UC Irvine
UC Irvine Previously Published Works

Title
The distributed model intercomparison project – Phase 2: Experiment design and summary results of the western basin experiments

Permalink
https://escholarship.org/uc/item/4d45j6x0

Authors
Smith, Michael
Koren, Victor
Zhang, Ziya
et al.

Publication Date
2013-12-01

DOI
10.1016/j.jhydrol.2013.08.040

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, availalbe at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Journal of Hydrology 507 (2013) 300–329

Contents lists available at ScienceDirect

Journal of Hydrology

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

The distributed model intercomparison project – Phase 2: Experiment design and summary results of the western basin experiments

Michael Smith [a,*], Victor Koren [a], Ziya Zhang [a,o], Fekadu Moreda [b], Zhengtao Cui [a], Brian Cosgrove [a], Naoki Mizukami [a,c], David Kitzmiller [a], Feng Ding [a,d], Seann Reed [a], Eric Anderson [a], John Schaake [a], Yu Zhang [a], Vazken Andréassian [e], Charles Perrin [e], Laurent Coron [e], Audrey Valéry [e,1], Behnaz Khakbaz [f], Soroosh Sorooshian [g], Ali Behrangi [h], Bisher Imam [i], Kuo-Lin Hsu [g], Ezio Todini [j], Gabriele Coccia [j], Cinzia Mazzetti [k], Enrique Ortiz Andres [l], Félix Francés [m], Ismael Orozco [m], Robert Hartman [n], Arthur Henkel [n], Peter Fickenscher [n], Scott Staggs [n]

[a] Office of Hydrologic Development, NOAA National Weather Service, 1325 East West Highway, Silver Spring, MD 20910, USA
[b] Water and Ecosystems Management, Research Triangle Institute – International, 3040 Cornwallis Road, Research Triangle Park, NC, USA
[c] Research Applications Laboratory, National Center for Atmospheric Research, P.O. Box 3000, Boulder, CO, USA
[d] ADNET Systems, Inc., 6720B Rockledge Drive, Suite 504, Bethesda, MD 20817, USA
[e] Irstea (former Cemagref) – UR HBAN, 1 Rue Pierre-Gilles de Gennes CS 10030, 92761 Antony Cedex, France
[f] URS Corporation, 3500 Porsche Way, Ontario, CA 91764, USA
[g] Center for Hydrometeorology and Remote Sensing (CHRS), Department of Civil and Environmental Engineering, University of California, Irvine, E-4130 Engineering Gateway, Irvine, CA 92697, USA
[h] Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr., MS 233-300, Pasadena, CA 91109, USA
[i] International Hydrological Programme (IHP), Division of Water Sciences, UNESCO, 1 Rue Miollis, 75732, Paris Cedex 15, France
[j] Department of Earth and Geo-Environmental Sciences, University of Bologna, 1 Piazza di Porta S. Donato, 40126 Bologna, Italy
[k] PROGEA s.r.l., Via Santo Stefano 6, 40125 Bologna, Italy
[l] Idrologia & Ambiente s.r.l, Riviera di Chiaia 72, 80122 Napoli, Italy
[m] Research Institute of Water Engineering and Environment, Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia, Spain
[n] Visiting Scientist Programs, University Corporation for Atmospheric Research, P.O. Box 3000, Boulder, CO 80307, USA

ARTICLE INFO

Article history:
Received 20 November 2012
Received in revised form 25 July 2013
Accepted 28 August 2013
Available online 4 September 2013
This manuscript was handled by Konstantine P. Georgakakos, Editor-in-Chief, with the assistance of Baxter E. Vieux, Associate Editor

Keywords:
Hydrologic model
Rainfall-runoff
Distributed model
Lumped model
Calibration
Simulation

SUMMARY

The Office of Hydrologic Development (OHD) of the U.S. National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) conducted the two phases of the Distributed Model Intercomparison Project (DMIP) as cost-effective studies to guide the transition to spatially distributed hydrologic modeling for operational forecasting at NWS River Forecast Centers (RFCs). Phase 2 of the Distributed Model Intercomparison Project (DMIP 2) was formulated primarily as a mechanism to help guide the U.S. 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 forced by high quality operational 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 the testing, evaluation, and improvement of distributed models.

DMIP 2 contained experiments in two regions: in the DMIP 1 Oklahoma basins, and second, in two basins in the Sierra Nevada Mountains in the western USA. This paper presents the overview and results of the DMIP 2 experiments conducted for the two Sierra Nevada basins. Simulations from five independent groups from France, Italy, Spain and the USA were analyzed. Experiments included comparison of lumped and distributed model streamflow simulations generated with uncalibrated and calibrated parameters, and simulations of snow water equivalent (SWE) at interior locations. As in other phases of DMIP, the participant simulations were evaluated against observed hourly streamflow and SWE data and compared with simulations provided by the NWS operational lumped model. A wide range of statistical measures are used to evaluate model performance on a run-period and event basis. Differences between uncalibrated and calibrated model simulations are assessed.

* Corresponding author. Tel.: +1 301 713 0640x128; fax: +1 301 713 0963.
E-mail address: michael.smith@noaa.gov (M. Smith).
[1] Present address: EDF-DTG, 21, Avenue de l'Europe, 38040 Grenoble Cedex 9, France.

0022-1694/$ - see front matter Published by Elsevier B.V.
http://dx.doi.org/10.1016/j.jhydrol.2013.08.040

Results indicate that in the two study basins, no single model performed best in all cases. In addition, no distributed model was able to consistently outperform the lumped model benchmark. However, one or more distributed models were able to outperform the lumped model benchmark in many of the analyses. Several calibrated distributed models achieved higher correlation and lower bias than the calibrated lumped benchmark in the calibration, validation, and combined periods. Evaluating a number of specific precipitation-runoff events, one calibrated distributed model was able to perform at a level equal to or better than the calibrated lumped model benchmark in terms of event-averaged peak and runoff volume error. However, three distributed models were able to provide improved peak timing compared to the lumped benchmark. Taken together, calibrated distributed models provided specific improvements over the lumped benchmark in 24% of the model-basin pairs for peak flow, 12% of the model-basin pairs for event runoff volume, and 41% of the model-basin pairs for peak timing. Model calibration improved the performance statistics of nearly all models (lumped and distributed). Analysis of several precipitation/runoff events indicates that distributed models may more accurately model the dynamics of the rain/snow line (and resulting hydrologic conditions) compared to the lumped benchmark model. Analysis of SWE simulations shows that better results were achieved at higher elevation observation sites.

Although the performance of distributed models was mixed compared to the lumped benchmark, all calibrated models performed well compared to results in the DMIP 2 Oklahoma basins in terms of run period correlation and %Bias, and event-averaged peak and runoff error. This finding is noteworthy considering that these Sierra Nevada basins have complications such as orographically-enhanced precipitation, snow accumulation and melt, rain on snow events, and highly variable topography. Looking at these findings and those from the previous DMIP experiments, it is clear that at this point in their evolution, distributed models have the potential to provide valuable information on specific flood events that could complement lumped model simulations.

Published by Elsevier B.V.

1. Introduction

1.1. Overview

The Office of Hydrologic Development (OHD) of the U.S. National Oceanic and Atmospheric Administration's (NOAA) National Weather Service (NWS) led two phases of the Distributed Model Intercomparison Project (DMIP) as cost-effective studies to guide the transition into spatially distributed hydrologic modeling for operational forecasting (Smith et al., 2012a; Smith et al., 2004) at NWS River Forecast Centers (RFCs). DMIP 1 focused on distributed and lumped model intercomparisons in basins of the southern Great Plains (Reed et al., 2004; Smith et al., 2004). DMIP 2 contained tests in two geographic regions: continued experiments in the U.S. Southern Great Plains (Smith et al., 2012a,b) and tests in two mountainous basins in the Sierra Nevada Mountains, hereafter called DMIP 2 West. Since the conclusion of DMIP 1, the NWS has used a distributed model for basin outlet forecasts (e.g., Jones et al., 2009) as well as for generating gridded flash flood guidance over large geographic domains (Schmidt et al., 2007). The purpose of this paper is to present the DMIP 2 West experiments and results.

Advances in hydrologic modeling and forecasting are needed in complex regions (e.g., Hartman, 2010; Westrick et al., 2002). Experiments are needed in the western USA and other areas where the hydrology is dominated by complexities such as snow accumulation and melt, orographically-enhanced precipitation, steep and other complex terrain features, and sparse observational networks. The need for advanced models in mountainous regions is coupled with the requirements for more data in these areas. Advanced models cannot be implemented for operational forecasting without commensurate analyses of the data requirements in mountainous regimes.

A major component of the NWS river forecast operations is the national snow model run (NSM) by the NWS National Operational Hydrologic Remote Sensing Center (NOHRSC; Rutter et al., 2008; Carroll et al., 2001). For over a decade, NOHRSC has executed the NSM in real time at an hourly, 1 km scale over the contiguous US (CONUS) to produce a large number of gridded snow-related variables.

1.2. Science questions

DMIP 2 West was originally formulated to address several major science questions (Smith et al., 2006). They are framed for the interest of the broad scientific community with a corollary for the NOAA/NWS. These science questions and issues are highly intertwined but are listed separately here for clarity.

1.2.1. Distributed vs. lumped approaches in mountainous areas

Can distributed hydrologic models provide increased streamflow simulation accuracy compared to lumped models in mountainous areas? If so, under what conditions? Are improvements constrained by forcing data quality? This was one of the dominant questions in DMIP 1 and the DMIP 2 experiments in Oklahoma. Smith et al. (2012a,b) and Reed et al. (2004) showed improvements of deterministic distributed models compared to lumped models in non-snow, generally uncomplicated basins. The specific question for the NOAA/NWS mission is: under what circumstances should NOAA/NWS use distributed hydrologic models in addition to lumped models to provide hydrologic services in mountainous areas? While many distributed models have been developed for mountainous areas (e.g., Garen and Marks, 2005; Westrick et al., 2002; Wigmosta et al., 1994), there remains a gap in our understanding of how much model complexity is warranted given data constraints, heterogeneity of physical characteristics, and modeling goals (e.g., McDonnell et al., 2007). Several major snow model intercomparison efforts have been conducted in recent years such as Phases 1 and 2 of the Snow Model Intercomparison Project (SnowMIP; Rutter et al., 2009; Etchevers et al., 2004) and the Project for Intercomparison of Land Surface Process models (PILPS; Slater et al., 2001). In addition, several comparisons of temperature index and energy budget snow models have been conducted (e.g., Debele et al., 2009; Franz et al., 2008a,b; Lei et al., 2007; Walter et al., 2005; Fierz et al., 2003; Essery et al., 1999; WMO, 1986a,b). Comprehensive studies such as the Cold Land Processes Experiment (CLPX; Liston et al., 2008) have also been performed. However, to the best of our knowledge, there have been few specific tests of lumped and distributed modeling approaches in mountainous basins with a focus on improving river simulation

and forecasting. One such study was conducted by Braun et al. (1994), who found that finer spatial modeling scales did not lead to performance gains.

1.2.2. Estimation of models inputs and model sensitivity to existing data

What are the advantages and disadvantages associated with distributed vs. lumped modeling in hydrologically complex areas using *existing* NWS operational precipitation and temperature forcing data? Current NWS RFC lumped streamflow models in mountainous areas rely on networks of surface precipitation and temperature gages to derive mean areal averages of point precipitation and temperature observations as input forcings for model calibration and real time forecasting. The density of these networks varies greatly, with most networks featuring sparse coverage at high elevations. Even for lumped hydrologic modeling, there are uncertainties in the precipitation and temperature observations used by the NWS RFCs in mountainous areas (Hartman, 2010; Carpenter and Georgakakos, 2001). Beyond network density issues, there are problems with observation times, missing data, distribution of multi-day precipitation accumulations, and other difficulties. It is not known if these data uncertainties preclude the application of distributed models, giving rise to the question: can the existing observational networks support operational distributed modeling? Nonetheless, some have attempted to apply distributed models using such existing data (e.g., Shamir and Georgakakos, 2006). The intent in DMIP 2 West was to set up and run the models using Quantitative Precipitation Estimates (QPE) derived from the relatively dense existing gage network. Follow-on experiments would use QPE fields derived from reduced networks to investigate the appropriate density for modeling.

1.2.3. Model complexity and corresponding data requirements

The NOAA/NWS corollary is: what can be improved over the current lumped model approach used in the NWS for operational river forecasting? Is there a dominant constraint that limits the performance of hydrologic simulation and forecasting in mountainous areas? If so, is the major constraint the quality and/or amount of forcing data, or is the constraint related to a knowledge gap in our understanding of the hydrologic processes in these areas? In other words, given the current level of new and emerging data sets available to drive advanced distributed models, can improvements be realized? Or, do we still not have data of sufficient quality in mountainous areas? Additionally, what data requirements can be specified for the NOAA/NWS to realize simulation and forecasting improvements in mountainous areas?

There is a considerable range in the recent literature on the subjects of model complexity and corresponding data requirements for hydrologic modeling in mountainous areas. We provide a sample here to indicate the range of issues and findings.

For hydrologic models driven solely by precipitation and temperature, there are the issues of the gage density and location required to achieve a desired simulation accuracy (e.g., Guan et al., 2010; Tsintikidis et al., 2002; Reynolds and Dennis, 1986). The gage density issue also affects merged-data precipitation estimates (e.g., satellite–radar–gage) because the gage information is very often used to adjust the other (radar or satellite) observations (e.g., Nelson et al., 2010; Boushaki et al., 2009; Guirguis and Avissar, 2008; Young et al., 2000).

As new precipitation/runoff models with increased complexity are developed, commensurate observations of the forcing variables are also required (e.g., Dong et al., 2005; Fierz et al., 2003; Waichler and Wigmosta, 2003). Numerous studies have identified the difficulties in collecting the data required to match model complexity (Rice and Bales, 2010; Valéry et al., 2009; Feng et al., 2008; Daly et al., 2008; Franz et al., 2008a; Shamir and Georgakakos, 2006; Bales et al., 2006; Garen and Marks, 2005; Walter et al., 2005; Simpson et al., 2004; Cline et al., 1998).

Addressing the data limitations in mountainous areas noted by Garen and Marks (2005) and others (at least in terms of coverage), the number of radar, model-based, and satellite-derived products is rapidly growing. Efforts are ongoing to improve the ability of weather radars to observe precipitation in mountainous areas (e.g., Kabeche et al., 2010; Gourley et al., 2009; Westrick et al., 1999). Model-based data include the North American Regional Reanalysis – (NARR, Mesinger et al., 2006), Rapid Update Cycle (RUC; Benjamin et al., 2004), and Real Time Mesoscale Analysis (RTMA; De Pondeca et al., 2011). Much work has gone into satellite estimates of precipitation in remote regions (e.g., Kuligowski et al., 2013; Behrangi et al., 2009; Kuligowski, 2002). As these data sets emerge and become more common, users are cautioned to avoid the expectation that increased data *resolution* in new data sets will translate into increased data *realism* and *accuracy* (Guentchev et al., 2010; Daly, 2006).

1.2.4. Rain-snow partitioning

Can improvements to rain-snow partitioning be made? Partitioning between rainfall and snowfall plays a major role in determining both the timing and amount of runoff generation in high altitude basins (Guan et al., 2010; White et al., 2010, 2002; Gourley et al., 2009; Lundquist et al., 2008; Kienzle, 2008; McCabe et al., 2007; Maurer and Mass, 2006; Westrick and Mass, 2001; Kim et al., 1998). The question for the NOAA/NWS is: can distributed models provide improved representation of the spatial variability of rain/snow divisions?

Traditionally, surface temperature observations have been used to determine the form of precipitation, although such data are not the most reliable indicators of surface precipitation type (Minder and Kingsmill, 2013; Minder et al., 2011; Maurer and Mass, 2006). Recently, as part of the western Hydrometeorologic Testbed (HMT-West; Zamora et al., 2011; Ralph et al., 2005; hmt.noaa.gov), instrumentation such as vertically pointing wind profilers and S-Band radars have been used to detect freezing levels by locating the bright-band height (BBH, Minder and Kingsmill, 2013; White et al., 2010, 2002).

1.2.5. Scale issues

What are the dominant hydrologic scales (if any) in mountainous area hydrology? Understanding the variations of snowpacks and the timing and volume of snowmelt that generate streamflow has grown in recent periods but is complicated by difficult scale issues (e.g., Simpson et al., 2004). Blöschl (1999) describes three scales related to snow: process, measurement (observational data), and modeling scale. Process scale is the variability of a snow related variable. Measurement scale covers spacing, extent, and 'support' or area of integration related to an instrument. Modeling scale describes the spatial unit to which the model equations are applied (e.g., grid cell size in a distributed model). Several studies have investigated the impacts of modeling scale (e.g., Merz et al., 2009; Leydecker et al., 2001; Cline et al., 1998). However, to the best of our knowledge, there is scant literature on modeling scales that jointly considers snow and runoff processes. One exception is the work of Dornes et al. (2008), who found a spatially distributed approach provided better late season ablation rates and runoff hydrographs than a spatially aggregated model.

For forecasting agencies like NOAA/NWS, the scale question can be restated as: is there an appropriate *operational* modeling scale in mountainous areas that captures the essential rain/snow/runoff processes and provides adequate information for forecasting, water resources management, and decision support? For example, can the 4 km grid scale used in the non-mountainous DMIP 1 and 2 test basins be used instead of the current elevation zones for

operational forecasting? Or is this 4 km scale too coarse to capture the large terrain variations and resultant hydrometeorological impacts on modeling? Shamir and Georgakakos (2006) used a 1 km grid modeling scale in the American basin, but concluded that significant improvement in simulation quality would result by better representations of the spatial variability of precipitation and temperature, especially at the lower elevations of the snowpack. Some have commented on the difficulty, or even impossibility, of finding an optimum element size that effectively comprises measurement, process, and modeling scales (Dornes et al., 2008; Blöschl, 1999). DMIP 2 West intended to examine simulation performance vs. modeling scale to infer appropriate model spatial resolution.

1.2.6. Internal consistency of distributed models

Another question posed in DMIP 2 West: 'Can distributed models reproduce processes at interior locations (points upstream of basin outlet gages) in mountainous areas?' Inherent in this question is the ability of distributed models to simulate (and therefore hopefully *forecast*) hydrologic variables such as SWE, soil moisture, and streamflow at points other than those for which observed streamflow data exist. Successful simulation of such variables at interior points supports the idea that the models achieve the right answer (i.e., basin outlet streamflow simulations) for the right reason, i.e., because they are correctly modeling processes in the basin interior (Kirchner, 2006).

The remainder of this paper is organized as follows. Section 2 presents the methodology of the DMIP 2 West experiments, including data derivation and an overview of the modeling instructions. The results and discussion are presented in Section 3. We present our conclusions in Section 4, while recommendations for future work are offered in Section 5.

2. Methodology

2.1. Participating institutions, models, and submissions

Five groups submitted simulations for analysis. As with other phases of DMIP, the level of participation varied. Some participants submitted all requested simulations, while others submitted only a subset. Table 1 lists the simulations submitted by each DMIP 2 West participant.

The models used in DMIP 2 West feature a range of model structures and approaches to hydrological modeling. Appendix A presents information on the models used in DMIP 2 West, as well as references that describe the models in more detail. The greatest differences amongst the models seem to be in the precipitation-runoff approaches. The snow models all used relationships based on temperature rather than the full energy budget equations. Threshold temperatures are used to partition rain and snow. It should be kept in mind that the results herein reflect the appropriateness of model structure and other factors such as user expertise, parameter estimation, and calibration. It was not the intent of DMIP 2 West to diagnose simulation improvements from specific model structures but rather to examine the performance of the models as applied by the participants.

Focusing on precipitation-runoff models, the University of Bologna (UOB) used the TOPKAPI model (Coccia et al., 2009), which is based on the idea of combining kinematic routing with a topographic representation of the basin. Three non-linear reservoirs are used to generate subsurface, overland, and channel flow. TOPKAPI includes components that represent infiltration, percolation, evapotranspiration, and snowmelt. The NWS Office of Hydrologic Development (OHD) used the HL-RDHM model (Koren et al., 2006; 2004). HL-RDHM uses the Sacramento Soil Moisture Accounting model (SAC-SMA; Burnash, 1995) applied to grid cells. Kinematic wave equations are used to route runoff over hillslopes and through the channel system. The University of California at Irvine (UCI) also used the SAC-SMA model but applied it to sub-basins. Kinematic wave routing was used for channel routing of the runoff volumes. The Technical University of Valencia (UPV) used the TETIS model (Vélez et al., 2009; Francés et al., 2007). TETIS is a 6-layer conceptual model linked to a kinematic channel routing module. The GR4J model (Perrin et al., 2003) was used by CEMAGREF (CEM). GR4J is a parsimonious 4-parameter lumped model. For DMIP 2 West, it was applied to 5 elevation zones.

2.2. Benchmarks and performance evaluation

Two benchmarks (e.g., Seibert, 2001; Perrin et al., 2006) were used to assess model performance. Observed hourly streamflow data from the U.S. Geological Survey (USGS) were used as 'truth.' Simulations from the NWS operational precipitation/runoff model (hereafter referred to as the lumped (LMP) benchmark) were used as the second benchmark. This model was selected to address the science question regarding the improvement of distributed models compared to lumped models. In addition, the LMP model was chosen for consistency with the DMIP 1 and DMIP 2 Oklahoma experiments (Smith et al., 2012b; Reed et al., 2004).

The LMP model actually consists of several NWS components linked together. The NWS Snow-17 model (Anderson, 2006, 1976) is used to model snow accumulation and melt. Rain and melt water from the Snow-17 model is input to the SAC-SMA model. Runoff volumes are transformed into discharge using unit hydrographs. This model combination is typically applied over two elevations zones above and below the typical rain/snow elevation. Flow from the upper elevation zone is routed through the lower zone with a lag/k method. Unit hydrographs are used to convert runoff to discharge in upper and lower zones. Mean areal precipitation (MAP) and temperature (MAT) time series for the elevation zones were defined from the gridded values on the DMIP 2 ftp site. More information on elevation zone modeling can be found in Anderson (2002) and Smith et al. (2003). In the North Fork American River basin, an elevation of 1,524 m was used to divide the

Table 1
Participants' submissions. An 'x' indicates that a simulation was submitted for analysis.

Participant and DMIP 2 Acronym	Modeling Instruction 1 North Fork American River		Modeling Instruction 2 Outlet at Gardnerville, NV	Modeling Instruction 2 Interior flow point at Markleeville, CA	Modeling Instruction 3 Markleeville, CA	
	Flow	SWE (2 sites)	Flow	Flow	Flow	SWE (4 sites)
NWS/OHD Distributed (OHD) USA	x	x	x	x	x	x
NWS/OHD Lumped (LMP) USA	x	x	x	x	x	x
U. Bologna (UOB) Italy	x	x	x	x	x	x
CEMAGREF (CEM) France	x	x	x	x	x	x
U. Ca at Irvine (UCI) USA	–	–	–	–	x	x
Technical U. Valencia (UPV) Spain	x	x	x	x	x	x

upper and lower elevation zones for the LMP model. The upper basin comprises 37% of the total basin area, while the lower basin comprises 65%. In the East Fork of the Carson River we used an elevation value of 2134 m to separate the upper and lower zones. These elevation values agree with the configuration of the NWS California-Nevada River Forecast Center (CNRFC) operational models.

