Available online at www.sciencedirect.com
SCIENCE DIRECT

Journal of Hydrology xx (0000) xxx–xxx

Journal of Hydrology
www.elsevier.com/locate/jhydrol

Overall distributed model intercomparison project results

Seann Reed, Victor Koren, Michael Smith*, Ziya Zhang, Fekadu Moreda, Dong-Jun Se, DMIP Participants[1]

National Institute of Water and Atmospheric Research, New Zealand

Received 7 May 2003; revised 25 September 2003; accepted 29 March 2004

Abstract

This paper summarizes results from the Distributed Model Intercomparison Project (DMIP) study. DMIP simulations from twelve different models are compared with both observed streamflow and lumped model simulations. The lumped model simulations were produced using the same techniques used at National Weather Service River Forecast Centers (NWS-RFCs) for historical calibrations and serve as a useful benchmark for comparison. The differences between uncalibrated and calibrated model performance are also assessed. Overall statistics are used to compare simulated and observed flows during all time steps, flood event statistics are calculated for selected storm events, and improvement statistics are used to measure the gains from distributed models relative to the lumped models and calibrated models relative to uncalibrated models. Although calibration strategies for distributed models are not as well defined as strategies for lumped models, the DMIP results show that some calibration efforts applied to distributed models significantly improve simulation results. Although for the majority of basin-distributed model combinations, the lumped model showed better overall performance than distributed models, some distributed models showed comparable results to lumped models in many basins and clear improvements in one or more basins. Noteworthy improvements in predicting flood peaks were demonstrated in a basin distinguishable from other basins studied in its shape, orientation, and soil characteristics. Greater uncertainties inherent to modeling small basins in general and distinguishable inter-model performance on the smallest basin (65 km^2) in the study point to the need for more studies with nested basins of various sizes. This will improve our understanding of the applicability and reliability of distributed models at various scales.

Keywords: Distributed hydrologic modeling; Model intercomparison; Radar precipitation; Rainfall–runoff; Hydrologic simulation

1. Introduction

By ingesting radar-based precipitation products and other new sources of spatial data describing the land surface, there is potential to improve the quality and resolution of National Weather Service (NWS) river and stream forecasts through the use of distributed models. The Distributed Model Intercomparison Project (DMIP) was initiated to evaluate the capabilities of existing distributed hydrologic models forced with operational quality radar-based precipitation forcing. This paper summarizes DMIP results. The results provide insights into the simulation capabilities of 12 distributed models and suggest

* Corresponding author. Address: Hydrology Lab., Office of Hydrologic Development, Research Hydrologists, WOHD-12 NOAA/National Weather Service, 1325 East-West Highway, 20910, SIlver Spring, MD, USA.

E-mail address: michael.smith@noaa.gov (M. Smith).

[1] See Appendix A.

0022-1694/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.jhydrol.2004.03.031

HYDROL 14503—11/6/2004—21:20—SIVABAL—106592 – MODEL 3 – pp. 1–34

areas for further research. Smith et al. (2004b) provide a more detailed explanation of the motivations for the DMIP project and a description of the basins modeled. As discussed by Smith et al. (2004b), although the potential benefits of using distributed models are many, the actual benefits of distributed modeling in an operational forecasting environment, using operational quality data are largely unknown. This study analyzes model simulation results driven by observed, operational quality, precipitation data.

The NWS hydrologic forecasting requirements span a large range of spatial and temporal scales. NWS River Forecast Centers (RFCs) routinely forecast flows and stages for over 4000 points on river systems in the United States using the NWS River Forecast System (NWSRFS). The sizes of basins typically modeled at RFCs range anywhere from 300 to 5000 km^2. For flash-floods on smaller streams and urban areas, basin-specific flow or stage forecasts are only produced at a limited number of locations; however, Weather Forecast Offices (WFOs) evaluate the observed and forecast precipitation data and Flash Flood Guidance (FFG) (Sweeney, 1992) provided by RFCs to produce flash-flood watches and warnings. Lumped models are currently used at RFCs for both river forecasting and to generate FFG.

