasonova, O.N., Noilhan, J., Schaake, J., Schlosser, A., Shao, Y., Shmakin, A.B., Verseghy, D., Warrach, K., Wetzel, P., Xue, Y., Yang, Z.-L., Zeng, Q., 1998. The project for intercomparison of land surface parameterization schemes (PILPS) phase 2(c) Red–Arkansas River basin experiment: 1. Experiment description and summary intercomparisons. Global Planet. Change 19, 115–136.

Xu, C.-Y., Vandewiele, G.L., 1994. Sensitivity of monthly rainfall–runoff models to input errors and data length. Hydrol. Sci. J. 39 (2), 157–176.

Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: application to the NWS distributed hydrologic model. Water Resour. Res. 44, W09417. doi:10.1029/2007WR006716.

Young, C.B., Bradley, A.A., Krajewski, W.F., Kruger, A., 2000. Evaluating NEXRAD multi-sensor precipitation estimates for operational hydrologic forecasting. J. Hydrometeorol. 1, 241–254.

Zhang, L., Dawes, W.R., Walker, G.R., 2001. Response of mean annual evapotranspiration to vegetation changes at catchment scale. Water Resour. Res. 37 (3), 701–708.

Zhang, Y., Reed, S., Kitzmiller, D., 2011. Effects of gauge-based readjustment of historical multi-sensor precipitation on hydrologic simulations. J. Hydrometeorol. 12 (3), 429–443. doi:10.1175/2010JHM1200.1.

Zhang, Z., Koren, V., Reed, S., Smith, M., Zhang, Y., Moreda, F., Cosgrove, B., in preparation. SAC-SMA *a priori* parameter differences and their impact on distributed hydrologic model simulations. J. Hydrol.