The LMP model also included the ability to compute the elevation of the rain/snow line using the MAT data and typical lapse rates. This rain/snow elevation is then used in conjunction with the area-elevation curve for the basin in Snow-17 to determine how much of the basin receives rain vs. snow (Anderson, 2006).

A major consideration in any model evaluation or intercomparison is the question: what constitutes a "good" simulation or an acceptable level of simulation accuracy (Seibert, 2001). This is the subject of much discussion (e.g., Bennett et al., 2013; Ritter and Muñoz-Carpena, 2013; Puspalatha et al., 2012; Ewen, 2011; Confalonieri et al., 2010; Andréassian et al., 2009; Clarke, 2008; Gupta et al., 2008; Moriasi et al., 2007; Schaefli and Gupta, 2007; Shamir and Georgakakos, 2006; Krause et al., 2005; Seibert, 2001). These references and others indicate that there is not yet an agreed-upon set of goodness-of-fit indicators for hydrologic model evaluation. Moreover, it has been difficult to specify ranges of values of the goodness-of-fit indicators that determine whether a model simulation is acceptable, good, or very good, although suggested ranges have recently emerged (Ritter and Muñoz-Carpena, 2013; Moriasi et al., 2007; Smith et al., 2003). One potential cause of these difficulties is that one must consider the quality of the input data when judging simulation results. What is a "poor" simulation for a basin with excellent input data may be considered good for a basin having poor quality input data (Ritter and Muñoz-Carpena, 2013; Moriasi et al., 2007; Seibert, 2001). As a result, the interpretation of goodness-of-fit indices continues to be a subjective process.

With this in mind, and consistent with DMIP 1 and 2 (Reed et al., 2004; Smith et al., 2012b) we use a number of performance criteria computed over different time periods to evaluate the simulations compared to the benchmarks. These include measures of hydrograph shape (modified correlation coefficient r_{mod}; McCuen and Snyder, 1975) and volume (%Bias), water balance partitioning, cumulative runoff error, and specific indices to measure the improvement compared to the LMP benchmark. Appendix D presents the statistical measures used herein and target ranges of these measures are given in the discussion below. In addition, we relate our results to those achieved in the relatively simple non-snow DMIP Oklahoma basins (Smith et al., 2012b; Reed et al., 2004).

2.3. Definitions

For consistency with the results of the DMIP 1 and 2 experiments in Oklahoma, we adopt the definition of Reed et al. (2004) 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. Interested readers are referred to Kampf and Burges (2007) for a detailed discussion of definitions and classifications regarding distributed hydrologic models.

A parent basin is defined as a watershed for which explicit calibration can be 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 RFCs. Interior points are locations within the parent basins where simulations are generated without explicit calibration (hereafter also referred to as 'blind' simulations).

Statistics are computed over two types of time intervals. The term "overall" refers to multi-year run periods such as the calibration, validation, and combined calibration/validation periods. Event statistics are computed for specific precipitation/runoff events.

2.4. Study basins

2.4.1. Description

Two sub-basins in the American and Carson River watersheds located near the border of California (CA) and Nevada (NV), USA, were selected as test basins (Fig. 1). Although these basins are geographically close, their hydrologic regimes are quite different due to their mean elevation and location on either side of the Sierra Nevada divide (Simpson et al., 2004). The Carson River basin is a high-altitude basin with a snow dominated regime, while the American River drains an area that is lower in elevation with precipitation falling as rain and mixed snow and rain (Jeton et al., 1996). These two basins were selected to represent the general hydrologic regimes of western mountainous areas with the hope that our modeling results would be relevant to other mountainous areas. Table 2 presents a summary of the characteristics of the American and Carson River basins.

In the American River basin, we selected the North Fork sub-basin above the USGS gage 11427000 shown in Fig. 2. This gage is at the North Fork dam forming Lake Clementine. Hereafter, this basin is referred to as NFDC1, using the NWS CNRFC basin acronym. This

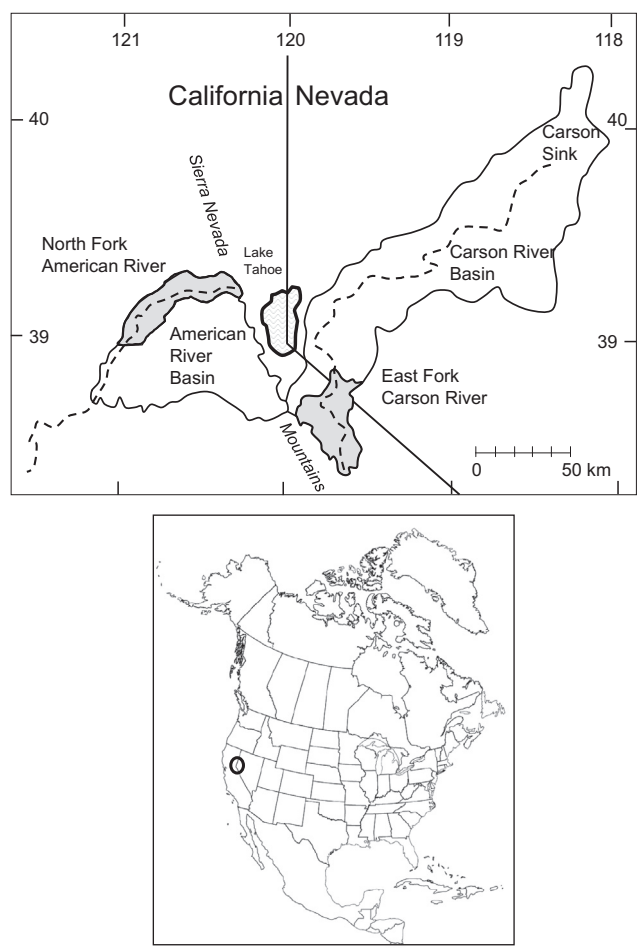

Fig. 1. Location of the American and Carson River Basins in California and Nevada, USA (top) and in relation to North America (circle, bottom).

Table 2
Characteristics of the two study basins.

	East Fork Carson River above Gardnerville, NV	North Fork American River above North Fork Dam
Area	922 km^2	886 km^2
Median altitude	2417 m	1270 m
Annual rainfall	560–1244 mm	813–1651 mm
Min, max temperature	0, 14 (°C)	3, 18 (°C)
Precipitation form	Mostly snow	Snow and rain
Aspect	Leeward	Windward
Soil	Shallow sandy and clay soil	Clay loams and coarse sandy loans
Geology	Volcanic rock and granodiorite	Meta sedimentary rock and granodiorite
Vegetation	Rangeland in lower altitude and conifer forests in upper altitude	Pine–oak woodlands, shrub rangeland, ponderosa pine forest, and subalpine forest
USGS gauge	1030900 near Gardnerville, NV	11427000 at North Fork Dam

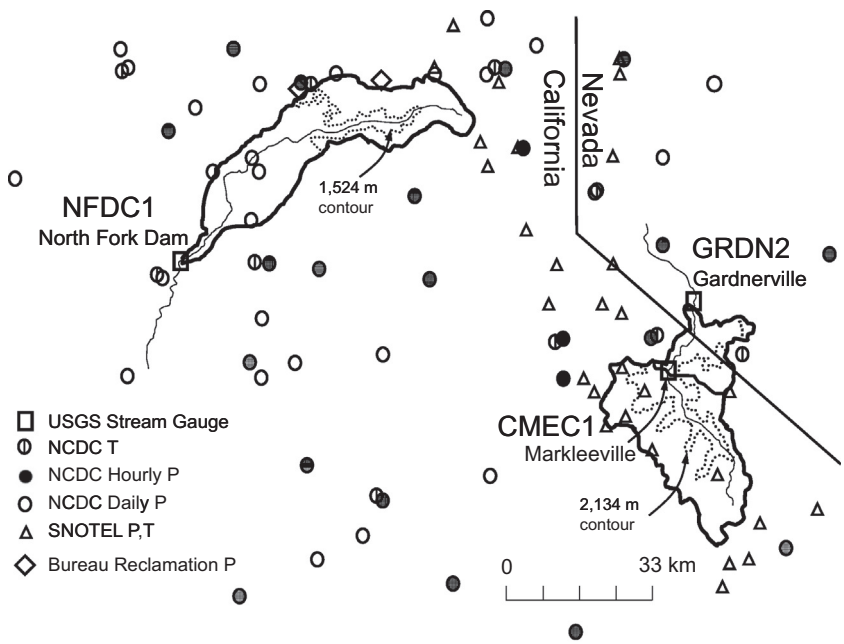

Fig. 2. Location of precipitation (*P*), temperature (*T*), and streamflow gaging stations. Also shown are the 1524 m and 2134 m elevation contours which separate the upper and lower elevation zones.

basin is 886 km^2 in area and rests on the western, windward side of the Sierra Nevada crest. Precipitation is dominated by orographic effects, with mean annual precipitation varying from 813 mm at Auburn (elev. 393 m above msl) to 1651 mm at Blue Canyon (elev. 1676 m above msl) (Jeton et al., 1996). Precipitation occurs as a mixture of rain events and rain–snow events. The basin mean annual precipitation is 1532 mm and the annual runoff is 851 mm (Lettenmaier and Gan, 1990). Streamflow is about two-thirds wintertime rainfall and snowmelt runoff and less than one-third springtime snowmelt runoff (Dettinger et al., 2004). The basin is highly forested and varies from pine-oak woodlands, to shrub rangeland, to ponderosa pine, and finally to sub-alpine forest as one moves up in elevation. Much of the forested area is secondary-growth due to the extensive timber harvesting conducted to support the mining industry in the late 1800s (Jeton et al., 1996). Soils in the basin are predominately clay loams and coarse sandy loams. The geology of the basin includes metasedimentary rocks and granodiorite (Jeton et al., 1996).

In the Carson River basin, the East Fork sub-basin shown in Fig. 2 was selected for DMIP 2 West. Hereafter, the CNRFC identifier GRDN2 is used for the basin above the gage at Gardnerville, NV, and CMEC1 is used for the interior basin above the stream gage at Markleeville, CA. The Carson River terminates in the Carson Sink. The East Fork Carson River generally flows from south to north so that its average slope is not as steep as it could be if it were to face directly east–west. GRDN2 is a high altitude basin, with a drainage area of 714 km^2 above USGS stream gage 10-308200 near Markleeville, CA and 922 km^2 above USGS stream gage 10–309000 at Gardnerville, NV. Elevations in the GRDN2 basin range from 1650 m near Markleeville to about 3400 m at the basin divide. Mean annual precipitation varies from 559 mm at Woodfords (elev. 1722 m) to 1,244 mm near Twin Lakes (elev. 2438 m). Fig. 3 shows the rugged, heavily forested terrain in both basins.

2.4.2. Rationale for basin selection

Several factors underscore the selection of the NFDC1 and GRDN2 basins for use in DMIP 2 West. First, both basins are largely unregulated (Georgakakos et al., 2010; Simpson et al., 2004; Dettinger et al., 2004; Carpenter and Georgakakos, 2001; Jeton et al., 1996). Second, these basins are geographically close, yet they present an opportunity to study different hydrologic regimes. Moreover, their proximity allowed for more expedient data processing by DMIP 2 West organizers and participants.

Fig. 3. Terrain and landcover in the study basins. North Fork of the American River looking downstream from near the Iowa Hill Bridge (upper left). North Fork of the American River looking southeast from the airport at Blue Canyon (upper right; photo courtesy of Brooks Martner). Forested conditions at the Blue Lakes SNOTEL site in the East Fork Carson River basin (bottom).

Third, at the time of the design of DMIP 2 West, these basins had already been studied by numerous researchers. For example, these basins have been used for parameter estimation experiments (Leavesley et al., 2003; Jeton and Smith, 1993), climate change studies (Lettenmaier and Gan, 1990; Jeton et al., 1996; Carpenter and Georgakakos, 2001), and mountainous area hydrometeorology investigations (Wayand et al., 2013; Simpson et al., 2004; Lundquist and Cayan, 2002; Cayan and Riddle, 1993; Kim et al., 1998). Several authors used these basins to study the dynamics of the precipitation generation process in mountainous areas (Wang and Georgakakos, 2005; Dettinger et al., 2004; Hay and Clark, 2003; Tsintikidis et al., 2002; Pandey et al., 1999, 2000; Reynolds and Dennis, 1986). Data sets from the HMT-West experiments in NFDC1 have been used to partition rain and snow (e.g., Minder and Kingsmill, 2013; Mizukami et al., 2013; White et al., 2010; Lundquist et al., 2008), improve radar estimates of QPE in the mountains (e.g., Gourley et al., 2009; Matrosov et al., 2007), identify land-falling atmospheric rivers (e.g., Ralph and Dettinger, 2011; Kingsmill et al., 2006), enhance physics in numerical weather models for improved QPF (e.g., Jankov et al., 2007), and understand soil moisture movement in mountainous soils (Zamora et al., 2011).

2.5. Calibration

Participants were free to calibrate their models using strategies and statistical measures of their choice as this process is usually model-dependent. This provision was also an aspect of the Oklahoma experiments in DMIP 1 and 2 (Smith et al., 2012a,b; Reed et al., 2004) and is similar to the Model Parameter Estimation Experiment (MOPEX; Duan et al., 2006) and SnowMIP-2 (Rutter et al., 2009). Appendix B presents a brief description of the strategies followed by the participants to calibrate their models. Three models (OHD, UOB, and UPV) used spatially variable *a priori* parameters and adjusted the parameter grids uniformly using scalar factors. CEM and UCI maintained the spatially constant parameters in each computational area.

2.6. Run periods

Specific periods were prescribed for model calibration and validation. An initial one-year 'warm up' 'or 'spin-up' period was provided to allow models to equilibrate after a complete annual wetting/drying cycle. Table 3 presents the computational periods. The warm-up and calibration periods for the East Fork of the

Table 3
Time spans of the 'warm up', calibration, and validation periods.

Basin	'Warm Up'	Calibration	Validation
North Fork American R. at L. Clementine (NFDC1)	October 1, 1987 to September 30, 1988	October 1, 1988 to September 30, 1997	October 1, 1997 to September 30, 2006
East Fork Carson R. at Markleeville, CA (CMEC1)	October 1, 1989 to September 30, 1990	October 1, 1990 to September 30, 1997	October 1, 1997 to September 30, 2006
East Fork Carson R. at Gardnerville, NV (GRDN2)	October 1, 1989 to September 30, 1990	October 1, 1990 to September 30, 1997	October 1, 1997 to September 30, 2006

Carson and North Fork of the American were slightly different due to the availability of the observed hourly USGS streamflow data.

2.7. Simulations and modeling instructions

Participants followed specific modeling instructions to generate simulations of streamflow and SWE in order to address the science questions (See http://www.nws.noaa.gov/oh/hrl/dmip/2/docs/sn_modeling_instructions.pdf and Table 1). Modeling Instruction 1 was for NFDC1. Participants generated hourly simulated streamflow at the basin outlet gage using calibrated and uncalibrated model parameters. There were no interior "blind" streamflow gages in NFDC1. During the same run to generate the outlet streamflow hydrographs, participants also generated simulations of snow water equivalent at two locations where snow pillows are operated by the U.S. Bureau of Reclamation (USBR): Blue Canyon and Huysink. Modeling Instruction 2 focused on the GRDN2 basin. Participants generated uncalibrated and calibrated streamflow simulations at the GRDN2 outlet. During the same run to generate GRDN2 outlet simulations, participants generated streamflow simulations at the interior USGS gage at Markleeville, CA. The gage at Markleeville was considered a blind simulation point for this test with no explicit calibration. Hereafter, we refer to this test as CMEC1-2. In addition, the basin above the USGS gage at Markleeville, CA was also treated as an independent basin (Modeling Instruction 3). This test is referred to as CMEC1-3. As part of Modeling Instruction 3, participants also generated SWE simulations for four U.S. Natural Resources Conservation Service (NRCS) snowpack telemetry (SNOTEL) sites (Serreze et al., 1999). Explicit instructions for modeling scale were not provided. Rather, it was hoped that the participants' models would inherently represent a sufficiently broad range of modeling scales from which to make inferences on appropriate model scale.

2.8. Data assimilation

Assimilation of observed streamflow or other data to adjust model states was not allowed in DMIP 2 West. The simulations were generated by running models continuously over the warm-up, calibration, and validation periods.

2.9. Data

Participants were required to use the hourly precipitation and temperature grids posted on the DMIP 2 West web site. This gridded data was derived from the same types of in situ measurements used by RFCs to construct MAP and MAT time series for model calibration and operational forecasting. As in the other phases of DMIP, basic forms of many other physical basin features and meteorological variables were provided to promote participation.

2.9.1. Precipitation

2.9.1.1. Precipitation data sources. DMIP 2 West used precipitation data collected by the NWS Cooperative Observer Network (COOP; NRC, 1998) available from the National Climatic Data Center (NCDC). Also used were daily precipitation observations from the SNOTEL network. Problems with the COOP and SNOTEL data are well known and include difficulties in distributing multi-day accumulations (Eischeid et al., 2000), data-entry, receiving, and reformatting errors (Reek et al., 1992), observer errors (e.g., NRC, 1998), and dealing with varying daily station observation times (e.g., Hay et al., 1998).

Hourly and daily stations from the COOP and SNOTEL networks were selected inside and near the basins. After initial screening for period-of-record (at least 5–10 years) and windward-leeward effects, 41 stations for NFDC1 and 69 for GRDN2 were selected for further analysis. The location of the precipitation and temperature stations is shown in Fig. 2.

2.9.1.2. Generation of gridded QPE. Considerable effort was expended to generate a multi-year, hourly, 4 km gridded QPE data set (Smith et al., 2010). One goal in the development of the QPE data was to generate spatially varying gridded precipitation data based on the same types of in situ point measurements currently used by NWS RFCs for lumped model calibration and real time operational forecasting. In this way we could address the science question: can distributed models be operationally implemented with currently available precipitation and temperature data?

An initial QPE data set for 1987–2002 was used to launch the DMIP 2 West experiments (Moreda et al., 2006), but was found to contain a large inconsistency when extended from 2002 to 2006 (Mizukami and Smith, 2012; Smith et al., 2009). Developing an alternative procedure and QPE data set delayed DMIP 2 West experiments by nearly two years. Finally, an approach was developed using modified NWS procedures as shown in Fig. 4. The method consists of three major steps: (1) data quality control and generation of hourly point precipitation time series, (2) spatial interpola-

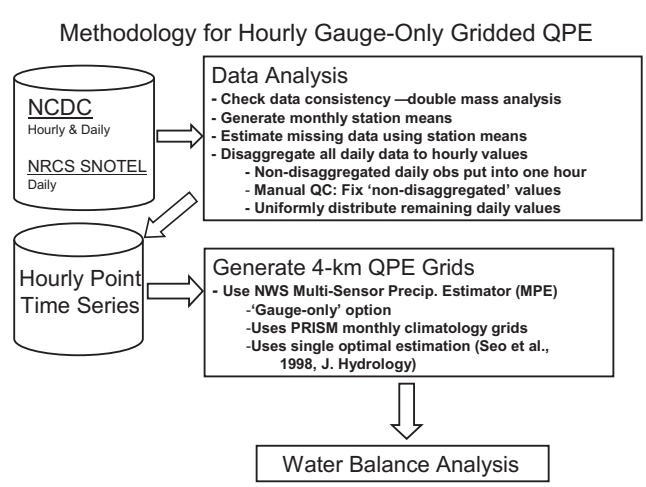

Fig. 4. Methodology for generating gage-only gridded QPE for DMIP 2 West.

tion of the point time series to a 4 km grid, and (3) water balance analyses. Details of these steps can be found on the DMIP 2 West web site: http://www.nws.noaa.gov/oh/hrl/dmip/2/wb_precip.html.

Step 1: Data quality control and generation of point hourly data time series.

The goal of this step was to generate quality-controlled, serially complete, hourly precipitation time series at all the gage locations, including daily stations. The term serially complete means that precipitation was estimated for each station over the entire analysis period of 1987–2006 (no missing values; Eischeid et al., 2000). The hourly and daily data were quality controlled (QC) using standard NWS procedures described in Smith et al. (2003) and Anderson (2002). Double mass analysis was used to identify and correct for human influences. Missing precipitation data were estimated using weighted observations from the closest station in each of four quadrants. To account for orographic influences, long-term ratios of monthly station means were used to condition estimates of missing data by the ratio of long term monthly mean of the estimator station to the current station.

Multi-day gage accumulations were distributed over the preceding flagged days using daily amounts from surrounding stations. Daily precipitation accumulations were subsequently distributed to hourly values using the distributions of hourly stations surrounding each daily station. A great deal of effort was expended to correct problems associated with distributing daily observations (Smith et al., 2010). In particular, two types of problems were addressed: (1) distributing multi-day accumulations to each individual day (e.g., Eischeid et al., 2000) and (2) distributing daily totals into hourly values. The precipitation records from 1988 to 2006 were examined month by month to identify and correct such errors.

Step 2. Generation of gridded QPE.

A major task is the estimation of precipitation at ungaged gridded points (Tsintikidis et al., 2002). The multi-sensor precipitation estimation (MPE: Seo, 1998) algorithm was used to spatially distribute the point hourly time series onto the DMIP 2 West ~ 4 km Hydrologic Rainfall Analysis Project (HRAP; Green and Hudlow, 1982; Reed and Maidment, 1999) grid. Data selected from 27 stations in and around NFDC1 and 62 stations for GRDN2 were used to derive the final gridded QPE data sets.

MPE uses PRISM data to adjust the interpolation of point precipitation to grids. The 800 m resolution PRISM monthly climatological precipitation data derived for 1971–2000 (Daly et al., 2008) was selected to be consistent with the river forecasting operations at CNRFC (Hartman, 2010). The DMIP 2 West QPE data did not include any corrections for gage under-catch (Yang et al., 1998a,b). Treatment of gage under-catch is usually model-dependent so participants were free to make adjustments as they chose.

Step 3. Water balance analysis.