Given the prominence of lumped models in current operational systems, a key question addressed by DMIP is whether or not a distributed model can provide comparable or improved simulations relative to lumped models at RFC basin scales. In addition, the potential benefits of using a distributed model to produce hydrologic simulations at interior points are examined, although with limited interior point data in this initial study. Statistics comparing distributed model simulations to observed flows and statistics comparing the performance of distributed model and lumped model simulations are presented in this paper. Previous studies on some of the DMIP basins have shown that depending on basin characteristics, the application of a distributed or semi-distributed model may or may not improve outlet simulations over lumped simulations (Zhang et al., 2003; Koren et al., 2003a; Boyle et al., 2001; Carpenter et al., 2001; Vieux and Moreda, 2003; Smith et al., 1999).

There is no generally accepted definition for distributed hydrologic modeling in the literature. For purposes of this study, we define a distributed model as any model that explicitly accounts for spatial variability inside a basin and has the ability to produce simulations at interior points without explicit calibration at these points. The scales of parent basins of interest in this study are those modeled by RFCs. This relatively broad definition allows us compare models of widely varying complexities in DMIP. Those with a stricter definition of distributed modeling might argue that some rainfall–runoff models evaluated in this study are not true distributed models because they simply apply conceptual lumped modeling techniques to smaller modeling units. It is true that several DMIP models use algorithms similar to those of traditional lumped models for runoff generation, but in many cases, methods have been devised to estimate the spatial variability of model parameters within a basin. Several DMIP modelers have also worked on methods to estimate spatially variable routing parameters. Therefore, all models do consider the spatial variations of properties within the DMIP parent basins in some way.

The parameter estimation problem is a bigger challenge for distributed hydrologic modeling than for lumped hydrologic modeling. Although some parameters in conceptual lumped models can be related to physical properties of a basin, these parameters are most commonly estimated through calibration (Anderson, 2003; Smith et al., 2003; Gupta et al., 2003). Initial parameters for distributed models are commonly estimated using spatial datasets describing soils, vegetation, and landuse; however, these so-called physically based parameter values are often adjusted through subsequent calibration to improve streamflow simulations. These adjustments may account for many factors, including the inability of model equations and parameterizations to represent the true basin physics and heterogeneity, scaling effects, and the existence of input forcing errors. Given that parameter adjustments are used to get better model performance, the distinction between physically based parameters and conceptual model parameters becomes somewhat blurred. Although calibration strategies for distributed models are not as well defined as those for lumped models, a number of attempts have been made to use physically based parameter estimates to aid or constrain calibration and/or simulate the effects of parameter uncertainty (Koren et al., 2003a; Leavesley et al., 2003;

Vieux and Moreda, 2003; Carpenter et al., 2001; Christiaens and Feyen, 2002; Madsen, 2003; Andersen et al., 2001; Senarath et al., 2000; Refsgaard and Knudsen, 1996; Khodatalab et al., 2004). In addition, Andersen et al. (2001) incorporate multiple sites into their calibration strategy and Madsen (2003) use multiple criteria (streamflow and groundwater levels) for calibrating a distributed model, techniques that are not possible with lumped models. A key to effectively applying these approaches is that valid physical reasoning goes into deriving the initial parameter estimates.

To get a better handle on the parameter estimation problem for distributed models, participants were asked to submit both calibrated and uncalibrated distributed model results. The improvements gained from calibration are quantified in this paper. Uncalibrated results were derived using parameters that were estimated without the benefit of using the available time-series discharge data. Some of the uncalibrated parameter estimates used by DMIP participants are based on direct objective relationships with soils, vegetation, and topography data while others rely more on subjective estimates from known calibrated parameter values for nearby or similar basins. Both these objective and subjective estimation procedures are physically based to some degree. Calibrated simulations submitted by DMIP participants incorporate any adjustments that were made to the uncalibrated parameters in order to produce better matches with observed hydrographs.

In the DMIP study area, data sets from a few nested stream gauges in the Illinois River basin (Watts, Savoy, Kansas, and Christie) are available to evaluate model performance at interior points. In an attempt to understand the models' abilities to blindly simulate flows at ungauged points, the DMIP modeling instructions did not allow use of data from interior points for model calibration. However, it is recognized that an alternative approach that uses interior point data in calibration may help to improve simulations at basin outlets (e.g. Andersen et al., 2001). Only one of these interior basins (Christie) is significantly smaller (65 km^2) than the basins typically modeled by RFCs using lumped models (300–5000 km^2). As discussed below, the results for Christie are distinguishable from the results for the larger basins because of lower simulation accuracy and the relative performance of different models is not the same in Christie as it is for larger basins.

In this paper, all model comparisons are made based on streamflow, an integrated measure of hydrologic response, at basin and subbasin outlets. The focus is on streamflow analysis because no reliable measurements of other hydrologic variables (e.g. soil moisture, evaporation) were obtained for this study, and because streamflow (and the corresponding stage) forecast accuracy is the bottom line for many NWS hydrologic forecast products. Use of only observed streamflow for evaluation does limit our ability to make conclusions about the distributed models' representations of internal watershed dynamics. Therefore, it is hoped that future phases of DMIP can include comparisons of other hydrologic variables.

Following this Section 1, a Section 2 briefly describes the participant models, the NWS lumped model runs used for comparison, and events chosen for analysis. Next, Section 3 focus on the overall performance of distributed models, comparisons among lumped and distributed models, and comparisons among calibrated and uncalibrated models at all gauged locations. The variability of model simulations at ungauged interior points and trends in variability with scale are also discussed. Overall statistics and event statistics defined by Smith et al. (2004b) are presented for different models and different basins.

2. Methods

2.1. Participant models and submissions

Twelve different participants from academic, government, and private institutions submitted results for the August 2002 DMIP workshop. Table 1 provides some information about participants and general characteristics of the participating models. The first column of Table 1 lists the main affiliations for each participant, and the two or three letter abbreviation for each affiliation shown in this column will be used throughout this paper to denote results submitted by that group. Since detailed descriptions of the DMIP models are available elsewhere in the literature or this issue (See Table 1, Column 3),

HYDROL 14503—11/6/2004—21:20—SIVABAL—106592 – MODEL 3 – pp. 1–34

Table 1
Participant information and general model characteristics

Participant	Modeling system name	Primary reference (s)	Primary application	Spatial unit for rainfall–runoff calculations	Rainfall–runoff/vertical flux model	Channel routing method
Agricultural Research Service (ARS)	SWAT	Neitsch et al. (2002) and Di Luzio and Arnold (2004)	Land management/ agricultural	Hydrologic response unit (HRU) (6–7 km^2)	Multi-layer soil water balance	Muskingum
University of Arizona (ARZ)	SAC-SMA	Khodatalab et al. (2004)	Streamflow forecasting	Subbasin (avg. size ~180 km^2)	SAC-SMA	Kinematic wave
Danish Hydraulics Institute (DHI)	Mike 11	Havno et al. (1995) and Butts et al. (2004)	Forecasting, design, water management	Subbasins (~150 km^2)	NAM	Full dynamic wave solution
Environmental Modeling Center (EMC)	NOAH Land Surface Model	http://www.emc.ncep.noaa.gov/mmb/gcp/noahlsm/README_2.2.htm	Land-atmosphere interactions for climate and weather prediction models, off-line runs for data assimilation and runoff prediction	~160 km^2 (1/8th degree grids)	Multi-layer soil water and energy balance	Linearized St Venant equation
Hydrologic Research Center (HRC)	HRCDHM	Carpenter and Georgakakos (2003)	Streamflow forecasting	Subbasins (59–85 km^2)	SAC-SMA	Kinematic wave
Massachusetts Institute of Technology (MIT)	tRIBS	Ivanov et al. (2004)	Streamflow forecasting, soil moisture prediction, slope stability	TIN (~0.02 km^2)	Continuous profile soil-moisture simulation with topographicaly driven, lateral, element to element interaction	Kinematic wave
Office of Hydrologic Development (OHD)	HL-RMS	Koren et al. (2003a,b)	Streamflow forecasting	16 km^2 grid cells	SAC-SMA	Kinematic wave
University of Oklahoma (OU)	r.water.fea	Vieux (2001)	Streamflow forecasting	1 km^2 or smaller	Event based Green-Ampt infiltration	Kinematic wave
University of California at Berkeley (UCB)	VIC-3L	Liang, et al. (1994) and Liang and Xi (2001)	Land-atmosphere interactions	~160 and ~80 km^2 (1/8th, 1/16th degree grids)	Multi-layer soil water and energy balance	One parameter simple routing
Utah State University (UTS)	TOPNET	Bandaragoda et al. (2004)	Streamflow forecasting	Subbasins (~90 km^2)	TOPMODEL	Kinematic wave
University of Waterloo, Ontario (UWO)	WATFLOOD	Kouwen et al. (1993)	Streamflow forecasting	1-km grid	WATFLOOD	Linear storage routing
Wuhan University (WHU)	LL-II	–	Streamflow forecasting	4-km grid	Multi-layer finite difference model	Full dynamic wave solution

only general characteristics of these models are provided in Table 1.