A check of the DMIP 2 West precipitation is shown in Fig. 5. This figure presents a Budyko-type plot (Budyko, 1974) of water balance components for a number of basins across the US, representing a broad range of basin climatologies. On the abscissa is plotted the ratio of observed long-term mean precipitation (P_{obs}) to potential evapotranspiration (PET) while on the ordinate we plot the ratio of observed streamflow (Q) to PET. PET was computed from the NOAA Evaporation Atlas (Farnsworth et al., 1982). The basins in the domain of the NWS Arkansas-Red Basin RFC (ABRFC) were taken from the work of Koren et al. (2006) and range in size from 20 km^2 to 15,000 km^2. This figure shows that the climatological

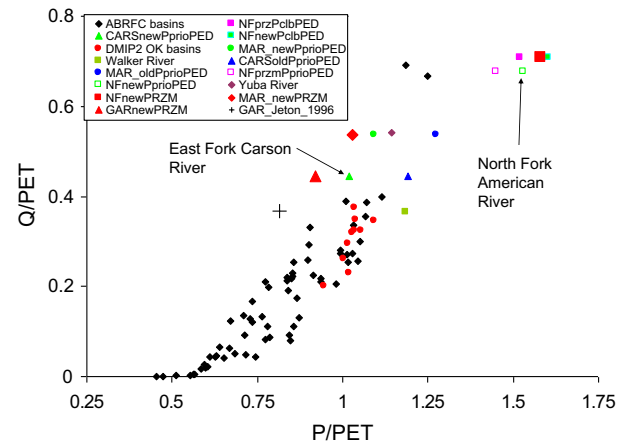

Fig. 5. Long term mean annual values of Q/PET and P/PET for the western basins. Also shown are basins in several climatic regions of the USA. The black diamonds are basins in the U.S. Southern Great Plains, of which the red circles are a subset used in DMIP 1 and 2 Oklahoma experiments.

variables of PET and precipitation for the DMIP 2 West basins agree with the trend established by the other basins, indicating that the mean annual precipitation values are reasonable.

2.9.2. Temperature

2.9.2.1. Temperature data sources. Hourly 4 km gridded temperature values were derived using daily maximum and minimum (hereafter tmax and tmin) temperature data available from the NWS COOP and SNOTEL networks shown in Fig. 2. The underlying interpolation procedure uses an inverse distance weighting algorithm. It also uses PRISM gridded monthly climatological analyses of daily maximum and minimum temperature (Daly et al., 1994). The procedure has the following major steps:

2.9.2.2. Generation of gridded hourly temperature.

Step 1. Generate gridded daily tmax and tmin.

Observation times estimated for COOP stations with missing temperature observation times were assumed to be the same as for the corresponding daily precipitation observations. The procedure used to estimate missing observation times for these stations is documented in Schaake et al. (2006).

A daily temperature processor was used to generate daily tmax and tmin grids for each day of the analysis period. Complex terrain in the DMIP 2 West study area generates spatial variations in temperature that are comparable to diurnal temperature variations. As a result, simple spatial interpolation of gage observations to grid locations did not by itself account for the complexity of the actual spatial variations. Therefore, monthly PRISM climatological grids of daily mean tmax and tmin were used as part of the interpolation process.

The spatial interpolation procedure for daily maximum and minimum temperature analysis is as follows. COOP and SNOTEL sites to be used for the given day were selected having at least 5 years of observations with less than 15% missing data. At each site, the algorithm computes the difference between the observed gage value for the given day and the monthly PRISM climatological mean values of tmax and tmin. These differences were interpolated to the HRAP grid used by DMIP 2 West using an inverse distance weighting interpolation procedure with the distance exponent equal to 1.0. Difference values for each of the nearest 2 gages in each of 4 quadrants surrounding each HRAP grid point are used. For each HRAP grid point, the PRISM mean value was added to

the analyzed difference value to get the grid point value of the analyzed maximum or minimum daily air temperature.

Step 2. Generate hourly temperature grids.

The hourly temperature processor uses the daily tmax and tmin grids to generate hourly temperature grids for each hour of each day. This procedure uses the Parton and Logan (1981) algorithm to estimate hourly temperatures from daily max–min values. As a check of the procedure, gridded climatologies of daily tmax and tmin for January for the period 1961–1990, over an area including the American and Carson River basins were generated (not shown). The gage analysis and PRISM climatologies were nearly identical.

2.9.3. Form of precipitation

Participants were free to determine the form of precipitation (rain or snow) for each time step in their modeling. The procedures followed by the participants are shown in Appendix A.

2.9.4. Potential evaporation

Participants were allowed to determine the values of potential evaporation data for their models. As in the DMIP 2 experiments in Oklahoma, estimates of climatological monthly PE for both basins were provided. Koren et al. (1998) used information from seasonal and annual Free Water Surface (FWS) evaporation maps in NOAA Technical Report 33 (Farnsworth et al., 1982) and mean monthly station data from NOAA Technical Report 34 (Farnsworth et al., 1982) to derive parameters for an equation that predicts the seasonal variability of mean daily free water surface (FWS) evaporation. These parameters were used to derive the mean monthly FWS evaporation estimates for DMIP 2 West basins. The data for NFDC1 are in the same range as the estimates derived by Carpenter and Georgakakos (2001) for the entire American River basin.

A link to the NARR data (Mesinger et al., 2006) was also provided, along with guidelines and processing codes for participants to compute other estimates of PE. The NARR project provides relative humidity, wind speed, air temperature, and radiative flux data.

2.9.5. Analysis of simulated hydrographs

As a final check of the DMIP 2 West forcings, the precipitation, temperature, and climatological PE forcing data were tested in hourly lumped and distributed simulations over the project period. Snow correction factors used in Snow-17 for these simulations to compensate for gage undercatch were derived during the calibration of the benchmark LMP model. All suspect simulated hydrograph 'spikes' that did not seem to be consistent with the observed data were investigated (Smith et al., 2010). Cumulative streamflow simulation error plots were also examined.

2.9.6. Digital elevation data

Participants were not required to use any particular digital elevation model (DEM). 15 arc-s and 1 arc-s DEM data were provided via the DMIP 2 West website to encourage participation. The 15 arc-s national DEM was derived by resampling 3 arc-s DEMs (1:250,000 scale) distributed by the U.S. Geological Survey. 1 arc-s DEM data were available from the USGS National Elevation Dataset (NED).

2.9.7. Flow direction data

Flow direction grid files at a 30 m resolution were provided for the convenience of any participants who wished to use them. These 30 m grids were used to define the basin boundaries. The basis for these flow direction grids was the 30 m DEM data from the USGS NED data server. The DEM data were projected and filled, and commercial software was used to calculate flow directions using the D8 algorithms of Jenson and Domingue (1988). Flow directions were also provided at a 400 m resolution. Moreover, we also defined flow directions for coarse resolution model cells that matched or aligned with the grid of available radar-based forcing data (the HRAP grid) using the algorithm described by Reed (2003).

2.9.8. Vegetation and land use data

DMIP 2 West provided a 1 km gridded vegetation/land use dataset covering both basins. These data were originally developed by Hansen et al. (2000). Thirteen classes of vegetation were defined in these data.

2.9.9. Hourly observed streamflow data

Instantaneous hourly flow data were acquired from the USGS. The data included some corrections and shifts but were defined as provisional. Unlike the mean daily flow data available from the USGS National Water Information Service (NWIS) web site, the instantaneous observations had not undergone rigorous quality control. However, OHD performed some rudimentary quality control steps. These involved: (1) downloading the approved mean daily flow data from the NWIS web site for the same time periods, (2) computing mean daily flow from the hourly data, (3) visually comparing the derived and approved daily flow time series, hourly streamflow data, mean areal precipitation data, and basic simulations for each basin, and (4) setting any suspicious data in the hourly time series to missing values.

2.9.10. Soils information

State Soil Geographic (STATSGO) texture data covering the two basins were taken from data sets originally derived by Miller and White (1998). These were made available as a grid for each of 11 soil layers. In addition, a link was provided to finer resolution county-level soil information called the Soil Survey Geographic (SSURGO; Zhang et al., 2012) data set. The SSURGO data are typically available at a scale of at least 1:24,000. They are approximately ten times the resolution of STATSGO data in which the soil polygons can be on the scale of 100 – 200 km^2. SSURGO data have been recently used to derive *a priori* estimates of model parameters (e.g., Zhang et al., 2012, 2011). Participants were free to use information from either soil data set to derive any necessary model-specific parameters.

2.9.11. Areal extent of snow cover

NWS RFCs use snow covered area (SCA) data in the calibration of the operational hydrologic models. For DMIP 2 West, SCA data from the NWS NOHRSC were extracted for the basins. The data consisted of gridded 'snap shots' of SCA on available days, with values in each cell indicating clouds, snow, or no snow.

2.9.12. Snow water equivalent

Observed SWE data were also made available to DMIP 2 West participants. SWE data for two USBR sites in NFDC1 were downloaded from the California Data Exchange (CDEC). Data from four SNOTEL sites in the GRDN2 basin were also provided. The sites are listed in Appendix C. These data spanned the calibration period. Participants were allowed to use these data in the calibration of their models.

2.9.13. Cross sections

Due to the remote nature of the North Fork basin, we were only able to provide cross section data for one location in the North Fork basin. These data were derived from as-built bridge plans for the Iowa Hill Bridge near Colfax, CA provided by personnel from California State Parks.

3. Results and discussion

Following the format of the DMIP 2 Oklahoma results (Smith et al., 2012b), we present the results of the experiments from general to specific in order to address the science questions in a coherent manner. A number of statistical measures were used to assess the simulations. It would be impossible to present and discuss all of the analyses that were performed, but the results of the most important and relevant ones are presented below.

3.1. Overall water balance

A critical aspect of hydrologic modeling is the partitioning of precipitation into runoff and evapotranspiration/losses. This is especially important in mountainous areas given the large uncertainty in precipitation, temperature, and other meteorological observations. Following other experiments (e.g., Lundquist and Loheide, 2011; 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 the participants' models to partition precipitation into runoff, evaporation, and losses.

The water balance quantities for each model were computed using the general continuity equation:

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

where S is storage, P_{obs} is observed mean annual basin-average precipitation in mm, E is evaporation in mm, L represents the inter-catchment groundwater transfer (losses or gains) and R_{model} is the depth of model runoff in mm over the basin. We computed these quantities on an annual basis over a multi-year period and assumed 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:

$$E + L = P_{obs} - R_{model} \tag{2}$$

Fig. 6 can be interpreted as follows. Each diagonal 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. On each diagonal, a model's plotting symbol can be projected to the x and y axes to yield that model's basin-averaged mean annual runoff and evaporation plus losses. 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 should plot at the same point if they have the same partitioning of water. From Fig. 6, with the exception of one outlier (CEM for NFDC1), it can be seen that the models partition precipitation reasonably well. There is slightly more spread in the results for NFDC1 (156 mm) than in the GRDN2 (80 mm) and CMEC1 (105 mm) basins, which may be due to the difficulty in modeling the rain-snow dominated NFDC1.

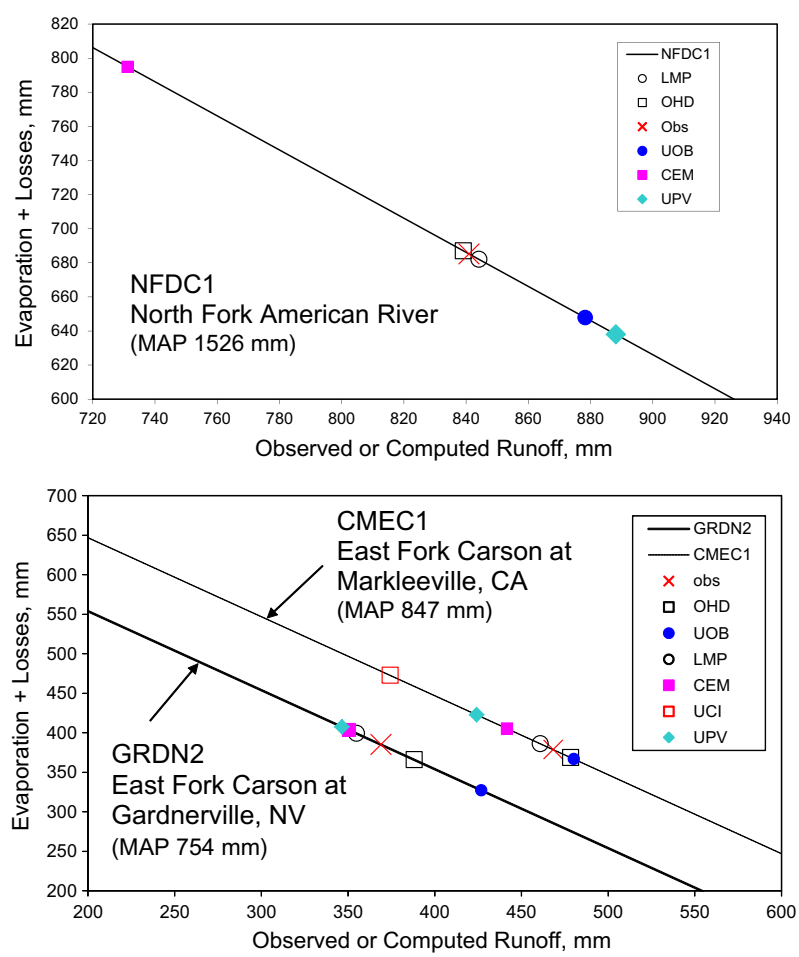

Fig. 6. Partitioning of precipitation into runoff and evaporation + losses: NFDC1 (top) and CMEC1 and GRDN2 (bottom).

3.2. Long-term cumulative simulation error

Overall model performance in terms of long-term runoff simulation error as performed by Reed et al. (2004) and Smith et al. (2012b) was also analyzed. This analysis was meant to examine the consistency of the precipitation estimates and subsequent impacts on multi-year hydrologic model simulations. Consistent precipitation data have been proven to be necessary for effective model calibration (Smith et al., 2012b and references therein). For example, in some cases in the DMIP 2 Oklahoma experiments, the improvements gained by distributed models compared to lumped models were negated when the models were calibrated using inconsistent precipitation data from DMIP 1 (Smith et al., 2012b). Fig. 7 presents the cumulative runoff error plots for the NFDC1, GRDN2, and CEMC1-3 basins. In general, the plots are linear, indicating that the precipitation, temperature, and evapotranspiration forcings are temporally consistent. The under-prediction from CEM shown in Fig. 7 follows from the partitioning shown in Fig. 6: the CEM model generated less runoff volume over time. The cumulative error plots in Fig. 7 are considerably improved compared to the OHD cumulative error plot using the original (flawed) DMIP 2 gridded QPE (Mizukami and Smith, 2012; Smith et al., 2009; Moreda et al., 2006). Moreover, the results in Fig. 7 span approximately the same error range for two basins in the DMIP 2 Oklahoma experiments (see Fig. 10 in Smith et al., 2012b).

3.3. Comparison of distributed and lumped model results

In this section we begin to address the science question: can distributed hydrologic models provide increased streamflow simulation accuracy compared to lumped models in mountainous areas? Fig. 8 presents the overall performance of the calibrated models in terms of r_{mod} (McCuen and Snyder, 1975), computed hourly for the calibration, validation, and combined periods. The term 'overall' means that the statistic was computed for each hour over the entire period specified. The r_{mod} measure was used herein to provide consistency with the DMIP 2 Oklahoma results (Smith et al., 2012b) and DMIP 1 results (Reed et al., 2004). The r_{mod} statistic is a goodness-of-fit measure of hydrograph shape. In Fig. 8, the results are organized in order of increasing computational element size (i.e., ranging from 250 m for UOB to two elevation zones for LMP). Recall that there are two sets of results for the basin CMEC1: one is for an independent calibration of the basin (CMEC1-3) and the other is for the use of the Markleeville gage as a 'blind' interior simulation point (CMEC1-2).

Looking collectively at the results in the top panel of Fig. 8, no single model performed best for all the basins in the calibration period. The models had the most uniform r_{mod} value for the snow-dominated basin GRDN2, and the most spread in r_{mod} for NFDC1. All models except LMP showed a decrease in r_{mod} for the blind simulation test CMEC1-2. Not surprisingly, all models showed an improvement in r_{mod} when the CMEC1 basin was explicitly calibrated (compare CMEC1-2 with CEMC1-3). Compared to LMP, the UOB model provided the only improved r_{mod} values for this calibration period for NFDC1, while the OHD, UOB, and CEM models provided equal-to or improved values for GRDN2. No model was able to provide an improvement over the lumped model for the blind test at CMEC1-2. Only the OHD model provided improved r_{mod} values for the CMEC1-3 test.

The r_{mod} values for the validation period are shown in the middle panel of Fig. 8. Models typically perform slightly worse in validation periods compared to the calibration period. In our case, mixed results were realized. Some models improved in this period (e.g., CEM in the CMEC1-2 and CMEC1-3 cases), while others showed the expected decline in r_{mod}. Only the OHD model provided improved r_{mod} values compared to LMP for the validation period.

The bottom panel of Fig. 8 shows the model performance over the combined calibration and validation periods. The LMP and OHD models had the highest r_{mod} for this period in all the basins. The CEM model achieved the next-highest values for three out of four basin tests. Only the OHD model provided improved r_{mod} values compared to LMP (cases of GRDN2 and CMEC1–3) in this period. Very similar results to Fig. 8 (not shown) were achieved when using the Nash–Sutcliffe statistic instead of r_{mod}.

The %Bias for all the calibrated models is shown in Fig. 9. The %Bias statistic is used to compare the overall simulated and observed runoff volumes. Target values of the %Bias measure vary in the literature from ±5% for NWS model calibration (Smith et al., 2003) to ±25% (Moriasi et al., 2007).

A wide range of values can be seen looking collectively at the plots. For the calibration period, some models were able to achieve a near zero bias for several of the basins (e.g., CEM in GRDN2, LMP in CMEC1-3, OHD in NFDC1). Other models had a consistent positive or negative bias for all the basins and all periods (UOB, CEM). As might be expected, the %Bias values were larger for the CMEC1-2 test compared to the explicitly calibrated CMEC1-3 test. Recall that CMEC1-2 simulations were generated as interior points within GRDN2 with no specific calibration (Modeling Instruction 2). In terms of %Bias, no model performed best in all cases, nor did any distributed model consistently outperform the LMP benchmark.

While no one model performed consistently better than the others in Fig. 8 and Fig. 9, all models achieved relatively high values of the two statistics. For example, the values of r_{mod} (0.63–0.95) and %Bias (−20.1 to 5.2) for the combined calibration–validation period in Figs. 8 and 9 fall well within the ranges of r_{mod} and %Bias in the non-snow DMIP 2 Oklahoma experiments (Figs. 3 and 4 of Smith et al., 2012b).

3.4. Analysis of precipitation/runoff events

To further investigate the science question of the performance of distributed and lumped models, statistics were computed for 68 events in the NFDC1 and 92 events in the GRDN2 and CMEC1 basins. These events were selected from the combined calibration and validation period. Event statistics were computed because our experience has shown that the overall run-period statistics can mask the improvement of distributed models over lumped models for individual events (Smith et al., 2012b; Reed et al., 2004). We use the same two measures as in DMIP 1 and DMIP 2 (Reed et al., 2004; Smith et al., 2012b respectively) to evaluate the model performance for events: the event absolute % runoff error and the event absolute % peak error. These measures evaluate the models' ability to simulate runoff volumes and peak flow rates. Eqs. (7a) and (7b) of Appendix D present the formula for % runoff error and % peak error, respectively. Low values of % runoff error and % peak error are desired.

The event statistics for the NFDC1 basin are shown in Fig. 10. Each plotting symbol represents the average measure of a specific model for the 68 events. As much as possible, the same plotting symbols in Smith et al. (2012b) and Reed et al. (2004) are used. The LMP and OHD models have near-identical performance with the lowest values of the two statistics. The OHD model provides a slight improvement in % peak error over the LMP model but at the expense of a slightly worse runoff volume. Next in order of performance is the UOB model, followed by the UPV and CEM models.

The event statistics for GRDN2 are plotted in Fig. 11. All models have errors whose magnitudes are similar to Fig. 10. One difference is that there is more spread between the OHD and LMP results here compared to the NFDC1 basin. In this case the OHD model provides

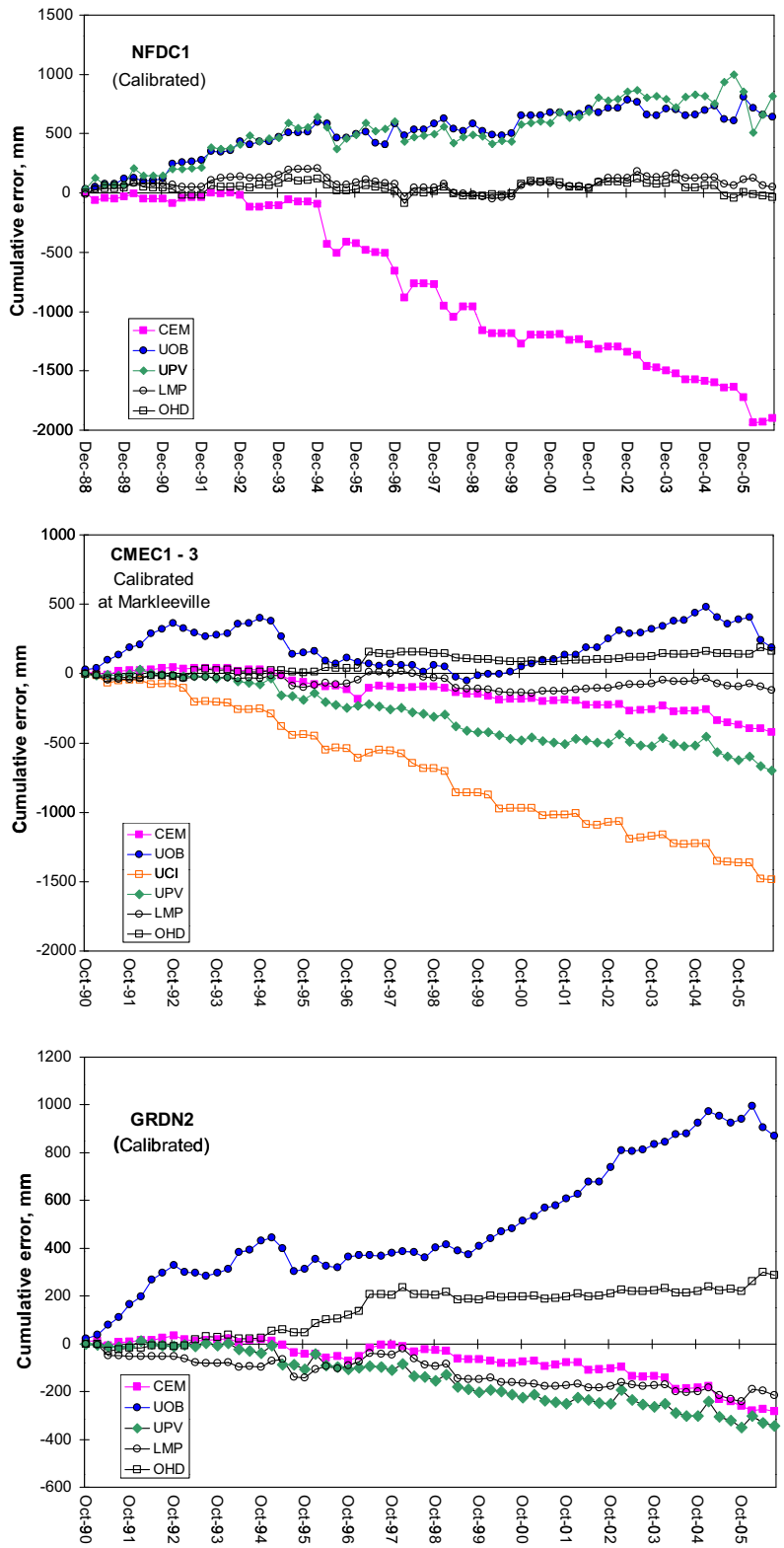

Fig. 7. Cumulative runoff error (mm) for the combined calibration and validation periods. NFDC1 (top), CMEC1-3 (middle), and GRDN2 (bottom).

improvement compared to the LMP model for both runoff volume error and peak error.