Table 1 highlights both differences and similarities among modeling approaches. Some models only consider the water balance, while others (e.g. UCB, EMC, and MIT) calculate both the energy and water balance at the land surface. The sizes of the water balance modeling elements chosen for DMIP applications range from small triangulated irregular network (TIN) modeling units ($\sim 0.02\ \mathrm{km}^2$) to moderately sized subbasin units ($\sim 100\ \mathrm{km}^2$). Some models account directly or indirectly for the effects of topography on the soil-column water balance while others only explicitly use topographic information for channel and/or overland flow routing calculations. There tend to be fewer differences in the choice of a basic channel routing technique than the choice of a rainfall–runoff calculation method. Many participants use a kinematic wave approximation to the Saint-Venant equations while only a few use a more complex diffusive wave or fully dynamic solution. The methods used to estimate parameters and subdivide channel networks in applying these routing techniques do vary and are described in the individual participant papers and the references provided. It should be kept in mind that the accuracy of simulations presented in this paper reflect not only the appropriateness of the model structure, parameter estimation procedures, and computational schemes of the individual models, but also the skill, experience, and time commitment of the individual modelers to these particular basins.

The level of DMIP participation varied among participants and is indicated in Table 2. Some participants were able to submit all 30 simulations requested in the modeling instructions (i.e. both calibrated and uncalibrated results for all model points), while others submitted more limited results. An 'x' in Table 2 indicates that a flow time series was received for the specified basin and case. Table 2 shows that 198 out of a possible 360 time series files (30 cases $\times$ 12 models) were submitted and analyzed (55%). Given that research funding was not provided for participation in DMIP (aside from a small amount of travel money), this high level of participation is encouraging. Results analyzed in this paper are based on simulation time-series submitted to the NWS Office of Hydrologic Development (OHD). It is expected that individual participants may include more updated or comprehensive results for their models in other papers in this special issue.

In order to encourage as much participation as possible, there was some flexibility allowed in the types of submissions accepted for DMIP. Footnotes in Table 2 indicate some of the non-standard submissions that were accepted. Due to non-standard and/or partial submissions, some graphics and tables presented in this paper cannot include all participant models; however, they do reflect all submissions usable for the type of analysis presented. For example, all models were run in continuous simulation mode with the exception of the University of Oklahoma (OU) event simulation model. It is difficult to objectively compare event and continuous simulation models because event simulation models must include some type of scheme to define initial soil moisture conditions, an inherent feature in continuous simulation models. Overall statistics could not be computed for the OU results, but event statistics were computed when possible.

The University of California at Berkeley (UCB) submitted daily rather than hourly simulation results so only limited analyses (overall bias) of UCB results are included in this paper.

To be fair to all participants, it was agreed at the August 2002 workshop that analysis of any results submitted after the workshop should be clearly marked if they were to be included in this paper. Although the Massachusetts Institute of Technology (MIT) group was only able to submit simulations covering a part of the DMIP simulation time period prior to the August 2002 workshop, MIT was able to submit simulations covering the entire DMIP period in January 2003. Since the final simulations from MIT are not much different than the initial simulations during the overlapping time period, and use of the entire time period for analyses makes statistical comparisons more meaningful, statistics from the January 2003 MIT submissions are presented in this paper.

For those modelers who did submit calibrated results, calibration strategies varied widely in their level of sophistication, the amount of effort required, and the amount of effort invested specifically for the DMIP project. No target objective functions were prescribed for calibration so, for example,