Fig. 12 shows the calibrated event results for the CMEC1-2 and CMEC1-3 tests. The former calls for the CMEC1 basin to be simulated as a blind interior point (Modeling Instruction 2) within the GRDN2 basin. The latter test calls for explicit calibration of the CMEC1 basin as an independent headwater basin (Modeling Instruction 3). The plotting symbols are shown in different sizes to distinguish the results for each participant. The arrows show how explicit calibration of the CMEC1 basin impacts the statistics.

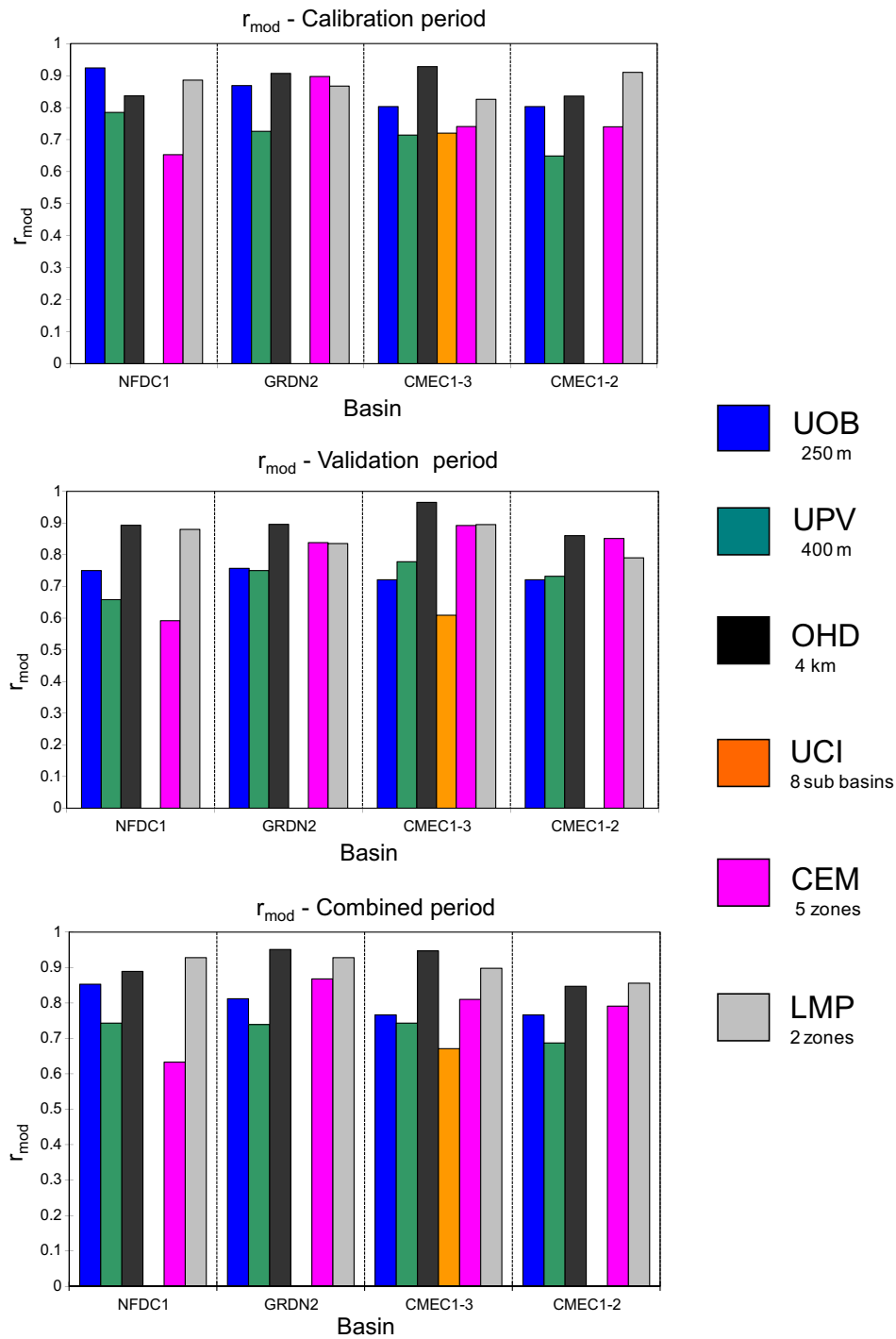

Fig. 8. Calibrated r_{mod} for three periods: calibration (top), validation, (middle), and combined periods (bottom). For each basin, models are arranged in order of increasing model scale. *Note*: UCI only submitted simulations for CEMC1-3.

Note that UCI only submitted simulations for the CMEC1-3 test, and the UOB used the same simulations for the CMEC1-2 and CMEC1-3 tests. The CMEC1-2 results (small plotting symbols) span approximately the same range of error values as in NFDC1 (Fig. 10) and GRDN2 (Fig. 11). Thus, in this case, the distributed models calibrated at the basin outlet achieved about the same event simulation performance at this interior location. Explicit calibration of the CMEC1 basin (i.e., CMEC1-3) improved the event statistics for the LMP, UPV, and OHD models as one would expect. The OHD model generated a 'blind' simulation that was slightly better than the LMP model in terms of % peak error (23% vs. 26%). These results may be influenced by the large size of the interior basin (714 km^2) compared to the parent basin (922 km^2).

Summarizing the results of Figs. 10–12, only one distributed model (OHD) was able to perform at a level near or better than the LMP benchmark for absolute % runoff error and absolute % peak error. However, all models achieved levels of performance equivalent to those in one of the non-snow Oklahoma basins in DMIP 2.

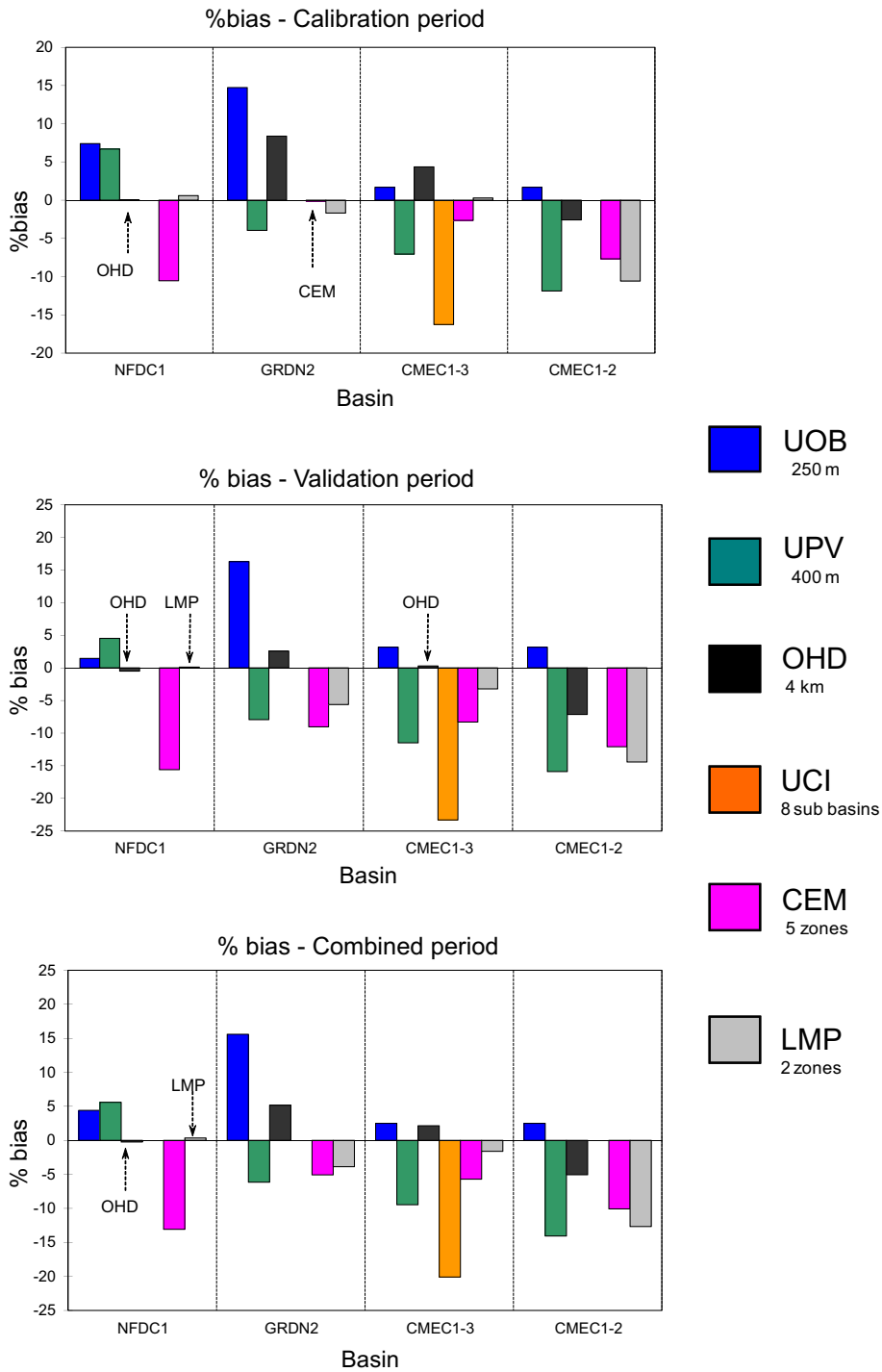

Fig. 9. Calibrated %Bias for three periods: calibration (top), validation, (middle), and combined periods (bottom). *Note*: UCI only submitted simulations for CEMC1-3.

For example, the range of values of absolute % runoff error (17.2–47.3) and absolute % peak error (15.5–28.1) agree very well with the results for the ELDO2 basin in Oklahoma (see Fig. 5 of Smith et al., 2012b).

3.5. Improvement of distributed models over lumped models

This analysis specifically addresses the question whether calibrated distributed models can provide improved event simulations compared to lumped models in mountainous regions. Using the same events from Section 3.4, three specific measures of improvement were computed: % improvement in peak flow, % improvement in runoff event volume, and improvement in peak time error in hours (Eqs. (8a)-(8c) in Appendix D). The calibrated simulations from the LMP model were used as the benchmark.

Fig. 13 presents the three measures of improvement for the calibrated models, along with the inter-model average of the values. It is desirable to achieve values greater than zero in each of the plots. Each data point is the average value of the measure for a specific model in a specific basin over many events. Looking collectively at the plots, it can be seen that the OHD model provides improvement in peak flow and volume, as seen previously in Figs. 10–12.

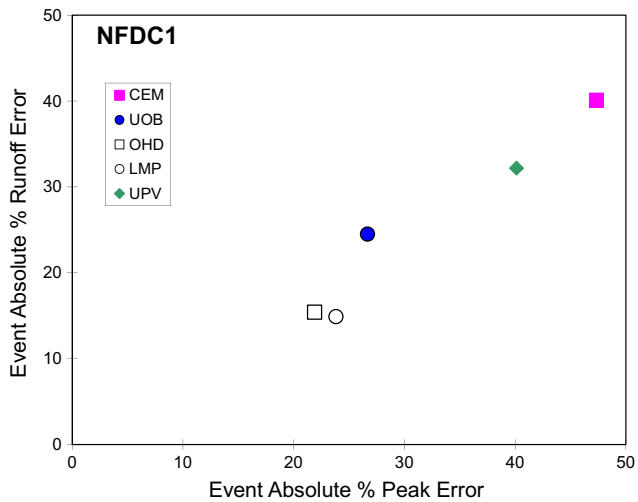

Fig. 10. Average runoff volume errors and peak errors for 68 events in the NFDC1 basin. Results are from calibrated models from the combined calibration and validation periods.

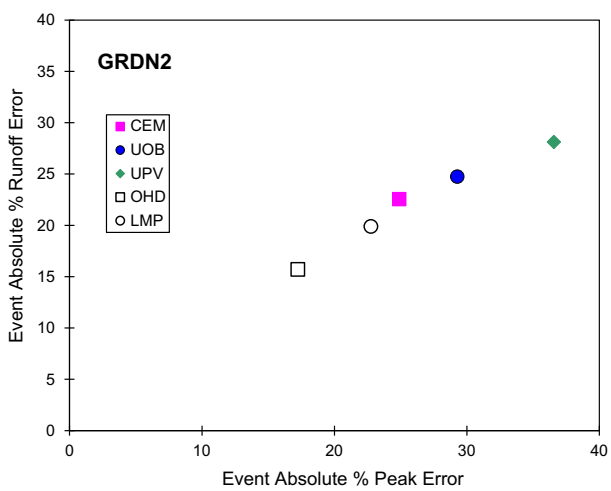

Fig. 11. Average runoff volume errors and peak errors for 92 events in the GRDN2 basin. Results are from calibrated models from the combined calibration and validation periods.

Interestingly three models show improvement in peak timing: OHD, CEM and UPV.

Taken as a group, distributed models provided improved peak flow simulations in 24% of the 17 model-basin pairs, improved runoff volume in 12% of the 17 model-basin pairs, and peak timing improvements in 41% of the 17 model-basin pairs. These values correspond to 24%, 18%, and 28%, respectively, achieved in the DMIP 2 Oklahoma tests (Smith et al., 2012b). However, caution is advised as the DMIP 2 Oklahoma results were based on a much larger number of model-basin pairs (148 vs. 17).

Visual inspection of the hydrograph simulations revealed that the UOB model also provided improved simulations for a few certain events, but these improvements were overwhelmed by other events in the average improvements statistics and not visible in Fig. 13.

While the results in this section may be discouraging, they are entirely consistent with the results from the Oklahoma experiments in DMIP 1 and 2. In DMIP 1, there were more cases when a lumped model out-performed a distributed model than vice versa (Reed et al., 2004). The results of the DMIP 2 Oklahoma

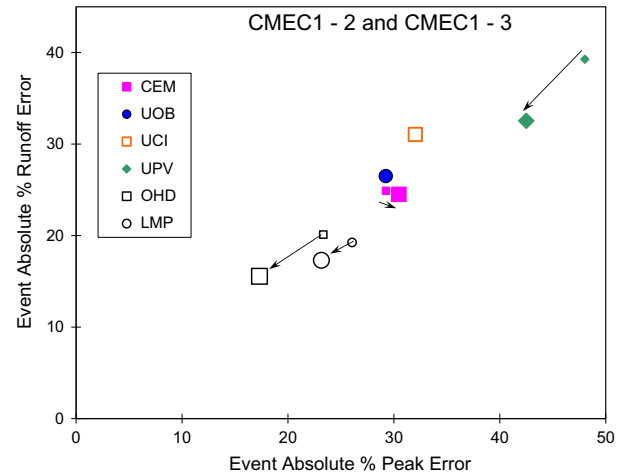

Fig. 12. Average runoff volume errors and peak errors for 92 events in the CMEC1 basin. Results are from calibrated models from the combined calibration and validation periods. These results are for Modeling Instruction 3 (CMEC1-3) in which CMEC1 was treated as an independent headwater basin calibrated at Markleeville, NV. These are the larger plotting symbols. Also shown are results from Modeling Instruction 2 (CMEC1-2), in which CMEC1 is treated as a 'blind' interior point within the calibrated GRDN2 parent basin. These are shown by the smaller symbols for each participant. The arrows show the direction from 'blind' to explicitly calibrated results.

experiments showed a greater number of cases of distributed model improvement than in DMIP 1 (Smith et al., 2012b).

The convention of Reed et al. (2004) is used herein to identify the cases in which the 'improvement' in Fig. 13 was negative but near zero. This shows that distributed models can perform nearly as well as a calibrated lumped model. The UCI, UOB, and UPV models show 'improvement' values less than −5% for peak volume and flow, and less than 1 h for peak timing.

3.6. Specific examples of distributed model improvement

To complement the event statistics presented in Section 3.5, we provide an analysis of two events in December, 1995 and January, 1996 to diagnose the source of distributed model improvement. While other events and models could have been selected, this case is used to intercompare distributed (OHD) and lumped (LMP) models that share the same precipitation/runoff physics.

Fig. 14 shows the results for the December, 1995 event in NFDC1. At least some of the improvement from the OHD model (and other distributed models in other cases) may result from improved definition of the rain/snow line, and the subsequent impacts on runoff generation. As an illustration, Fig. 15 shows the time evolution of the hourly rain/snow line as computed by the LMP model for the month of December, 1995. The diurnal variation of the rain/snow line is quite evident. The grey zone marks the four-day period from December 11th to December 14th, during which the LMP rain/snow line drops dramatically in elevation. Fig. 16 shows how the LMP model and the OHD model simulate the rain/snow line using the gridded temperature data at 15Z on December 12, 1995. The white line denotes the rain/snow line at 1758 m computed by the LMP model. Areas higher in elevation (to the right in the figure) than this receive snow, while areas below this line receive rain. The OHD model receives rain over a larger area than LMP as denoted by the red grid cells. Fig. 17 shows the precipitation and runoff for this event from the LMP and OHD models. The top panel shows that the OHD model generates more runoff than the LMP model. The middle two panels of Fig. 17 show how the LMP and OHD models partition total

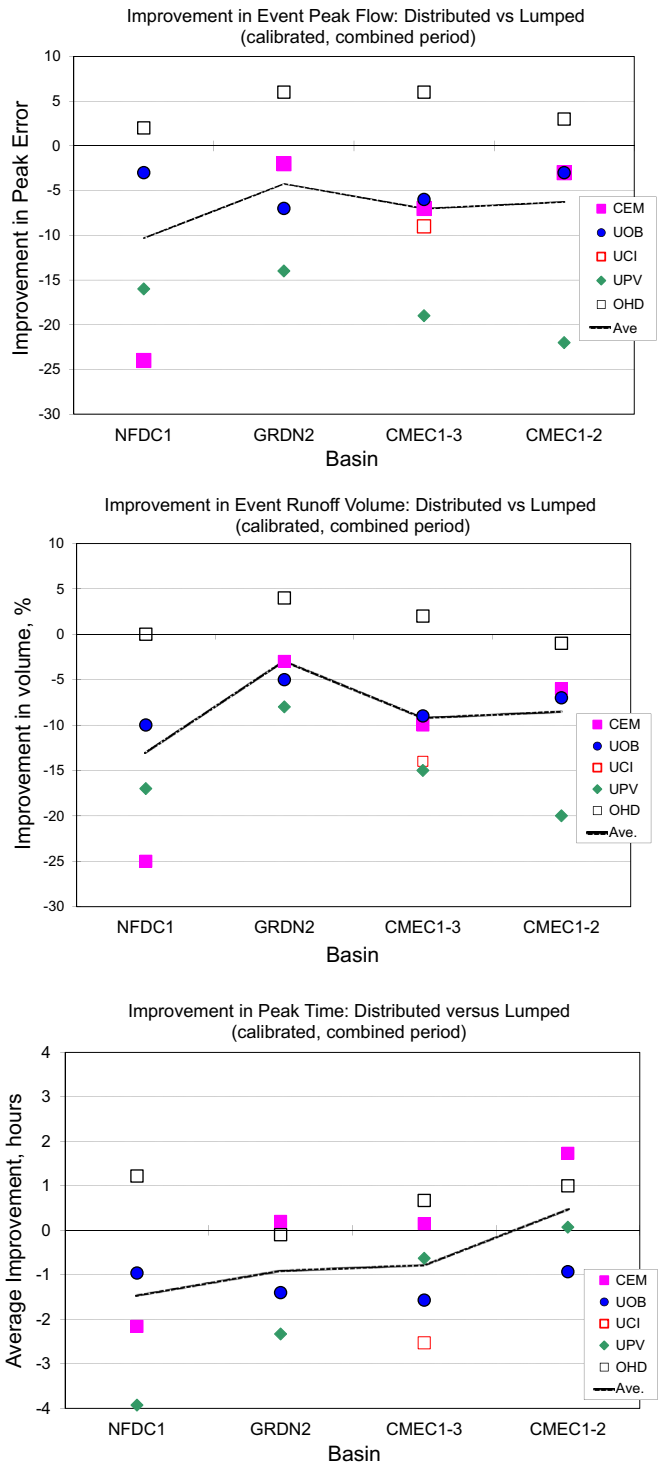

Fig. 13. Improvement of distributed models over lumped models: peak flow (top), event runoff volume (middle), and peak time (bottom).

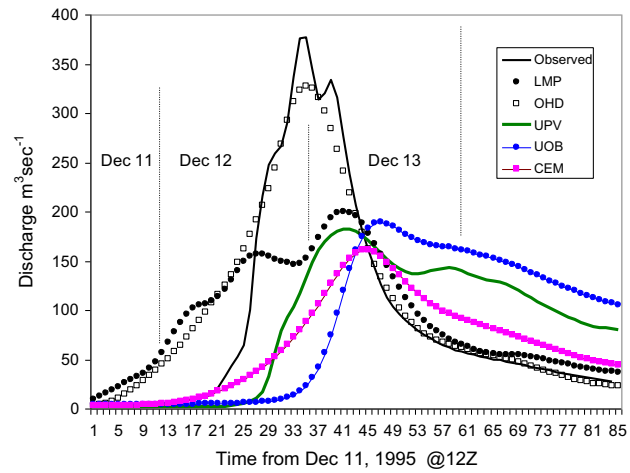

Fig. 14. Runoff event of December 12 and 13, 1995 in the North Fork. Simulated and observed hourly hydrographs are shown.

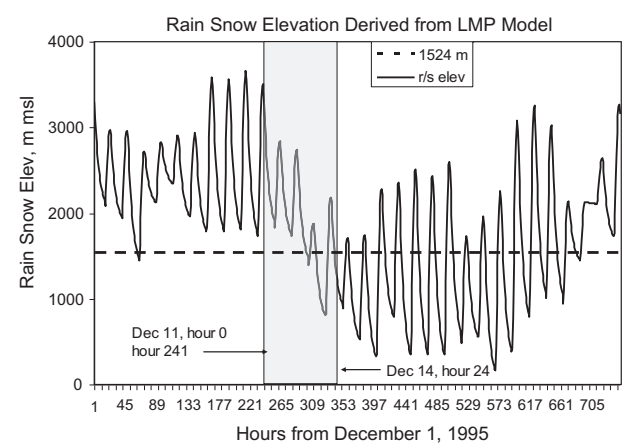

Fig. 15. Time evolution of the elevation of the hourly rain/snow line as computed by the LMP model for December, 1995. The grey shaded area corresponds to the period of December 11–14.

precipitation into different amounts of rain and snow. The bottom panel shows the arithmetic difference between the two rainfall time series.

The OHD model partitions total precipitation into a greater percentage of rain than the LMP model, and more runoff is generated by the OHD model compared to the LMP model. However, our analysis did not determine the exact cause of the increased runoff. The increase could be simply due to the greater amount of total rain input into the OHD model. Another cause could be that distributed models (OHD) preserve the precipitation intensity in each grid as shown in Fig. 18 rather than averaging it over the LMP elevation zones. As a result, fast-responding surface runoff is generated in several cells and is routed down the channel network. In the lumped model, the precipitation is spatially averaged over the elevation zones, resulting in a delayed runoff response. Another cause may be the difference in channel routing schemes between the OHD model (kinematic wave) and LMP (unit hydrographs and lag/k routing).