Table 2
Level of participation

Model	Christie		Kansas		Savoy4		Savoy5		Eldon		Blue		Watts4		Watts5		Tiff City		Tahlequah	
	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc
Gaged Locations																				
ARS	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
ARZ					×	×							×	×						
DHI											×									
EMC		×		×		×		×		×		×		×		×		×		×
HRC			×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
MIT[a]	×					×			×		×	×		×						
OHD	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
OU[b]			×	×			×	×			×	×			×	×			×	×
UCB[c]											×									
UTS	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
UWO	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×	×
WHU[d]											×									
	Eldp1		Blup1		Blup2		Wttp1		Tifp1											
	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc	Cal	Unc										
Ungaged locations																				
ARS	×	×	×	×	×	×	×	×	×	×										
ARZ							×	×												
DHI			×		×															
EMC		×		×		×		×		×										
HRC	×	×	×	×	×	×	×	×	×	×										
MIT[a]	×		×	×	×	×		×												
OHD	×	×	×	×	×	×	×	×												
OU[b]			×	×	×	×														
UCB[c]																				
UTS	×	×	×	×	×	×	×	×	×	×										
UWO	×	×	×	×	×	×	×	×	×	×										
WHU[d]																				

[a] Time series submitted in January 2003 that cover the entire DMIP study period are analyzed for this paper to make statistical comparisons more meaningful.
[b] Simulations submitted only for selected events.
[c] Results have a daily time step.
[d] Calibration is based on only 1 year of observed flow (1998). Results submitted January 2003.

some participants may have placed more emphasis on fitting flood peaks than obtaining a zero simulation bias for the calibration period. This is not a big concern in evaluating DMIP results because a variety of statistics are considered and results indicate that models with good results based on one statistical criterion typically have good results for other statistical criteria as well. Discussion of participant parameter estimation and calibration strategies is beyond the scope of this paper but information about participant-specific procedures can be found in the references listed in Table 1.

2.2. Lumped model

To provide a 'standard' for comparison, both calibrated and uncalibrated lumped simulations were generated at OHD for all of the gauged DMIP locations. Techniques used to generate lumped simulations are the same as those used for operational forecasting at most NWS River Forecast Centers (RFCs). The Sacramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al., 1973; Burnash, 1995) is used for rainfall–runoff calculations and the unit hydrograph model is used for channel flow routing. For the DMIP basin calibration runs, SAC-SMA parameters were estimated using manual calibration at OHD following the strategy typically used at RFCs and described by Smith et al. (2003) and Anderson (2003). As defined by Smith et al. (2004b), the calibration period was June 1, 1993 to May 31, 1999. Model parameters routinely used for operational forecasting in the DMIP basins by the Arkansas-Red Basin RFC (ABRFC) could not be used directly to produce lumped simulations because these parameters are based on 6-h calibrations (hourly simulations are the standard in DMIP) with gauged-based rainfall, and it is well known that SAC-SMA model results are sensitive to the time step used for model calibration (Koren et al., 1999; Finnerty et al., 1997).

Lumped SAC-SMA parameters derived for the DMIP basins are given in Table 3. No snow model was included in the lumped runs for these basins because snow has a very limited effect on the hydrology of the DMIP basins. For the lumped DMIP runs, constant climatological mean monthly values for potential evaporation (PE) (mm/day) were used. In the SAC-SMA model, evapotranspiration (ET) demand is defined as the product of PE and a PE adjustment factor, which is related to the vegetation state. During manual calibration, PE adjustment factors are initially assigned based on regional knowledge but may be adjusted during the calibration process to remove seasonal biases. The ET demand values used for calibrated lumped DMIP runs are also given in Table 3.

Because climatological mean ET demand values were used for lumped runs, the only observed input forcing required to produce the lumped model simulations was hourly rainfall. Hourly time series of lumped rainfall to force lumped model runs were obtained by computing the areal averages from hourly multi-sensor rainfall grids (the same rainfall grids used to drive the distributed models being tested). Areal averages for a basin were computed using all rainfall grid cells with their center point inside the basin. Algorithms used to develop the multi-sensor rainfall products used in this study are described by Seo and Breidenbach (2002), Seo et al. (2000), Seo et al. (1999) and Fulton et al. (1998). There are some known biases in the cumulative precipitation estimates during the study period that