The contributions of the spatial variability of runoff generation and routing were also investigated. Hourly OHD surface and subsurface routing volumes shown (as accumulations) in Fig. 18 were averaged over the entire basin, then used as spatially uniform input to each 4 km grid into the OHD routing network. In this case, the resulting hydrograph (not shown) was very similar to the OHD hydrograph in Fig. 14 with the result that routing constitutes only 9% of the difference in root mean square (rms) error statistics between the OHD and LMP hydrographs. This small contribution seems reasonable given the central location of the runoff volumes. However, this is only an approximate comparison as the OHD and LMP runoff volumes prior to routing were not equal (Fig. 17, top panel).

In another case of distributed model simulation improvement (January 25, 1996; not shown), liquid precipitation and surface runoff were concentrated near the basin outlet. The routing contri-

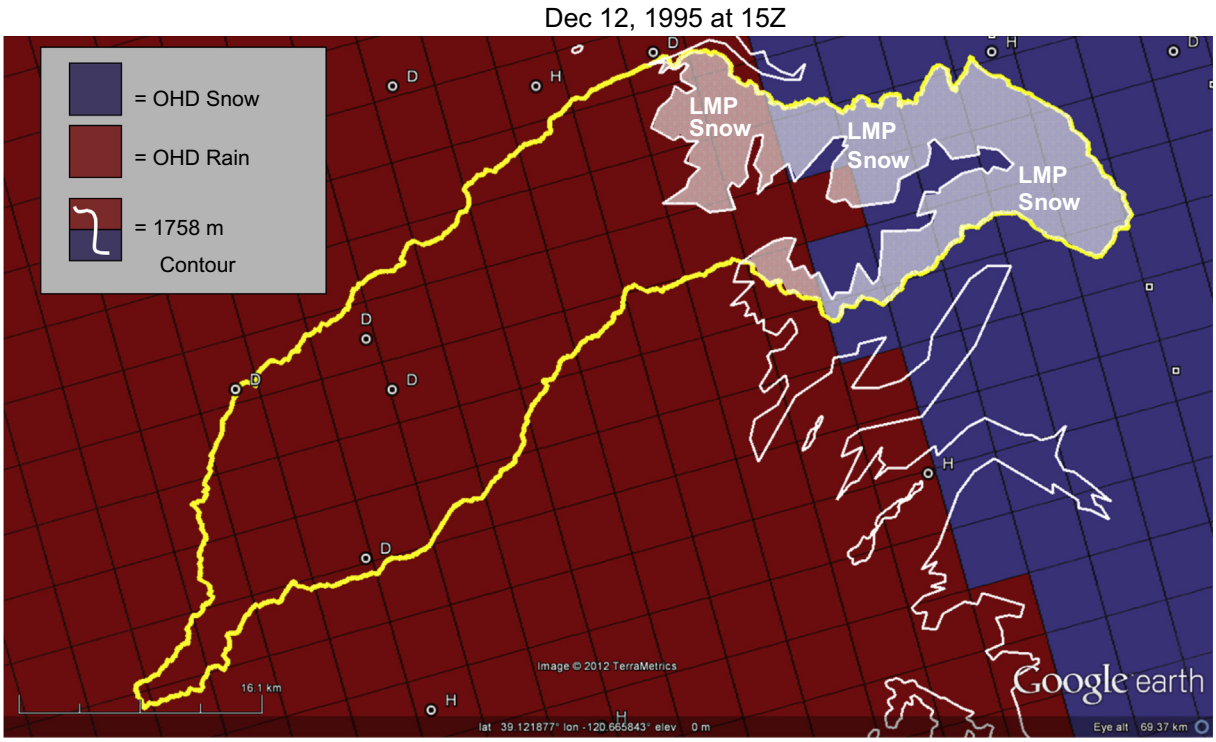

Fig. 16. Difference between the LMP model snow areas (to the right of the white 1758 m contour line) and the OHD distributed model snow area (red grid cells) for the hour ending at 15Z on December 12, 1995. The NFDC1 basin outline is shown in yellow. The grid shows the outline of the ~4 km HRAP cells.

bution to the difference in rms error between the OHD and LMP simulations in this case was much larger at ~36%. In this case, capturing the spatial variability of the liquid precipitation and subsequent runoff may be more important than differences in precipitation partitioning.

From our limited analysis, it is not clear which factor, or combination of factors, was the source of improvement of the OHD model. Beyond simply identifying the differences in precipitation partitioning and the spatial variability of precipitation and runoff, it was difficult to isolate the source of distributed model improvement. Further complications arise due to different initial conditions in the OHD and LMP model for the events studied. In one case, the spatial distribution of precipitation appeared to play a major role, while in another case, modeling the rain/snow areas seemed to have a large impact. Dornes et al. (2008) were able to attribute the benefits of their distributed model in two events to capturing the impact of topography on shortwave radiation for the modeling of snow accumulation and melt. However, their study did not address complexities arising from mixed rain and snow.

3.7. Effect of model parameter calibration

As in the DMIP 1 and 2 Oklahoma experiments, participants were instructed to submit simulations using calibrated and uncalibrated model parameters. This analysis was designed to assess the efficacy of *a priori* model parameters as well as schemes to calibrate hydrologic models in mountainous areas. The results shown are for the combined calibration and validation periods. The r_{mod} and %Bias measures are presented to provide an overall view of the impacts of parameter calibration.

The calibration results for NFDC1, GRDN2, and CEMC1-3 are presented in Fig. 19. The r_{mod} and %Bias measures for uncalibrated and calibrated results are shown connected by an arrow indicating the directional change in values (e.g., Viney et al., 2009). For NFDC1, calibration improved the r_{mod} and %Bias measures for several models: UPV, LMP, UOB, and OHD. The results for CEM do not show any gain in either measure, indicating that the calibration process focused on minimizing other error criteria. Indeed, the CEM model was calibrated using root mean square error (RMSE) calculated on root square transformed flows, which may explain why the change in r_{mod} is not so satisfactory.

For GRDN2, three of the four models achieved improved r_{mod} values, but at the expense of increasing (positive or negative) the overall %Bias (UOB, LMP, and OHD). The most consistent improvements from calibration were seen in the CMEC1-3 test in the bottom panel of Fig. 19. All models realized gains in the r_{mod} and %Bias statistics with the exception of UPV. The uncalibrated r_{mod} results for the OHD and CEM models were better than the calibrated results for the three remaining distributed models.

Although not shown here, parameter calibration resulted in clear improvements in nearly all cases of other statistics such as the Nash–Sutcliffe efficiency and root mean square error.

3.8. Analysis of interior processes: streamflow and SWE

This part of the DMIP 2 West experiments was designed to investigate how distributed models represent basin-interior processes in mountainous areas. Participants generated simulations of two variables: streamflow and SWE.

Only one interior flow point was available (specified as a "blind" test with no explicit calibration; CMEC1-2). The results for calibrated simulations are shown in Fig. 8, Fig. 9 and Fig. 10. The multi-year r_{mod} values for CMEC1-2 in Fig. 8 are only slightly lower than the CMEC1-3 case in which specific calibration was allowed. More of a difference between CMEC1-2 and CMEC1-3 is visible in Fig. 9, where larger multi-year %Bias values can be seen for CMEC1-2. This suggests that specific calibration at the Markleeville gage corrects for locally-generated biases that are not readily removed when calibrating using downstream information at Gardn-

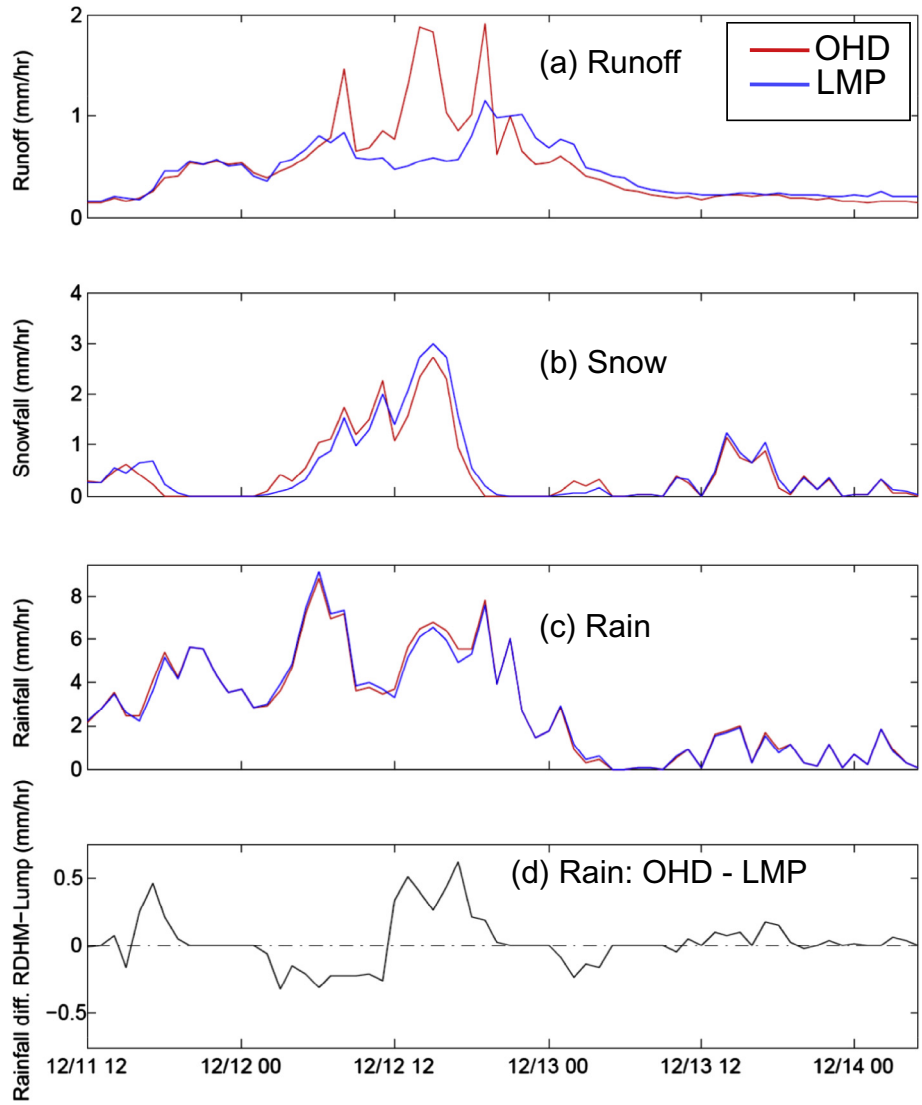

Fig. 17. Total precipitation and runoff for Dec, 1995 event in NFDC1 by the LMP and OHD models. All quantities are basin-average values. (a) Runoff generated by the LMP and OHD models. (b) Snowfall registered by both models. (c) Rain for both models. (d) Arithmetic difference in OHD rain and LMP rain shown in panel c.

erville. Fig. 12 shows that specific calibration of the interior point CMEC1-3 led to improved values of peak and runoff volume error statistics for events compared to CMEC1-2 for two models (UPV and OHD). As stated earlier in Section 3.4, the CMEC1-2 results may be influenced by the large size of the CMEC1 basin compared to the parent GRDN2 basin.

Participants were also requested to generate uncalibrated and calibrated hourly SWE simulations at two instrumented points in the NFDC1 basin and four instrumented points in the CMEC1 basin. The SWE simulations represent the grid or other computational area at which the model was run. Given the uncertainties involved in comparing point to grid values of SWE, our goal was to understand the general ability of the models to simulate the character of the snow accumulation and ablation processes (e.g., Shamir and Georgakakos, 2006). Shamir and Georgakakos (2006) defined a "good" SWE simulation as one that fell between simulated uncertainty bounds and also had consistent agreement with sensor observations.

We computed an average value of the %Bias of the simulated SWE compared to the observed SWE. These average values represent the entire snow accumulation and ablation periods (approximately October to June) for all years at each of the six snow gage sites. Table 4 presents the results of this overall analysis for calibrated models. For the two stations in the NFDC1 basin, %Bias values were greater in absolute magnitude for the Blue Canyon site compared to the Huysink site for all four models. This agrees with Shamir and Georgakakos (2006) who found that the largest uncertainty was for that part of the snow pack located where the surface air temperature is near the freezing level. With an elevation of 1609 m, the Blue Canyon site is near the elevation of 1524 m typically used by the CNRFC to delineate rain and snow in its river forecast operations.

The %Bias values at the four SNOTEL sites in or near the higher elevation CMEC1-3 basin are generally less than those of the stations in the lower elevation NFDC1 basin. Here, the UOB and OHD models achieved the lowest values of the %Bias measure.

Large %Bias values can be seen in Table 4, highlighting the difficulties of simulating SWE in mountainous areas. For example, the UCI model featured a large over-simulation of SWE for the Spratt Creek SNOTEL site. This is due to the fact that the SWE sim-

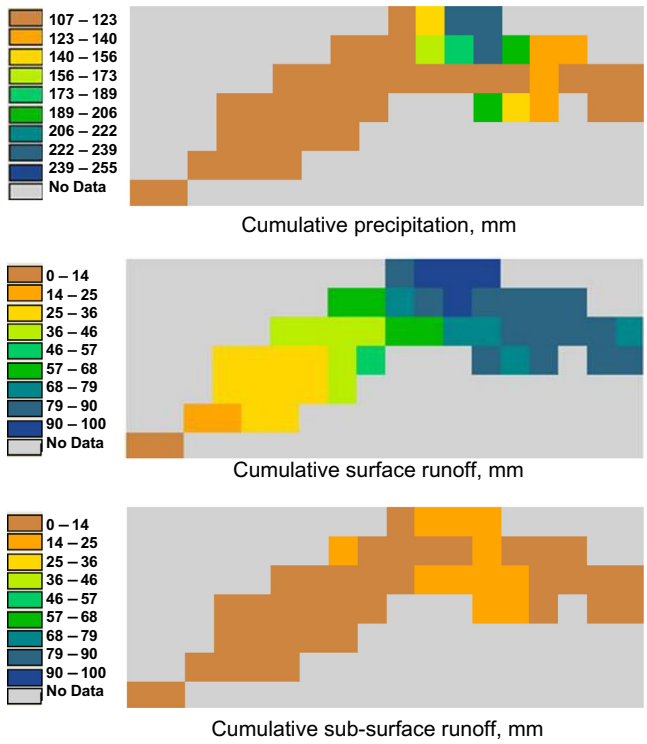

Fig. 18. Spatial variability of precipitation and OHD runoff components for the December 1995 event in NFDC1. The top panel shows the cumulative precipitation (rain and snow, mm), the middle panel shows the cumulated fast-responding surface runoff (mm), and the bottom panel shows the cumulated slow-responding sub-surface runoff (mm).

ulation for this site was generated by the UCI sub-basin model in which the average elevation of the sub-basin containing the site is over 400 m higher than the site itself—thus leading to overly frequent snowfall events. The large %Bias values may also reflect the scale mismatch between point SWE observations and simulations of SWE generated over a grid, sub-basin, or elevation zone.

Fig. 20 presents the simulated and observed SWE for a large snow year for the Blue Canyon site in the NFDC1 basin and the Blue Lakes site in the CMEC1-3 basin. The dates plotted are October 1, 1992 to July 31, 1993. The Blue Lakes site accumulates about twice as much SWE as does the Blue Canyon site. Up to day 151, all models accumulate snow at a rate similar to that observed for Blue Lakes. However, the onset of melt is different amongst models, and all models melt off the snow more quickly than is observed. Problems with snow accumulation are more evident for the Blue Canyon site, perhaps as a result of difficulties in tracking the rain/snow line, intermodel differences in treating precipitation gage undercatch due to wind, and differences in how models determine the form of precipitation.

One issue that arises in modeling studies is whether errors (uncertainty) in precipitation and temperature forcings mask the simulation differences resulting from model physics. Fig. 21 illustrates how intermodel spread in SWE simulations compares to differences in SWE caused by errors in the temperature forcing. We use the work of Lei et al. (2007), who compared the responses of the snow model in the Noah LSM (Koren et al., 1999) and the Snow-17 model to random and systematic errors in temperature, solar radiation, and other meteorologic forcing variables available in the NARR (Mesinger et al., 2006). SWE simulations for the Snow-17 model with different levels of random error in the temperature data are shown in the top panel for water year 1999. The middle panel shows the SWE simulations from the Noah model for the same water year and same levels of random temperature error. The spread in the DMIP 2 West SWE simulations at the Blue Lakes SNOTEL site for water year 1999 is shown in the bottom panel of Fig. 21. It can be seen from this figure that the spread in model-generated SWE could be as great as the spread caused by

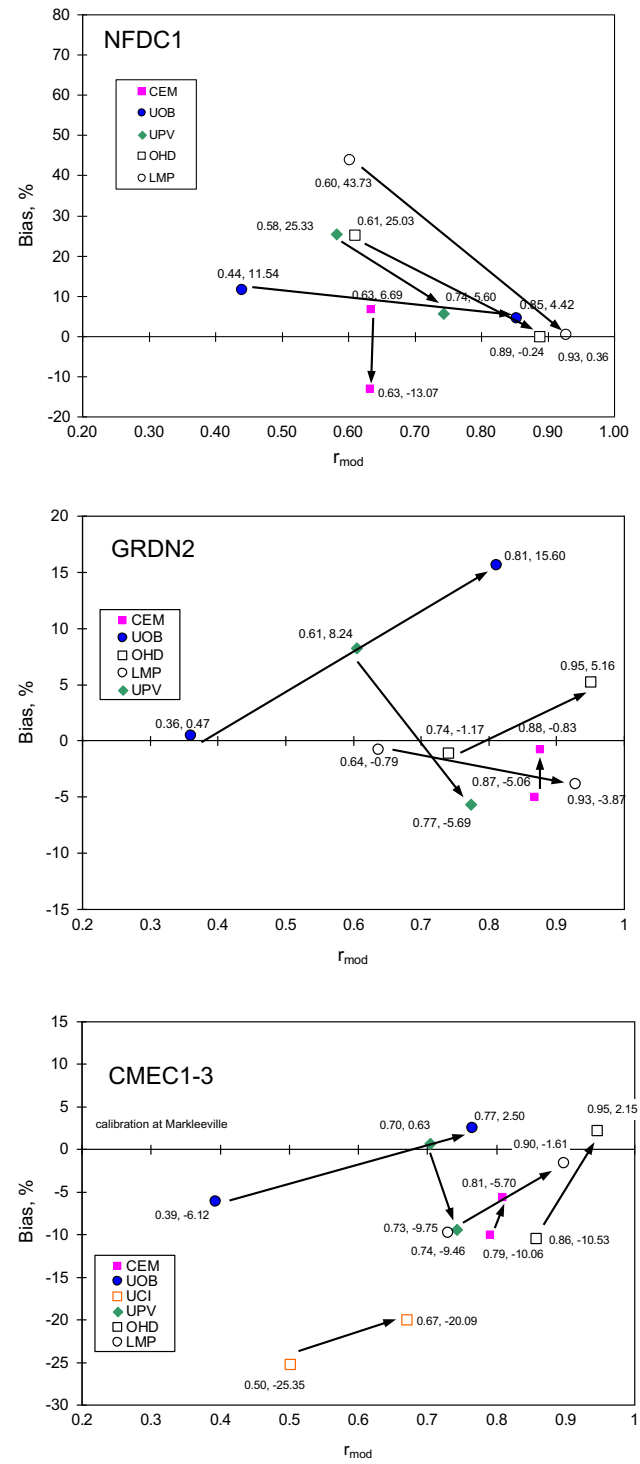

Fig. 19. Improvement in r_{mod} and %Bias achieved via model calibration: NFDC1 (top), GRDN2 (middle), and CMEC1-3 (bottom). Arrows indicate the direction from uncalibrated to calibrated statistics. Results are from the combined calibration and validation periods.

Table 4
Computed vs. observed SWE at six sites. Values shown are the multi-year average %Bias of simulated SWE compared to observed SWE.

Basin	Station	Elev. m	Participant				
			CEM	UOB	OHD	UCI	UPV
NFDC1	Blue Canyon	1609	−93.74	−84.93	−57.96	–	113.49
	Huysink	2012	−87.30	−56.90	−34.68	–	−13.60
CMEC1-3	Blue Lakes	2456	−86.80	−7.04	13.85	−15.89	−11.67
	Ebbett's Pass	2672	−69.06	−11.39	−14.85	−31.69	−21.72
	Poison Flats	2358	−79.89	−9.66	−15.36	21.49	−37.55
	Spratt Creek	1864	−77.49	−10.08	−27.80	560.34	29.63

random errors in the temperature data, depending on the level of data error.

The timing of the overall snow accumulation and melt was evaluated by computing the difference between the observed and simulated SWE centroid dates (SCD; Kapnick and Hall, 2010). The SCD is computed using Eq. (3):

$$SCD = \frac{\sum t_i \times SWE_i}{\sum SWE_i} \tag{3}$$

where SWE is the daily observed or simulated SWE in mm, t is the number of the day from the beginning of snow accumulation, and i denotes an individual SWE value.

Fig. 22 shows the difference SCD for each site for each of the years of the combined calibration and validation periods. Each box represents the 25–75% quartile range of SCD differences in days, while the red line is the median of the values. In Fig. 22, it is desirable to have an SCD difference of zero. Looking collectively at the plots, there is a general trend for the models to have an early SCD compared to the observed SCD. The values for the Huysink site in the NFDC1 basin have the largest and most consistent departure

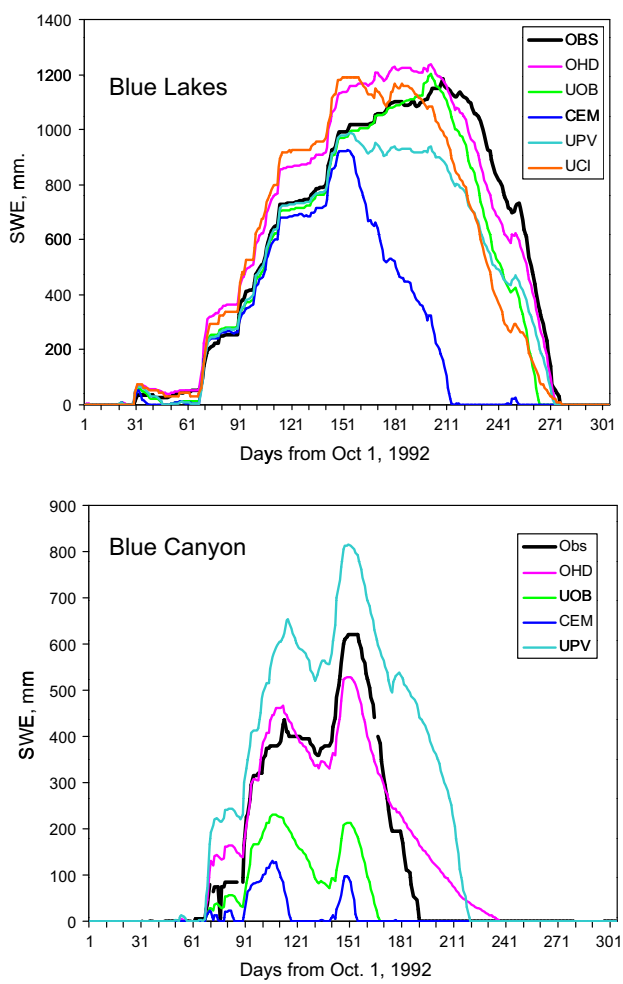

Fig. 20. Simulated and observed SWE for water year 1993: Blue Lakes SNOTEL site (top) and the Blue Canyon site operated by the Bureau of Reclamation (bottom).

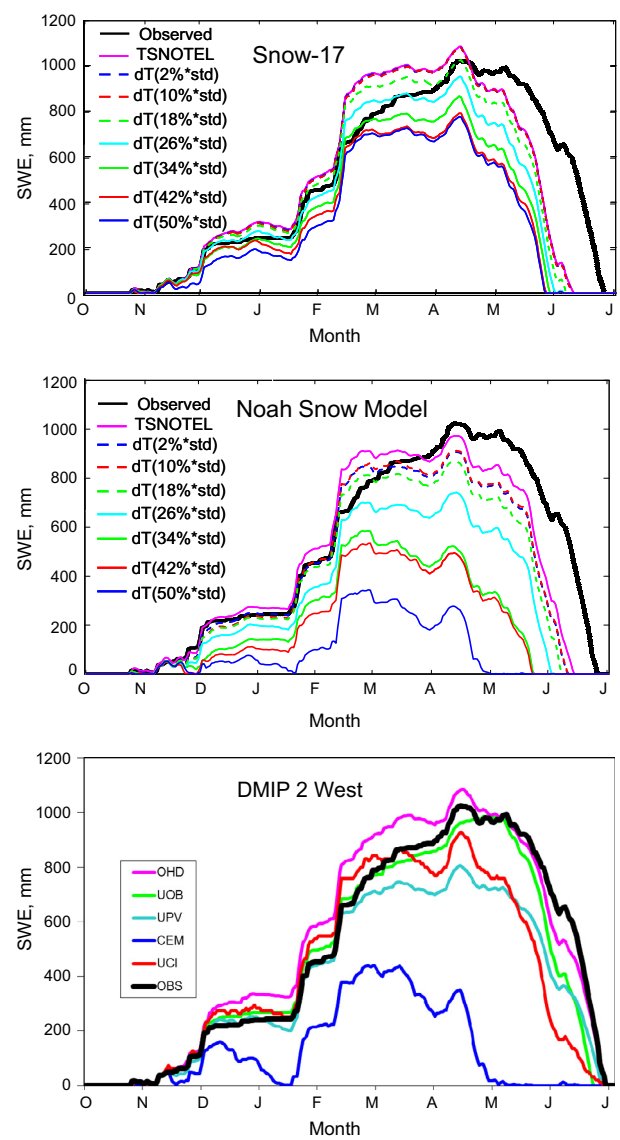

Fig. 21. SWE simulations for the Blue Lakes SNOTEL station for water year 1999. SWE simulations from Snow-17 (top) and Noah snow model (middle) using different levels of random temperature error are taken from Lei et al. (2007). DMIP 2 West simulated SWE is shown in the bottom panel. The observed SWE is the same in all three cases and is shown as a heavy black line.

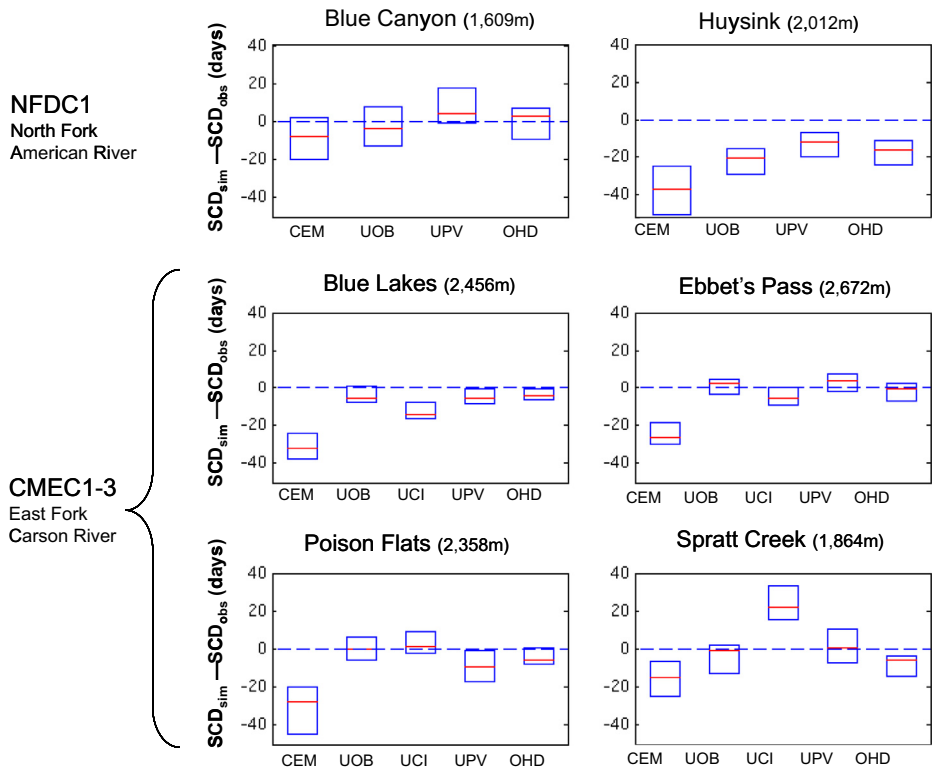

Fig. 22. Difference in days between observed and simulated SCD each year for SWE measurement sites. Blue boxes represent the 25–75% quartiles, while the red line denotes the median.

from zero. There is a tendency in the NFDC1 basin for the SCD to be slightly earlier in time compared to the observed value. The participants' SCD values for the Huysink site are consistently about 10–15 days earlier than the Blue Canyon station. Consistent with Shamir and Georgakakos (2006), better results were achieved at higher elevation snow sites. The best SCD results are for the two highest elevation stations, Blue Lakes and Ebbet's Pass. At these sites, four models achieved the best timing as evidenced by the smallest spread in the 25–75% quartile range and SCD differences less than 10 days. Larger errors in timing (Fig. 22) and bias (Table 4) could be the result of scale mismatch between the point observations and the models' computational element size. For example, the larger errors seen in the CEM snow simulations may be the result of using a coarse modeling scale (5 elevation zones).

The statistics presented here should be viewed with caution as considerable uncertainty exists in the representativeness of point SNOTEL (and other) observations of SWE to the surrounding area (e.g., Shamir and Georgakakos, 2006; Dressler et al., 2006; Garen and Marks, 2005; Simpson et al., 2004; Pan et al., 2003). Of course, point-to-point comparisons can be made when such data are available (e.g. Rutter et al., 2009; Rutter et al., 2008). However, there is great variability amongst results even when models are run at the point scale and compared to research-quality point observations. For example, the SnowMIP 2 project examined the performance of 33 snow models of various complexities at four point sites (Rutter et al., 2009). One of the conclusions from SnowMIP 2 was that it was more difficult to model SWE in forested sites compared to open sites. Moreover, there was no 'best' model or subset of models, and models that performed well at forested sites did not necessarily perform well (in a relative sense) at open sites. Along these lines, Mizukami and Koren (2008) noted discrepancies between satellite-estimated forest cover and the description of cover contained in a station's metadata. Such discrepancies could impact model simulations.

3.9. Modeling scale

DMIP 2 West intended to examine the science question of appropriate model scale. In the project formulation phase, it was hoped that there would be a sufficient number of models to make inferences between simulation results and model scale. Unfortunately, the number of participants (five plus LMP) did not allow us to investigate this issue as fully as hoped. Nonetheless, a brief discussion is provided here given the void in the literature on this aspect of modeling.

The spatial modeling scales ranged from grids of 250 m (UOB), to 400 m (UPV) to 4 km (OHD) to elevation zones (five in CEM; two in LMP) and to sub-basins (8 sub basins with average size 90 km^2, UCI). Runoff statistics reflect the integration of many processes including snow accumulation and melt, rain on snow, rain on bare ground, and hillslope and channel routing. Accordingly, we acknowledge that the results portray a mix of both model physics and modeling scales and that definitive conclusions regarding modeling scale cannot be made. Nonetheless, it is interesting that no trends between performance and modeling scale can be seen in the calibrated r_{mod} and %Bias plots of Fig. 8 and Fig. 9, respectively, for the calibration, validation, and total simulation periods. Results are plotted in Fig. 8 and Fig. 9 in order of increasing model resolution. Even where the model physics is similar (i.e., OHD, LMP, and UCI), there is no trend in these run-period statistics.

However, a different picture emerges when looking at the event statistics for calibrated models. Focusing on the two models with common physics which were run in all cases, the average

event statistics in Figs. 10–12 for calibrated models show that OHD provides better values than LMP. In all three basin cases, the OHD model provides lower peak error values, and in two out of three cases the runoff volume is better. Recall that LMP is actually the case of two elevation zones, above and below 1524 m in NFDC1 and 1724 m in GRDN2 and CMEC1. The event improvement statistics in Fig. 12 further illustrate the improvement of OHD compared to LMP. Uncalibrated run period and event statistics (not shown) also show improvements of OHD compared to LMP. Comparing models with the same physics, the OHD and LMP results agree with the scope and results of Dornes et al. (2008), who modeled the snow and runoff processes of a basin with lumped and distributed applications of the same precipitation-runoff model. In a limited way, the results herein support the expectation that higher resolution modeling scale will improve simulation performance.

3.10. Rain and snow partitioning

DMIP 2 West intended to address the important issue of rain-snow partitioning. Primarily this was to be addressed via the use of radar-detected observations of freezing level from HMT-West (Minder and Kingsmill, 2013; White et al., 2010, 2002) after participants set-up and ran their models with the baseline DMIP 2 West gridded precipitation and temperature data. Participants would then make additional simulations using the radar-based estimates of the rain-snow line to note the improvement. Delays in DMIP 2 West caused by the need to generate a new QPE data set precluded the use of HMT-West data in formal experiments with participants. Nonetheless, using the DMIP 2 West modeling framework, Mizukami et al. (2013) tested the OHD and LMP models with and without the HMT-West radar-derived rain-snow data for the 2005–2006 winter period. Mixed simulation results were seen; some runoff events were better simulated while other events were worsened. Interested readers are referred to Mizukami et al. (2013) for more information.

4. Conclusions

We present the major conclusions generally in order of the science questions listed in Section 1.2. Interspersed among these are additional conclusions and comments.

4.1. Distributed vs. lumped approaches in mountainous areas

Overall, no single model performed best in all basins for all streamflow evaluation statistics. Neither was any distributed model able to consistently outperform the LMP benchmark in all basins for all indices. Nonetheless, one or more distributed models were able to achieve better performance than the LMP benchmark in a number of the evaluations. These results are consistent with the findings of DMIP 1 (Reed et al., 2004) and DMIP 2 West (Smith et al., 2012b). We highlight several aspects of model performance below.

Considering the r_{mod} and %Bias measures computed for the multi-year calibration, validation, and combined calibration–validation periods, mixed results were achieved. No single model performed best in all periods in all basins. In addition, no distributed model consistently performed better than the benchmark LMP model. However, three models (OHD, UOB, and CEM) were able to outperform the LMP model for certain periods in certain basins.

The models were also inter-compared by evaluating specific precipitation/runoff events. Here, only one model (OHD) was able to perform at a level near to or better than the LMP benchmark for peak flow and runoff volume. However, three models (OHD, CEM, and UPV) achieved improvements in peak event timing compared to LMP, highlighting the potential of distributed models to capture spatially-variable precipitation and runoff processes. This evaluation of precipitation/runoff events showed that taken together, distributed models were able to provide improved peak flow values in 24% of the 17 model-basin pairs, improved runoff volume in 12% of the pairs, and improved peak timing in 41% of the pairs.

Even though the gains by distributed models over the LMP benchmark were modest, all models performed well compared to those in the less-hydrologically-complex Oklahoma basins in DMIP 2. For example, the r_{mod} and %Bias results of all models in the multi-year run-period tests are commensurate with those in the non-snow-dominated DMIP 2 Oklahoma basins (Smith et al., 2012b). Similarly, the event-averaged absolute % runoff error and absolute % peak error values agree well with the range of values for the DMIP 2 ELDO2 basin (Smith et al., 2012b). These results are noteworthy in that the DMIP 2 West basins have complexities such as orographic enhancement of precipitation, snow accumulation and melt, rain-on-snow events, and highly varied topography which are not present in the DMIP 2 Oklahoma basins. Looking at the results herein and the DMIP experiments overall, it is clear that at this point in their evolution, distributed models have the potential to provide valuable information on specific flood events that could complement lumped model simulations.

Based on these mixed results, care must be taken to examine a range of statistical measures, simulation periods, and even hydrograph plots when evaluating the performance of distributed models compared to lumped models. As in the DMIP 2 Oklahoma basins (Smith et al., 2012b), our results actually reflect the model/user combination and not the models themselves.

It proved difficult to determine the dominant factors which led to the improvement of the OHD distributed model over the benchmark LMP model for mixed rain/snow events. Our limited analyses on one of two study basins identified complex interactions of precipitation partitioning, spatial variability of liquid precipitation, runoff generation, and channel routing.

4.2. Estimation of models inputs and model sensitivity to existing data

The distributed models used gridded forms of the precipitation and temperature data widely used by NWS RFCs in mountainous areas for hydrologic model parameter calibration. In the study basins, the density of precipitation and temperature gauges was sufficient to develop useful gridded estimates of these variables over a 20-year span. A sufficient number of hourly (recording) rain gauges were available to distribute daily precipitation observations. These data were able to support effective model calibration and good simulations through the validation period, evidenced by %Bias values within or near the ±5% criteria for NWS model calibration (Smith et al., 2003), low cumulative runoff errors, and high values of r_{mod}.

For this study, careful quality control of the raw precipitation data was essential. This seemed especially warranted given the sensitivity of the hydrologic models noted in the development of the QPE data. Numerous errors, especially in the precipitation observations, were identified and corrected. The OHD and LMP models were sensitive to these errors for hourly time step simulations of mixed rain/snow events. Such errors manifested themselves as anomalous hydrograph peaks. The impact of such precipitation data errors may not be as evident in streamflow hydrographs that are dominated by snow melt.

4.3. Internal consistency

The ability of distributed models to simulate snow accumulation and melt was investigated at six SWE observation sites. The best results in terms of timing and volume were seen at the higher elevation stations. Larger errors in simulated SWE were apparent at a station near the typical elevation separating rain and snow. In addition, larger errors in timing and volume of snow accumulation and melt were seen in the distributed models featuring larger computational element sizes. This result may reflect the scale mismatch between the point observations and the computational element size. Our findings should be viewed in light of the considerable uncertainty that exists in the SWE observations and their representativeness of the surrounding area.

A limited test with one interior flow point showed that some distributed models calibrated at the outlet were able to achieve valid simulations of streamflow at the interior location. In particular, the overall r_{mod} statistics for the blind interior point CMEC1-2 were commensurate with those achieved through explicit calibration at the interior point (CMEC1-3). However, these good results may hinge on the large size of the interior basin compared to the parent basin.

4.4. Scale issues

Scale issues continue to be perplexing in mountainous areas. While our study was limited in scope, the results address a void in the literature regarding modeling scales that consider snow accumulation, melt, and runoff generation. Even in the highly variable terrain of NFDC1, a range of modeling scales led to relatively good streamflow simulations. Among the models that shared the same snow/rainfall/runoff schemes, better event statistics were achieved at higher resolution modeling scales. Considering all the models, which admittedly represented a mix of physics, user knowledge, and model scales, it was surprising that more apparent trends did not appear given the range of modeling resolution from 250 m to two elevation zones. We were not able to pinpoint an optimal modeling scale.

4.5. Parameter calibration

While not an explicitly identified science question, our results show that parameter calibration led to improved goodness-of-fit statistics for nearly all model-basin pairs. This suggests that calibration strategies can be effective in areas with complex hydrology. It also suggests that calibration strategies are needed even with advances in model structure and the development of *a priori* parameter estimates.

5. Recommendations

While DMIP 2 West provided interesting and informative results, much work remains to further address the science questions posed in DMIP 2 West and other issues that have bearing on mountainous area hydrologic simulation and forecasting.

Our results should be further examined in the context of uncertainty in forcing data and model parameters. For SWE, one simple method would be to use the approach of Shamir and Georgakakos (2006). In their approach, the uncertainty bounds were defined by running the snow model on an adjacent south-facing grid cell and a nearby north-facing grid cell. The resultant SWE simulations formed the lower and upper uncertainty bounds, respectively. Another idea is to use the results from Molotch and Bales (2006) to understand the relationship between SNOTEL SWE observations and the SWE simulations generated over the computational units within the participants' models. The amount of forest cover at each SWE site should be derived so that our results can be placed in the context of the SnowMIP 2 results (Rutter et al., 2009).

Continued efforts are necessary to diagnose the causes of differences between distributed and lumped model simulations. These efforts will require detailed analyses, probably along the lines of the hydrologic and meteorological studies of Lundquist et al. (2008) and Minder et al. (2011), respectively. While additional work can be done with the data on hand, the advanced data available from the HMT-West program will undoubtedly aid in this process.

DMIP 2 West was formulated as a general evaluation of distributed and lumped models in complex terrain without specific tests to highlight the benefits of model structure. To address this limitation, experiments are recommended to uncover and diagnose the impacts of model structure on performance (e.g., Clark et al., 2011; Butts et al., 2004).

The benefits of using the HMT-West data sets of additional surface temperature (Lundquist et al., 2008) and precipitation, optical disdrometer, vertically pointing radar-based freezing level (Mizukami et al., 2013; Minder and Kingsmill, 2013; Lundquist et al., 2008; White et al., 2010, 2002), soil moisture, and gap-filling radar-derived QPE (e.g., Gourley et al., 2009) should continue to be explored. These data sets should also aid in the diagnosis of modeling improvements. DMIP 2 West was always intended to be a multi-institutional and multi-model evaluation of the QPE, disdrometer, soil moisture, radar-freezing level, and other observations afforded by the rich instrumentation deployments in HMT-West. The intent was to first generate streamflow and SWE simulations using the 'basic' DMIP 2 West gage-only QPE and temperature fields. After calibrating and running their models with the basic data, it was planned to have participants rerun their models using the HMT-West data (radar QPE, snow level, and soil moisture) to note the improvements gained by advanced observations. However, both DMIP 2 West and HMT-West experienced major delays, with the unfortunate result being that the HMT-West data sets could not be explored in formal DMIP 2 West experiments.

Based on our experience with deriving precipitation, temperature, and evaporation forcing data sets for DMIP 2 West, continued work in deriving these forcings in complex terrain is of near-paramount importance for model testing, development, and calibration. Continued work is needed to address gage network density issues in mountainous areas. This is true for gage-only QPE and for the use of rain gages to bias-adjust radar estimates of precipitation. In spite of the enormous effort involved, data sets covering a large number of basins would support additional experiments and lead to broader conclusions (e.g., Andréassian et al., 2009, 2006).

River Forecast Centers within the NWS should consider the use of the OHD model. Other operational forecasting agencies should consider the use of distributed models in complex terrain. For the foreseeable future, such models should be viewed as complements to existing lumped forecast models rather than outright replacements.

Acknowledgements

We acknowledge the participants' institutions for their extensive support of the DMIP 2 West modeling efforts. We thank Gary Estes, organizer of the California Extreme Precipitation Symposium, for his help with finding cross section information. Ken

Appendix A. Models used in DMIP 2 West (See Table A1)

Table A1
Participating groups and major model characteristics.

Participant	Modeling system name	Primary reference	Primary application	Spatial unit for rainfall-runoff calculations	Rainfall-runoff/vertical flux model	Snow model	Rain/snow partitioning method	Channel routing
CEMAGREF (CEM)	GR4J	Perrin et al. (2003), Mathevet (2005), Le Moine (2008), Valéry (2010)	Streamflow, hydrological simulation, and forecasting	Lumped with five elevation zones	Soil moisture accounting	CemaNeige (degree day type model)	Threshold temperature in each of five elevation zones	Unit hydrograph and non-linear channel routing
U. Bologna (UOB)	TOPKAPI-X	Liu and Todini (2001), Todini and Ciarapica (2002)	Real time flood forecasting	250 m grid cells	Non-linear reservoir with Dunne infiltration. Percolation by Clapp and Hornberger. ET by Thornthwaite scaled on an hourly basis	Mass and energy balance	Percent of precipitation as rain on each cell: $F(T) = \frac{1}{1+e^{-\frac{T-T_s}{\sigma}}}$ where $\boldsymbol{\sigma}$ = 0.3, $\boldsymbol{T}$ is air temperature, $\boldsymbol{Ts}$ is derived from historical data by plotting frequency of precipitation type vs. air temp.	Kinematic Wave for slopes greater than 0.1% and Modified Muskingum–Cunge for slopes lower than 0.1%
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	Snow-17	Threshold temperature	Kinematic wave
NWS Office of Hydrologic Development (LMP)	NWSRFS	Burnash (1995)	Streamflow, water resources forecasting	Elevation zones	SAC-SMA	Snow-17	Threshold temperature, lapsed, area elevation curve	Unit hydrograph, Lag/K
Technical U. Valencia (UPV)	TETIS v8	Francés et al. (2007), Valéry et al. (2009)	Streamflow for flood and continuous simulation	400 m grid cells	Japanese Tank model	Degree-day	Threshold temperature	Geomorphologic Kinematic Wave
U. Ca Irvine (UCI)	Semi-distributed SAC-SMA and SNOW17	Ajami et al. (2004), Khakbaz et al. (2012)	Streamflow simulation	8 Sub-basins with the average area of 90 km^2	SAC-SMA	Snow-17	Threshold temperature	Kinematic wave

Appendix B. Calibration strategies for DMIP 2 West models (See Table B1)

Table B1
Calibration strategies for the DMIP 2 West models.