Table 3
SAC-SMA and ET demand parameters for 1-h Lumped calibrations

Parameter	Blue	Eldon, Christie	Tahlequah, Watts, Kansas, Savoy	Tiff City
Uztwm (mm)	45	50	40	70
Uzfwm (mm)	50	25	35	34
Uzk (day^{-1})	0.5	0.35	0.25	0.25
Pctim	0.005	0	0.005	0.002
Adimp	0	0	0.1	0
Riva	0.03	0.035	0.02	0.025
Zperc	500	500	250	250
Rexp	1.8	2	1.7	1.6
Lztwm (mm)	175	120	80	135
Lzfsm (mm)	25	25	27	21
Lzfpm (mm)	100	75	200	125
Lzsk (day^{-1})	0.05	0.08	0.08	0.12
Lzpk (day^{-1})	0.003	0.004	0.002	0.003
Pfree	0.05	0.25	0.1	0.15
Rserv	0.3	0.3	0.3	0.3
Month	ET Demand (mm/day)			
Jan	1.1	0.75	0.77	0.77
Feb	1.2	0.8	0.93	0.83
Mar	1.6	1.4	1.70	1.42
Apr	2.4	2.1	2.68	2.48
May	3.5	3.2	3.81	3.96
Jun	4.8	4.3	5.25	5.44
Jul	5.1	5.8	5.97	5.93
Aug	4.2	5.7	5.87	5.86
Sep	3.4	3.9	4.02	3.97
Oct	2.4	2.3	2.37	2.36
Nov	1.6	1.2	1.24	1.24
Dec	1.1	0.8	0.82	0.81

are discussed further in the results section (see also Johnson et al., 1999; Young et al., 2000; 'About the StageIII Data', http://www.nws.noaa.gov/oh/hrl/dmip/stageiii_info.htm; Wang et al., 2000; Guo et al., 2004). Smith et al. (2004a) discuss the spatial variability of the precipitation data over the DMIP basins independently of the hydrologic model application.

For gauged interior points (Kansas, Savoy, Christie, and Watts (when calibration is done at Tahlequah)), there are no fully calibrated lumped results. That is, no manual calibrations against observed streamflow were attempted at these points; however, we refer to lumped, interior point simulations using the calibrated SAC-SMA parameter estimates from parent basins as calibrated runs. As shown in Table 3, the calibrated SAC-SMA parameters for Eldon and Christie are the same, as are the parameters for Tahlequah, Watts, Kansas, and Savoy. There was an attempt to calibrate Tahlequah separately from Watts; however, since this analysis led to similar parameters for both Tahlequah and Watts, lumped simulation results used for analysis in DMIP were generated using the same SAC-SMA parameters for both Tahlequah and Watts.

To generate uncalibrated lumped SAC-SMA parameters for parent basins and interior points, areal averages of gridded a priori SAC-SMA parameters defined by Koren et al. (2003b) were used. Uncalibrated ET demand estimates were derived by averaging gridded ET demand estimates computed by Koren et al. (1998). Koren et al. (1998) produced 10-km mean monthly grids of PE and PE adjustment factors for the conterminous United States.

Hourly unit hydrographs for each of the parent basins (Blue, Tahlequah, Watts, Eldon, and Tiff City) were derived initially using the Clark time-area approach (Clark, 1945) and then adjusted (if necessary) during the manual calibration procedure. No manual adjustments were made to the Clark unit hydrographs for uncalibrated runs. Unit hydrographs for interior point simulations were derived using the same method but with no manual adjustment for both ‘calibrated’ and uncalibrated runs.

Fig. 1a and b show unit hydrographs used for the lumped simulations. Looking at the unit hydrographs for parent basins (Fig. 1a), the general trend that larger basins tend to peak later makes sense. Tahlequah is the largest basin, followed by Tiff City, Watts, Blue, and Eldon (See Smith et al. (2004b) for exact basin sizes). The shape of the Blue unit hydrograph is somewhat unusual because it has a flattened peak and no tail. The different hydrologic response characteristics for the Blue River are also seen in the observed data and distributed modeling results. The same sensible trend is evident in Fig. 1b for the smaller basins.

2.3. Events selected

For statistical analysis, between 16 and 24 storm events were selected for each basin. Tables 4–8 list