Group	Strategy
LMP	Systematic manual adjustment of parameters starting with baseflow and proceeding to fast response flow generation processes (Smith et al., 2003). Several statistical measures used at different points in the process to evaluate the fit of the simulation
OHD	Start with *a priori* parameters defined from soil texture. 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. Evaluate if this initial scaling is appropriate. Then use scalar multipliers to uniformly adjust each parameter field while maintaining spatial variability. Scalars are calibrated manually and/or automatically. Automatic calibration uses a multi-time scale objective function (Kuzmin et al., 2008)
CEM	The parameters were estimated using a steepest-descent type method combined with overall prior screening of the parameter space (see e.g. Mathevet, 2005). The resulting parameters were applied identically to all sub-basins
UPV	Use correction factors to globally modify each parameter map, assuming the prior spatial structure and thus reducing drastically the number of variables to be calibrated. In the used TETIS configuration, there were a total of nine correction factors: eight affecting the runoff production parameter maps and one for the stream network velocity. The TETIS model includes an automatic calibration module based on the SCE-UA algorithm (Duan et al., 1994). For this application, the objective function was the Nash–Sutcliffe efficiency index. The calibration was carried out in three steps: (1) calibration of rainfall-runoff parameters using the 1990 summer period; (2) calibration of snowmelt parameters using the 1992–93 and 1994–95 winter periods; (3) refinement of rainfall-runoff parameters using the period 1989–1993
UCI	The optimal parameter set was estimated through calibration of the lumped SAC-SMA and SNOW-17 models (over the entire watershed) using SCE-UA calibration algorithm (Duan et al., 1993) and MACS calibration scheme (Hogue et al., 2000). The resultant parameter set was then applied identically to all sub-basins in the distributed model configuration in order to generate streamflow at the outlet and interior points
UOB	Soils parameters derived from STATSGO texture classes in each grid. Adjusted horizontal and vertical saturated hydraulic conductivity and soil depth. Calibration process was carried out with a 'trial and error' methodology, focusing on the highest flood events in order to obtain the best model performance according to the Nash and Sutcliffe coefficient (Coccia et al., 2009)

Appendix C. Snow water equivalent stations and observation times (See Table C1)

Table C1
Characteristics of the six snow observation sites.

Basin	Station	Latitude/longitude	Elev.m	SNOTEL or CDEC[1] ID	Owner	Observation time
North Fork American	Huysink	39.282/−120.527	2012	HYS	U.S. Bureau Reclamation	0400 Local time; 1200 Z time
	Blue Canyon	39.276/−120.708	1609	BLC	U.S. Bureau Reclamation	0400 Local time; 1200 Z time
East Fork Carson	Blue Lakes	38.608/−119.924	2456	356	NRCS SNOTEL	2400 Local time; 0800 Z time (next day)
	Ebbetts Pass	38.550/−119.805	2672	462	NRCS SNOTEL	2400 Local time; 0800 Z time (next day)
	Poison Flats	38.506/−119.626	7736	697	NRCS SNOTEL	2400 Local time; 0800 Z time (next day)
	Spratt Creek	38.666/−119.818	6115	778	NRCS SNOTEL	2400 Local time; 0800 Z time (next day)

[1] California Data Exchange.

Howard of NSSL assisted with processing the PRISM data. The careful reviews and comments from the journal reviewers and editors have contributed to the clarity of this paper.

Appendix D. Statistical equations used in the analysis of DMIP 2 West results

D.1. Percent bias, PB (%)

PB is a measure of total volume difference between two time series. PB is computed as:

$$PB = \frac{\sum_{i=1}^{N}(S_i - O_i)}{\sum_{i=1}^{N} O_i} \cdot (100)$$

where S_i is the simulated discharge for each time step i, O_i is the observed value, and N is the total number of values within the time period of analysis.

D.2. Simulated or observed mean

$$\overline{Y} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

where Y is any type of data value.

D.3. Standard deviation, σ

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}{N-1}}$$

D.4. Correlation coefficient, r

While not used explicitly in the DMIP West results analysis, we present the formula for the correlation coefficient as background

for the discussion on the modified correlation coefficient. The correlation coefficient r is defined as:

$$r = \frac{N \cdot \sum_{i=1}^{N} S_i \cdot O_i - \sum_{i=1}^{N} S_i \cdot \sum_{i=1}^{N} O_i}{\sqrt{\left[N \cdot \sum_{i=1}^{N} S_i^2 - \left(\sum_{i=1}^{N} S_i\right)^2\right] \cdot \left[N \cdot \sum_{i=1}^{N} O_i^2 - \left(\sum_{i=1}^{N} O_i\right)^2\right]}}$$

D.5. Modified correlation coefficient, r_{mod} (McCuen and Snyder, 1975)

In this statistic, the normal correlation coefficient is reduced by the ratio of the standard deviations of the observed and simulated hydrographs. The minimum standard deviation (numerator) and maximum standard deviation (denominator) are selected so as to derive an adjustment factor less than unity:

$$r_{\text{mod}} = r \cdot \frac{\min\{\sigma_{sim}, \sigma_{obs}\}}{\max\{\sigma_{sim}, \sigma_{obs}\}}$$

D.6. Root mean square error (%)

$$\%\text{RMS} = \frac{\sqrt{\frac{\sum_{i=1}^{N}(S_i - O_i)^2}{N}}}{\overline{O}} \times 100$$

D.7. The following aggregate statistics were generated for selected individual events

a. Percent absolute event runoff error, E_r, %

This is the absolute value of the runoff bias from several events expressed as a percentage:

$$E_r = \frac{\sum_{i=1}^{N} |B_i|}{NY_{avg}} \cdot 100$$

b. Percent absolute peak error, E_p, %

This is the absolute value of error in peak discharge for several events expressed as a percentage:

$$E_p = \frac{\sum_{i=1}^{N} |Q_{pi} - Q_{psi}|}{NQ_{pavg}} \cdot 100$$

c. Percent absolute peak time error, E_t, h

This is the absolute value of the error in peak time for several events expressed as a percentage:

$$E_t = \frac{\sum_{i=1}^{N} |T_{\text{pi}} - T_{\text{psi}}|}{N} \cdot 100$$

where B_i is the runoff bias per ith flood event, mm; Y_{avg} the average observed flood event runoff, mm; $Q_{p,i}$ the observed peak discharge of the ith flood event, $m^3\,s^{-1}$; $Q_{ps,i}$ the simulated peak discharge of the ith flood event, $m^3\,s^{-1}$; $Q_{p,avg}$ the average observed peak discharge, $m^3\,s^{-1}$; $T_{p,i}$ the observed time to the ith peak, h; $T_{ps,i}$ the simulated time to the ith peak, h, and N the number of selected events.

D.8. Statistics to measure improvement over the LMP benchmark

a. Flood runoff improvement I_y, %

This statistic measures the improvement in computed runoff volume:

$$I_y = \frac{\sum_{i=1}^{N}(|Y_i - Y_{s,i}| - |Y_i - Y_{z,i}|)}{N \cdot Y_{avg}} \cdot 100$$

b. Peak flow improvement I_p, %

This statistic quantifies the gain in simulating the peak event discharge:

$$I_p = \frac{\sum_{i=1}^{N}(|Q_{p,i} - Q_{ps,i}| - |Q_{p,i} - Q_{pz,i}|)}{N \cdot Q_{p,avg}} \cdot 100$$

c. Peak time improvement I_t

This statistic measures the improvement in simulated peak time:

$$I_t = \frac{\sum_{i=1}^{N}(|T_{p,i} - T_{ps,i}| - |T_{p,i} - T_{pz,i}|)}{N}$$

where Y_i is the observed runoff volume of the ith flood, mm; $Y_{s,i}$ the (distributed model) simulated runoff volume of the ith event, mm; $Y_{z,i}$ the (lumped model) simulated runoff of the ith flood to compare with, mm; Y_{avg} the average observed flood event runoff volume of N events, mm; $Q_{p,i}$ the observed peak discharge of the ith event, $m^3\,s^{-1}$; $Q_{ps,i}$ the (distributed model) simulated peak discharge of the ith event, $m^3\,s^{-1}$; $Q_{pz,i}$ the (lumped model) simulated peak discharge, $m^3\,s^{-1}$; $Q_{p,avg}$ the average observed peak discharge of N events, $m^3\,s^{-1}$; $T_{p,i}$ the observed time of the ith peak, h; $T_{ps,i}$ the (distributed model) simulated time of the ith peak, h; $T_{pz,i}$ the (lumped model) simulated time to ith peak, h and N is the number of selected events.

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., 1976. National Weather Service River Forecast System, Snow Accumulation and Ablation Model, NOAA Technical Memorandum NWS HYDRO-17.

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, E.A., 2006. Snow Accumulation and Ablation Model –SNOW-17. <http://www.nws.noaa.gov/oh/hrl/nwsrfs/users_manual/part2/_pdf/22snow17.pdf>.

Andréassian, V., Hall, A., Chahinian, N., Schaake, J., 2006. Introduction and synthesis: Why should hydrologists work on a large number of basin data sets?, vol 307. IAHS Publ., pp. 1–5.

Andréassian, V., Perrin, C., Berthet, L., Le Moine, N., Lerat, J., Loumagne, C., Oudin, L., Mathevet, T., Ramos, H.M., Valéry, A., 2009. Crash tests for a standardized evaluation of hydrological models. Hydrol. Earth Syst. Sci. 13, 1757–1764. http://dx.doi.org/10.5194/hess-13-1757-2009.

Bales, R.C., Molotch, N.P., painter, T.H., Dettinger, M.D., Rice, R., Dozier, J., 2006. Mountain hydrology of the western United States. Water Resour. Res. 42, W08432. http://dx.doi.org/10.1029/2005WR004387.

Behrangi, A., Hsu, K.-L., Imam, B., Sorooshian, S., Huffman, G.J., Kuligowski, R.J., 2009. PERSIANN-MSA: a precipitation estimation method from satellite-based multispectral analysis. J. Hydrometerol. 10, 1414–1429. http://dx.doi.org/10.1175/2009JHM1139.1.

Benjamin, S.G., Devenyu, D., Weygandt, S.S., Brundage, K.J., Brown, J.M., Grell, G.A., Kim, D., Schwartz, B.E., Smirnova, T.G., Smith, T.L., Manikin, G.S., 2004. An hourly assimilation-forecast cycle: The RUC. Mon. Weather Rev. 132, 495–518. http://dx.doi.org/10.1175/1520-0493(2004)132%3C0495:AHACTR%3E2.0.CO;2.

Bennett, N.D., Croke, B.F.W., Guariso, G., Guillaume, J.H.A., Hamilton, S.H., Jakeman, A.J., Marsili-Libelli, S., Newham, L.T.H., Norton, J.P., Perrin, C., Pierce, S.A., Robson, B., Seppelt, R., Voinov, A.A., Fath, B.D., Andréassian, V., 2013. Characterizing performance of environmental models. Environ. Model. Softw. 40, 1–20. http://dx.doi.org/10.1016/j.envsoft.2012.09.011.

Blöschl, G., 1999. Scaling issues in snow hydrology. Hydrol. Process. 13, 2149–2175, 10.1002/(SICI)1099-1085(199910)13:14/15<2149::AID-HYP847>3.0.CO;2-8.

Boushaki, F.I., Hsu, K.-L., Sorooshian, S., Park, G.-H., 2009. Bias adjustment of satellite precipitation estimation using ground-based measurement: a case study evaluation over the southwestern United States. J. Hydrometeorol. 10, 1231–1242. http://dx.doi.org/10.1175/2009JHM1099.1.

Braun, L.N., Brun, E., Durand, Y., Martin, E., Tourasse, P., 1994. Simulation of discharge using different methods of meteorological data distribution, basin discretization, and snow modeling. Nord. Hydrol. 25 (1-2), 129–144. http://dx.doi.org/10.2166/nh.1994.009.

Budyko, M.I., 1974. Climate and Life. Academic Press, New York.

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.

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 simulation. J. Hydrol. 298, 242–266. http://dx.doi.org/10.1016/j.jhydrol.2004.03.042.

Carpenter, T.M., Georgakakos, K.P., 2001. Assessment of Folsom lake response to historical and potential future climate scenarios: 1. Forecasting. J. Hydrol. 249, 148–175.
Carroll, T., Cline, D., Fall, G., Nilsson, A., Li, L., Rost, A., 2001. NOHRSC Operations and the Simulation of Snow Cover Properties of the Coterminous U.S. Proceedings of the 69th Annual Meeting of the Western Snow Conference, Sun Valley, ID, Western Snow Conference, pp. 1–14.
Cayan, D.R., Riddle, L.G., 1993. The influence of precipitation and temperature on seasonal streamflow in California. Water Resour. Res. 29 (4), 1127–1140.
Clark, M.P., McMillan, H.K., Collins, D.B.G., Kavetski, D., Woods, R.A., 2011. Hydrological field data from a modeller's perspective: Part 2: process-based evaluation of model hypotheses. Hydrol. Process. 25, 523–543. http://dx.doi.org/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. http://dx.doi.org/10.1016/j.jhydrol.2008.01.026.
Cline, D., Elder, K., Bales, R., 1998. Scale effects in a distributed snow water equivalence and snowmelt model for mountain basins. Hydrol. Process. 12 (10–11), 1527–1536.
Coccia, G., Mazzetti, C., Ortiz, E.A., Todini, E., 2009. Application of the Topkapi model within the DMIP 2 project. In: Paper 4.2, 23rd Conference on Hydrology, 89th Meeting of the AMS, Phoenix, Arizona, January 10–16. <https://ams.confex.com/ams/89annual/techprogram/paper_149632.htm>.
Confalonieri, R., Bregaglio, S., Bocchi, S., Acutis, M., 2010. An integrated procedure to evaluate hydrological models. Hydrol. Process. 24, 2762–2770. http://dx.doi.org/10.1002/hyp.7683.
Daly, C., 2006. Guidelines for assessing the suitability of spatial climate data sets. Int. J. Climatol. 26, 707–721. http://dx.doi.org/10.1002/joc.1322.
Daly, C., Neilson, R.P., Phillips, D.L., 1994. A statistical-topographic model for mapping climatological precipitation over mountainous terrain. J. Appl. Meteorol. 33, 140–158.
Daly, C., Halbleib, M., Smith, J.I., Gibson, W.P., Dogget, M.K., Taylor, G.H., Curtis, J., Pasteris, P.P., 2008. Physiographically sensitive mapping of climatological temperature and precipitation across the conterminous United States. Int. J. Climatol.. http://dx.doi.org/10.1002/joc.1688.
De Pondeca, M.S.F.V., Manikin, G.S., DiMego, G., Benjamin, S.G., et al., 2011. The real-time mesoscale analysis at NOAA's National Centers for Environmental Prediction: current status and development. Weather Forecast. 26, 593–612. http://dx.doi.org/10.1175/WAF-D-10-05037.1.
Debele, B., Srinivasan, R., Gosain, A.K., 2009. Comparison of process-based and temperature-index snowmelt modeling in SWAT. Water Resour. Manage, 10.1007/s11269-009-9486-2.
Dettinger, M., Redmond, K., Cayan, D., 2004. Winter orographic precipitation ratios in the Sierra Nevada—large-scale atmospheric circulations and hydrologic consequences. J. Hydrometeorol. 5, 1102–1116.
Dong, X., Dohmen-Hansssen, M., Booij, M.J., 2005. Appropriate spatial sampling of rainfall for flow simulation. Hydrol. Sci. J. 50 (2), 279–298.
Dornes, P.F., Pomeroy, J.W., Pietroniro, A., Carey, S.K., Qinton, W.L., 2008. Influence of landscape aggregation in modeling snow-cover ablation and snowmelt runoff in a sub-arctic mountainous environment. Hydrol. Sci. J. 53 (3), 725–740.
Dressler, K.A., Fassnacht, S.R., Bales, R.C., 2006. A comparison of snow telemetry and snow course measurements in the Colorado River Basin. J. Hydrometeorol. 7, 705–712.
Duan, Q., Gupta, V.K., Sorooshian, S., 1993. A shuffled complex evolution approach for effective and efficient global minimization. J. Optim. Theory Appl. 76, 501–521.
Duan, Q., Sorooshian, S., Gupta, V.K., 1994. Optimal use of the SCE-UA global optimization method for calibrating watershed models. J. Hydrol. 158, 265–284.
Duan, Q.Y., Schaake, J.C., Koren, V.I., 1996. FIFE 1987 water budget analysis. J. Geophys. Res. 101 (D3), 7197–7207.
Duan, Q.Y., Schaake, J.S., 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. http://dx.doi.org/10.1016/j.jhydrol.2005.07.031.
Eischeid, J.K., Pasteris, P.A., Diaz, H.F., Plantico, M.S., Lott, N.J., 2000. Creating a serially complete, national daily time series of temperature and precipitation for the western United States. J. Appl. Meteorol. 39, 1580–1591.
Essery, R., Martin, E., Douville, H., Fernanndez, A., Brun, E., 1999. A comparison of four snow models using observations from an alpine site. Clim. Dyn. 15, 583–593.
Etchevers, P., Martin, E., Brown, R., Fierz, C., et al., 2004. Validation of the energy budget of an alpine snowpack simulated by several snow models (SnowMIP project). Ann. Glaciol. 38 (1), 140–158. http://dx.doi.org/10.3189/172756404781814825.
Ewen, J., 2011. Hydrograph matching method for measuring model performance. J. Hydrol. 408, 178–187. http://dx.doi.org/10.1016/j.jhydrol.2011.07.038.
Farnsworth, R.K., Thompson, E.S., Peck, E.L., 1982. Evaporation Atlas for the Contiguous 48 United States. NOAA Technical Report NWS 33, 27pp.
Feng, X., Sahoo, A., Arsenault, K., Houser, P., Luo, Y., Troy, T.J., 2008. The impact of snow model complexity at three CLPX sites. J. Hydrometeorol. 9, 1464–1481. http://dx.doi.org/10.1175/2008JHM860.1.
Fierz, C., Riber, P., Adams, E.E., Curran, A.R., Föhn, P.M.B., Lehning, M., Pluss, C., 2003. Evaluation of snow-surface energy balance models in alpine terrain. J. Hydrol. 282, 76–94. http://dx.doi.org/10.1016/S0022-1694(03)00255-5.
Francés, F., Vélez, J.I., Vélez, J.J., 2007. Split-parameter structure for the automatic calibration of distributed hydrological models. J. Hydrol. 332 (1), 226–240.
Franz, K.J., Hogue, T.S., Sorooshian, S., 2008a. Operational snow modelling: addressing the challenges of an energy balance model for National Weather Service forecasts. J. Hydrol. 360, 48–66.
Franz, K.J., Hogue, T.S., Sorooshian, S., 2008b. Snow model verification using ensemble prediction and operational benchmarks. J. Hydrometeorol. 9, 1402–1415. http://dx.doi.org/10.1175/2008JHM995.1.
Garen, D.C., Marks, D., 2005. Spatially distributed energy balance snowmelt modelling in a mountainous river basin: estimation of meteorological inputs and verification of model results. J. Hydrol. 315, 126–153.
Georgakakos, A.P., Yao, H., Georgakakos, K.P., 2010. Upstream Regulation Adjustments to Ensemble Streamflow Predictions. HRC Technical Report No. 7; Sponsored by NOAA/NWS/OHD, Award No. NA08NWS4620023.
Gourley, J.J., Jorgensen, D.P., Matrosov, S.Y., Flamig, Z.L., 2009. Evaluation of incremental improvements to quantitative precipitation estimates in complex terrain. J. Hydrometeorol. 10, 1507–1520.
Green, D.R., Hudlow, M.D., 1982. Hydrometeorologic grid mapping procedures. In: AWRA International Symposium on Hydrometeorology, Denver, CO., Available on Request from NWS/OHD.
Guan, B., Molotch, N.P., Waliser, D.E., Fetzer, E.J., Neiman, P.J., 2010. Extreme snowfall events linked to atmospheric rivers and surface air temperature via satellite measurements. Geophys. Res. Lett. 37, L20401. http://dx.doi.org/10.1029/2010GL044696.
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. http://dx.doi.org/10.1175/2010JAMC2484.1.
Guirguis, K.J., Avissar, R., 2008. An analysis of precipitation variability, persistence, and observational data uncertainty in the Western United States. J. Hydrometeorol. 9, 843–865. http://dx.doi.org/10.1175/2008JHM972.1.
Gupta, H.V., Wagener, T., Lui, Y., 2008. Reconciling theory with observations: elements of a diagnostic approach to model evaluation. Hydrol. Process. 22, 3802–3813. http://dx.doi.org/10.1002/hyp.6989.
Hansen, M.C., Defries, R.S., Townshend, J.R.G., Sohlbert, R., 2000. Global land cover classification at 1 km spatial resolution using a classification tree approach. Int. J. Rem. Sens. 21 (6&7), 1331–1364.
Hartman, R.K., 2010. Hydrologic forecasting in mountainous terrain. In: Presentation H21H-05, Presented at 2010 Fall Meeting, AGU, San Francisco, Calif., 13–17 December.
Hay, L.E., Clark, M.P., 2003. Use of statistically and dynamically downscaled atmospheric model output for hydrologic simulations in three mountainous basins in the western United States. J. Hydrol. 282, 56–75.
Hay, L, Viger, R., McCabe, G., 1998. Precipitation interpolation in mountainous regions using multiple linear regression. In: Preceedings '98 Conference on Hydrology, Water Resources, and Ecology in Headwaters, IAHS Publication 248, 33–38.
Hogue, T.S., Sorooshian, S., Gupta, H., Holz, A., Braatz, D., 2000. Multi-step automatic calibration scheme for river forecasting models. J. Hydrometerol. 1, 524–542.
Jankov, I., Schultz, P.J., Anderson, C.J., Koch, S.E., 2007. The impact of different physical parameterizations and their interactions on cold season QPF in the American River Basin. J. Hydrometeorol. 8, 1141–1151. http://dx.doi.org/10.1175/JHM630.1.
Jenson, S.K., Domingue, J.O., 1988. Extracting topographic structure from digital elevation data for geographic information system analysis. Photogramm. Eng. Rem. Sens. 54, 1593–1600.
Jeton, A.E., Smith, J.L., 1993. Development of watershed models for two Sierra Nevada basins using a geographic information system. Water Resour. Bull. 29 (6), 923–932.
Jeton, A.E., Dettinger, M.D., Smith, J.L., 1996. Potential Effects of Climate Change on Streamflow, Eastern and Western Slopes of the Sierra Nevada, California and Nevada. U.S. Geological Survey Water Resources Investigations Report 95-4260, 44pp.
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.
Kabeche, F., Figueras, J., Fradon, B., Hogan, R., Boumahmoud, A.-A., Illingworth, A., Tabary, P., 2010. Towards X-band polarimetric quantitative precipitation estimation in mountainous regions: the RHYTMME project. In: ERAD 2010 – the Sixth European Conference on Radar in Meteorology and Hydrology, Sibiu Romania, 6–10 September.
Kampf, S.K., Burges, S.J., 2007. A framework for classifying and comparing distributed hillslope and catchment hydrologic models. Water Resour. Res. 43, W05423. http://dx.doi.org/10.1029/2006WR005370.
Kapnick, S., Hall, A., 2010. Observed climate-snowpack relationships in California and their implications for the future. J. Clim. 23, 3446–3456.
Khakbaz, B., Imam, B., Hsu, K., Sorooshian, S., 2012. From lumped to distributed: calibration strategies for semi-distributed hydrologic models. J. Hydrol. 418–419, 61–77. http://dx.doi.org/10.1016/j.jhydrol.2009.02.021.
Kienzle, S., 2008. A new temperature based method to separate rain and snow. Hydrol. Process. 22, 5067–5085. http://dx.doi.org/10.1002/hyp.7131.
Kim, J., Miller, N.L., Guetter, A.K., Georgakakos, K.P., 1998. River flow response to precipitation and snow budget in California during the 1994/1995 winter. J. Clim. 11, 2376–2386.
Kingsmill, D.E., White, A.B., Neiman, P.J., Ralph, F.M., 2006. Synoptic and topographic variability of Northern California precipitation characteristics in land-falling winter storms observed during CALJET. Mon. Weather Rev. 134, 2072–2094. http://dx.doi.org/10.1175/MWR3166.1.

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. http://dx.doi.org/10.1029/2005WR004362.

Koren, V., Schaake, J., Duan, Q., Smith, M., Cong, S., 1998. PET Upgrades to NWSRFS. Internal NWS/OHD report. Available at: <http://amazon.nws.noaa.gov/ohd/publications/publications_google.php>.

Koren, V., Schaake, J., Mitchell, K., Duan, Q.-Y., Chen, F., Baker, J.M., 1999. A parameterization of snowpack and frozen ground intended for NCEP weather and climate models. J. Geophys. Res. 104 (D16), 19569–19585.

Koren, V., Reed, S., Smith, M., Zhang, Z., Seo, D.-J., 2004. Hydrology laboratory research modeling system (HL-RMS) of the U.S. 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. In: Predictions in Ungaged Basins: Promise and Progress (Proceedings of Symposium S7 Held During the Seventh IAHS Scientific Assembly at Foz do Iguacu, Brazil, April 2005. IAHS Publication 303, pp. 47–56.

Krause, P., Boyle, D.P., Bäse, F., 2005. Comparison of different efficiency criteria for hydrological model assessment. Adv. Geosci. 5, 89–97. http://dx.doi.org/10.5194/adgeo-5-89-2005.

Kuligowski, R.J., 2002. A self-calibrating real-time GOES rainfall algorithm for short-term rainfall estimates. J. Hydrometeorol. 3, 112–130. http://dx.doi.org/10.1175/1525-7541(2002) 003<0112:ASCRTG>2.0.CO;2.

Kuligowski, R.J., Li, Y., Zhang, Y., 2013. Impact of TRMM data on a low-latency, high-resolution precipitation algorithm for flash-flood forecasting. J. Appl. Met. Clim. 52, 1379–1393. http://dx.doi.org/10.1175/JAMC-D-12-0107.1.

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.

Le Moine, N., 2008. The Surface Catchment Seen from Underground: A Way to Improve the Performance and Realism of Rainfall-runoff Models? PhD Thesis, University Pierre et Marie Curie (Paris), CEMAGREF (Antony), 324pp. <http://webgr.irstea.fr/wp-content/uploads/2012/07/2008-LE_MOINE-THESE.pdf>.

Leavesley, G.H., Hay, L.E., Viger, R.J., Marskstrom, S.L., 2003. Use of *a priori* parameter-estimation methods to constrain calibration of distributed-parameter models. In: Duan et al. (Eds.), Calibration of Watershed Models. Water Science and Application. AGU Press, pp. 255–266.

Lei, F., Koren, V., Smith, M., Moreda, F., 2007. A sensitivity study of an energy-budget snow accumulation and ablation model. In: Proceedings of the Paper J6.4, 87th Annual Meeting of the AMS, 21st Conference on Hydrology, 13–18 January.

Lettenmaier, D.P., Gan, T.Y., 1990. Hydrologic sensitivities of the Sacramento-San Joaquin River basin, California, to global warming. Water Resour. Res. 26 (1), 69–86.

Leydecker, A., Sickman, J.O., Melack, J.M., 2001. Spatial scaling of the hydrological and biochemical aspects of high-altitude catchments in the Sierra Nevada, California, USA. Arct. Antarct. Alp. Res. 33 (4), 391–396.

Liston, G.E., Hiemstra, C.A., Elder, K., Cline, D.W., 2008. Mesocell study area snow distributions for the cold land processes experiment (CLPX). J. Hydrometeorol. 9, 957–976. http://dx.doi.org/10.1175/2008JHM869.1.

Liu, Z., Todini, E., 2001. Towards a comprehensive physically-based rainfall-runoff model. Hydrol. Earth Syst. Sci. 6, 859–881.

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., Mitchell, K.E., Houser, P.R., Wood, E.F., Schaake, J.C., Robock, A., Cosgrove, B.A., Sheffield, J., Duan, Q.-Y., Luo, L., Higgins, R.W., Pinker, R.T., Tarpley, J.D., 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. http://dx.doi.org/10.1029/2003JD003517.

Lundquist, J.D., Cayan, D.R., 2002. Seasonal and spatial patterns in diurnal cycles in streamflow in the western United States. J. Hydrometeorol. 3, 591–603.

Lundquist, J.D., Loheide, S.P., 2011. How evaporative water losses vary between wet and dry water years as a function of elevation in the Sierra Nevada, California, and critical factors for modeling. Water Resour. Res. 47, W00H09. http://dx.doi.org/10.1029/2010WR010050.

Lundquist, J.D., Neiman, P.J., Martner, B., White, A.B., Gottas, D.J., Ralph, F.M., 2008. Rain versus snow in the Sierra Nevada, California: Comparing Doppler profiling radar and surface observations of melting level. J. Hydrometeorol. 9, 194–211.

Mathevet, T., 2005. Which Lumped Rainfall-runoff Models at the Hourly Time Step? Empirical Development and Intercomparison of Models on a Large Set of Catchments. PhD Thesis, ENGREF (Paris), Cemagref (Antony), France, 463pp. <http://webgr.irstea.fr/wp-content/uploads/2012/07/2005-MATHEVET-THESE.pdf>.

Matrosov, S.Y., Clark, K.A., Kingsmill, D.E., 2007. A polarimetric radar approach to identify rain, melting-layer, and snow regions for applying corrections to vertical profiles of reflectivity. J. Appl. Meteorol. Climatol. 46, 154–166. http://dx.doi.org/10.1175/JAM2508.1.

Maurer, E.P., Mass, C., 2006. Using radar data to partition precipitation into rain and snow in a hydrologic model. J. Hydrol. Eng. 11 (3), 214–221.

McCabe, G.J., Clark, M.P., Hay, L.E., 2007. Rain-on-snow events in the western United States. Bull. Am. Meteorol. Soc., 319–328. http://dx.doi.org/10.1175/BAMS-88-3-319.

McCuen, R.H., Snyder, W.M., 1975. A proposed index for comparing hydrographs. Water Resour. Res. 11 (6), 1021–1024.

McDonnell, J.J., Sivapalan, M., Vache, K., Dunn, S., Grant, G., Hatterty, R., Hinz, C., Hooper, R., Kirchner, J., Roderick, M.L., Selker, J., Weiler, M., 2007. Moving beyond heterogeneity and process complexity: a new vision for watershed hydrology. Water Resour. Res. 43, W07301. http://dx.doi.org/10.1029/2006WR005467.

Merz, R., Parajka, J., Blöschl, G., 2009. Scale effects in conceptual hydrological modeling. Water Resour. Res. 45, W09405. http://dx.doi.org/10.1029/2009WR007872.

Mesinger, F., DiMego, G., Kalnay, E., Mitchell, K., Shafran, P.C., Ebisuzaki, W., Jovic, D., Woollen, J., Rogers, E., Berbery, E.H., Ek, M.B., Fan, Y., Grumbine, R., Higgins, W., Li, H., Lin, Y., Manikin, G., Parrish, D., Shi, W., 2006. North American Regional Reanalysis (NARR). Bull. Am. Meteorol. Soc. 87 (3), 343–360.

Miller, D.A., White, R.A., 1998. A conterminous United States multi-layer soil characteristics data set for regional climate and hydrology modeling. Earth Interact., 2, Paper No. 2. <http://EarthInteractions.org>.

Minder, J.R., Kingsmill, D.E., 2013. Mesoscale variations of the atmospheric snow-line over the northern Sierra Nevada: multi-year statistics, case study, and mechanisms. J. Atmos. Sci. 70 (3), 916–938. http://dx.doi.org/10.1175/JAS-D-12-0194.1.

Minder, J.R., Durran, D.R., Roe, G.H., 2011. Mesoscale controls on the mountainside snow line. J. Atmos. Sci. 68, 2107–2127. http://dx.doi.org/10.1175/JAS-D-10-05006.1.

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. http://dx.doi.org/10.1029/2003JD003823.

Mizukami, N., Koren, V., 2008. Methodology and evaluation of melt factor parameterization for distributed SNOW-17. In: Proceedings of the American Geophysical Union, Fall Meeting 2008, Abstract #H31J-08.

Mizukami, N., Smith, M., 2012. Analysis of inconsistencies in multi-year high space-time resolution precipitation grids over complex terrain and impacts on hydrologic modeling. J. Hydrol. 428–429, 129–141. http://dx.doi.org/10.1016/j.jhydrol.2012.01.030.

Mizukami, N., Koren, V., Smith, M., Kingsmill, D., Zhang, Z., Cosgrove, B., Cui, Z., 2013. The impact of precipitation type discrimination on hydrologic simulation: rain–snow partitioning derived from HMT-West radar-retected brightband height versus surface temperature data. J. Hydrometeorol 14, 1139–1158. http://dx.doi.org/10.1175/JHM-D-12-035.1.

Molotch, N.P., Bales, R.C., 2006. SNOTEL representativeness in the Rio Grande headwaters on the basis of physiographics and remotely sensed snow cover persistence. Hydrol. Process. 20, 723–739. http://dx.doi.org/10.1002/hyp.6128.

Moreda, F., Cong, S., Schaake, J., Smith, M., 2006. Gridded rainfall estimation for distributed modeling in western mountainous areas. In: Poster H23A, Spring Meeting of the AGU, May 23–27, Baltimore, MD.

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., Vieth, T.L., 2007. Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Trans. ASABE 50 (3), 885–900.

Nelson, B.R., Seo, D.-J., Kim, D., 2010. Multisensor precipitation reanalysis. J. Hydrometeorol. 11, 666–682. http://dx.doi.org/10.1175/2010JHM1210.1.

National Research Council (NRC), 1998. Future of the National Weather Service Cooperative Observer Network. National Weather Service Modernization Committee, Commission on Engineering and Technical Systems, National Academy Press, 73pp. <http://www.nap.edu/catalog/6197.html>.

Pan, M. et al., 2003. Snow process modeling in the North American Land Data Assimilation System (NLDAS): 2. Evaluation of model-simulated snow water equivalent. J. Geophys. Res. 108 (D22), 8850. http://dx.doi.org/10.1029/2003JD003994.

Pandey, G.R., Cayan, D.R., Georgakakos, K.P., 1999. Precipitation structure in the Sierra Nevada of California during winter. J. Geophys. Res. 104 (D10), 12019–12030.

Pandey, G.R., Cayan, D.R., Dettinger, M.D., Georgakakos, K.P., 2000. A hybrid orographic plus statistical model for downscaling daily precipitation in Northern California. J. Hydrometeorol. 1 (December), 491–506.

Parton, W.J., Logan, J.A., 1981. A model for diurnal variation in soil and air temperature. Agric. Meteorol. 23, 205–216.

Perrin, C., Michel, C., Andréassian, V., 2003. Improvement of a parsimonious model for streamflow simulation. J. Hydrol. 279, 275–289. http://dx.doi.org/10.1016/S0022-1694%2803%2900225-7.

Perrin, C., Andréassian, V., Michel, C., 2006. Simple benchmark models as a basis for criteria of model efficiency. Archiv für Hydrobiologie 17 (1–2), 221–244. http://dx.doi.org/10.1127/lr/17/2006/221 (Supplement 161/1-2, Large Rivers).

Puspalatha, R., Perrin, C., Le Moine, N., Andréassian, V., 2012. A review of efficiency criteria suitable for evaluating low-flow simulations. J. Hydrol. 420–421, 171–182. http://dx.doi.org/10.1016/j.jhydrol.2011.11.055.

Ralph, F.M., Dettinger, M.D., 2011. Storms, floods, and the science of atmospheric rivers. Eos Trans. Amer. Geophys. Union 92 (32), 265–272.

Ralph, F.M. et al., 2005. Improving short-term (0–48 h) cool-season quantitative precipitation forecasting: recommendations from a USWRP workshop. Bull. Am. Meteorol. Soc. 86, 1619–1632.

Reed, S., 2003. Deriving flow directions for coarse resolution (1–4 km) gridded hydrologic modeling. Water Resour. Res. 39 (9), 4-1–4-11.

Reed, S.M., Maidment, D.R., 1999. Coordinate transformations for using NEXRAD data in GIS-based hydrologic modeling. J. Hydrol. Eng. 4 (174). http://dx.doi.org/10.1061/(ASCE)1084-0699(1999)4:2(174).
Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., Participants, D.M.I.P., 2004. Overall distributed model intercomparison project results. J. Hydrol. 298 (1–4), 27–60.
Reek, T., Doty, S.R., Owen, T.W., 1992. A deterministic approach to the validation of historical daily temperature and precipitation data from the cooperative network. Bull. Am. Meteorol. Soc. 73 (6), 753–762. http://dx.doi.org/10.1175/1520-0477(1992)073<0753:ADATTV>2.0.CO;2.
Reynolds, D.W., Dennis, A.S., 1986. A review of the Sierra Cooperative Pilot Project. Bull. Am. Meteorol. Soc. 67 (5), 513–523.
Rice, R., Bales, R.C., 2010. Embedded-sensor network design for snow cover measurements around snow pillow and snow course sites in the Sierra Nevada of California. Water Resour. Res. 46, W03537. http://dx.doi.org/10.1029/2008WR007318.
Ritter, A., Muñoz-Carpena, R., 2013. Performance evaluation of hydrological models: statistical significance for reducing subjectivity in goodness-of-fit assessments. J. Hydrol. 480, 33–34. http://dx.doi.org/10.1016/j.jhydrol.2012.12.004.
Rutter, N., Cline, D., Long, L., 2008. Evaluation of the NOHRSC Snow Model (NSM) in a one-dimensional mode. J. Hydrometeorol. 9, 695–711.
Rutter, N., Essery, R., Pomeroy, J., Koren, V., et al., 2009. Evaluation of forest snow processes models (SnowMIP2). J. Geophys. Res. 114, 1–5. http://dx.doi.org/10.1029/2008JD011063.
Schaake, J.C., Cong, S., Duan, Q., 2006. The U.S. MOPEX Data Set, vol. 307. IAHS Publication, pp. 9–28.
Schaefli, B., Gupta, H.V., 2007. Do Nash values have value? Hydrol. Process. 21, 2075–2080. http://dx.doi.org/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.
Seibert, J., 2001. On the need for benchmarks in hydrological modeling. Hydrol. Process. 15 (6), 1063–1064.
Seo, D.-J., 1998. Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions. J. Hydrol. 208, 25–36.
Serreze, M.C., Clark, M.P., Armstrong, R.L., McGinnis, D.A., Pulwarty, R.S., 1999. Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. Water Resour. Res. 35 (7), 2145–2160.
Shamir, E., Georgakakos, K.P., 2006. Distributed snow accumulation and ablation modeling in the American River Basin. Adv. Water Resour. 29, 558–570.
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.
Simpson, J.J., Dettinger, M.D., Gehrke, F., McIntire, T.J., Huffurd, G.L., 2004. Hydrologic scales, cloud variability, remote sensing, and models: implications for forecasting snowmelt and streamflow. Weather Forecast. 19 (April), 251–276.
Slater, A.G., Schlosser, C.A., et al., 2001. The representation of snow in land surface schemes: results from PILPS 2(d). J. Hydrometeorol. 2, 7–25.
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 et al. (Eds.), Calibration of Watershed Models: Water Science and Applications. 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., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. J. Hydrol. 298 (1–4), 4–26.
Smith, M.B., Koren,V., Reed, S., Zhang, Z., Seo, D.-J., Moreda, F., Cui, Z., 2006. The Distributed Model Intercomparison Project: Phase 2 Science Plan. <http://www.nws.noaa.gov/oh/hrl/dmip/2/docs/dmip_2_plan_march10_06_update.pdf>.
Smith, M., Koren, V., Reed, S., Zhang, Z., Cui, Z., Mizukami, N., Moreda, F., Sheldon, S., 2009. Preliminary results of phase 2 of the distributed model intercomparison project. In: Presentation 4.1, 23rd Conference on Hydrology, 89th Meeting of the AMS, Phoenix, Arizona, January 10–15. <https://ams.confex.com/ams/89annual/techprogram/paper_148995.htm>.
Smith, M., Ding, F., Cui, Z., Koren, V., Mizukami, N., Zhang, Z., Cosgrove, B., Kitzmiller, D., Schaake, J., 2010. Development of gridded QPE datasets for mountainous area distributed hydrologic modeling. In: Presented at the ASCE EWRI Conference, Rhode Island, May 2010.
Smith, M.B., Koren, V., Reed, S., Zhang, Z., Zhang, Yu., Moreda, F., Cui, Z., Mizukami, N., Anderson, E.A., Cosgrove, B.A., 2012a. The distributed model intercomparison project – Phase 2: motivation and design of the Oklahoma experiments. J. Hydrol. 418–419, 3–16. http://dx.doi.org/10.1016/j.jhydrol.2011.08.055.
Smith, M., Koren, V., Zhang, Z., Zhang, Y., Reed, S., Cui, Z., Moreda, F., Anderson, E., Cosgrove, B., Mizukami, N., DMIP 2 Participants, 2012b. Overall results of the DMIP 2 Oklahoma experiments. J. Hydrol. 418–419, 17–48. http://dx.doi.org/10.1016/j.jhydrol.2011.08.056.
Timbal, B., Henderson-Sellers, A., 1998. Intercomparisons of land-surface parameterizations coupled to a limited area forecast model. Global Planet. Change 19, 247–260.
Todini, E., Ciarapica L., 2002. The TOPKAPI model. In: Singh, V.P., Frevert, D.K. (Eds.), Mathematical Models of Large Watershed Hydrology. Water Resources (chapter 12).
Tsintikidis, D., Georgakakos, K.P., Sperfslage, J.A., Smith, D.E., Carpenter, T.M., 2002. Precipitation uncertainty and rain gage network design within the Folsom Lake Watershed. ASCE J. Hydrol. Eng. 7 (2), 175–184.
Valéry, A., 2010. Modélisation précipitations – débit sous influence nivale. Élaboration d'un module neige et évaluation sur 380 bassins versants. Thèse de Doctorat, Cemagref (Antony), AgroParisTech (Paris), 405pp. <http://webgr.irstea.fr/wp-content/uploads/2012/07/2010-VALERY-THESE.pdf>.
Valéry, A., Andréassian, V., Perrin, C., 2009. Inverting the hydrologic cycle: When streamflow measurements help assess altitudinal precipitation gradients in mountain areas. IAHS-AISH Publication 333, pp. 281–286.
Vélez, J.J., Puricelli, M., López-Unzu, F., Francés, F., 2009. Parameter extrapolation to ungaged basins with a hydrological distributed model in a regional framework. Hydrol. Earth Syst. Sci. 13 (2), 229–246.
Viney, N.R., Bormann, H., Breuer, L., Bronstert, A., Croke, B.F.W., Frede, H., et al., 2009. Assessing the impact of land use change on hydrology by ensemble modeling (LUCHEM) II: Ensemble combinations and predictions. Advances in Water Resources 32 (2), 147–158. http://dx.doi.org/10.1016/j.advwatres.2008.05.006.
Waichler, S.R., Wigmosta, M.S., 2003. Development of hourly meteorological values from daily data and significance to hydrologic modeling at H.J. Andrews Experimental Forest. J. Hydrometeorol. 4, 251–263.
Walter, M.T., Brooks, E.S., McCool, D.K., King, L.G., Molnau, M., Boll, J., 2005. Process-based snowmelt modeling: does it require more input data than temperature-index modeling? J. Hydrol. 300, 65–75. http://dx.doi.org/10.1016/j.jhydrol.2004.05.002.
Wang, J., Georgakakos, K.P., 2005. Validation and sensitivities of dynamic precipitation simulation of winter events over the Folsom Lake Watershed: 1964–1999. Mon. Weather Rev. 133, 3–19.
Wayand, N., Hamlet, A., Hughes, M., Feld, S., Lundquist, J., 2013. Intercomparison of meteorological forcing data from empirical and mesoscale model sources in the N.F. American River Basin in northern Sierra Nevada, California. J. Hydrometeorol. 14, 677–699. http://dx.doi.org/10.1175/JHM-D-12-0102.1.
Westrick, K.J., Mass, C.J., 2001. An evaluation of a high-resolution hydrometeorological modeling system for prediction of a cool-season flood event in a coastal mountainous watershed. J. Hydrometeorol. 2, 161–180.
Westrick, K.J., Mass, C.F., Colle, B.A., 1999. The limitations of the WSR-88D radar network for quantitative precipitation measurement over the coastal Western United States. Bull. Am. Meteorol. Soc. 80 (11), 2289–2298.
Westrick, K.J., Storck, Pl., Mass, C.J., 2002. Description and evaluation of a hydrometeorological forecast system of mountainous watersheds. Weather Forecast. 17, 250–262.
White, A.B., Gottas, D.J., Strem, E.T., Ralph, F.M., Neiman, P.J., 2002. An automated brightband height detection algorithm for use with Doppler radar spectral moments. J. Atmos. Oceanic Technol. 19, 697.
White, A.B., Gottas, D.J., Henkel, A.F., Neiman, P.J., Ralph, F.M., Gutman, S.I., 2010. Developing a performance measure for snow-level forecasts. J. Hydrometeorol. 11, 739–753.
Wigmosta, M.S., Vail, L., Lettenmaier, D.P., 1994. A distributed hydrology-vegetation model for complex terrain. Water Resour. Res. 30, 1665–1679.
WMO, 1986a. Intercomparison of Models of Snowmelt Runoff. Operational Hydrology Report No. 23. World Meteorological Organisation, Geneva.
WMO, 1986b. Results of an Intercomparison of Models of Snowmelt Runoff, IAHS-AISH Publication 155, pp. 103–112.
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.
Yang, D., Goodison, B.E., Metcalfe, J.R., Golubevy, V.S., Bates, R., Pangburn, T., Hanson, C.L., 1998a. Accuracy of NWS 8″ standard non-recording precipitation gage: results and application of WMO intercomparisons. J. Atmos. Oceanic Technol. 15, 54–68.
Yang, D., Goodison, B.E., Ishida, S., 1998b. Adjustment of daily precipitation data at 10 climate stations in Alaska: application of world meteorological organization intercomparison results. Water Resour. Res. 34 (2), 241–256.
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.
Zamora, R.J., Ralph, F.M., Clark, E., Schneider, T., 2011. The NOAA hydrometeorology testbed soil moisture observing networks: design, instrumentation, and preliminary results. J. Atmos. Oceanic Technol. 28, 1129–1140. http://dx.doi.org/10.1175/2010JTECHA1465.1.
Zhang, Y., Zhang, Z., Reed, S., Koren, V., 2011. An enhanced and automated approach for deriving *a priori* SAC-SMA parameters from the Soil Survey Geographic Database. Comput. Geosci. 37 (2), 219–231. http://dx.doi.org/10.1016/j.cageo.2010.05.016.
Zhang, Z., Koren, V., Reed, S., Smith, M., Zhang, Y., Moreda, F., Cosgrove, B.A., 2012. SAC-SMA a priori parameter differences and their impact on distributed hydrologic model simulations. J. Hydrol. 420–421, 216–227. http://dx.doi.org/10.1016/j.jhydrol.2011.12.004.