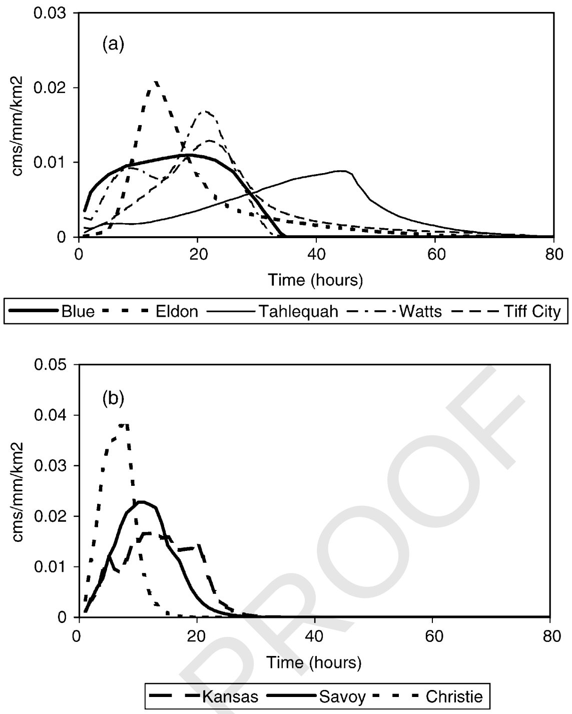

Fig. 1. Unit hydrographs for (a) parent basins, and (b) interior points.

events selected for Tahlequah and Watts, Kansas, Savoy, Eldon and Christie, and Blue, respectively. In some cases, the same time windows were selected for both interior points and parent basins (e.g. Eldon and Christie), while in other cases the time windows are slightly different to better capture the event hydrograph (e.g. Kansas and Savoy event windows are different than the parent basins Tahlequah and Watts). Fewer events were used for the Savoy analysis because the available Savoy observed flow data record does not start until October, 1995. For the Blue River, some seemingly significant events were excluded from the analysis because of significant periods of missing streamflow observations.

The selection of storms was partially subjective and partially objective. The method for selection was primarily visual inspection of observed streamflow and the corresponding mean areal rainfall values. Although the goal of forecasting floods tends to encourage analysis primarily of large events, we are also interested in studying model performance over a range of event sizes and the relationships between

Table 4
Selected events for tahlequah and watts

Event	Start time		End time		Tahlequah Peak ($m^3 s^{-1}$)	Watts Peak ($m^3 s^{-1}$)	Tahlequah volume (mm)	Watts volume (mm)
1	1/13/1995	0:00	1/26/1995	24:00	430	345	50.6	54.1
2	3/4/1995	16:00	3/11/1995	15:00	202	191	15.3	17.5
3	4/20/1995	0:00	4/30/1995	23:00	362	402	31.4	38.4
4	5/7/1995	0:00	5/14/1995	23:00	580	535	52.8	51.6
5	6/3/1995	0:00	6/19/1995	23:00	436	410	56.9	58.8
6	5/10/1996	16:00	5/17/1996	13:00	262	252	18.1	20.9
7	9/26/1996	0:00	10/4/1996	23:00	542	590	35	37
8	11/4/1996	12:00	11/14/1996	23:00	498	525	32.9	38.8
9	11/24/1996	1:00	12/5/1996	9:00	483	449	63.1	71.8
10	2/19/1997	2:00	2/25/1997	23:00	597	536	38.8	41.2
11	8/17/1997	0:00	8/23/1997	23:00	42	62	4.94	5.8
12	1/4/1998	0:00	1/16/1998	23:00	729	727	81.5	84.6
13	3/16/1998	0:00	3/26/1998	23:00	349	315	48.4	49.6
14	10/5/1998	0:00	10/11/1998	23:00	206	179	17	14.9
15	2/7/1999	0:00	2/15/1999	23:00	276	233	28.4	23.2
16	4/4/1999	0:00	4/10/1999	23:00	132	151	17.3	22.4
17	5/4/1999	0:00	5/11/1999	23:00	370	343	35.7	31.7
18	6/24/1999	0:00	7/6/1999	23:00	556	627	48.4	55.9
19	1/2/2000	0:00	1/9/2000	23:00	40	45	5.71	5.31
20	5/26/2000	0:00	6/1/2000	23:00	191	170	14.3	12.6
21	6/15/2000	13:00	7/10/2000	23:00	992	870	191	172

model structure and simulation performance over various flow ranges. Therefore, all of the largest storms were selected, several moderately sized storms, and a few small storms. To the degree possible, storms were selected uniformly throughout the study period (approximately the same number each year) and from different seasons.

Due to the subjective nature of defining the event windows and the fact that different OHD personnel selected event windows for different basins, there are some subtle differences in how much of the storm tails are included in the event windows. For example, Eldon event windows tend to include less of the hydrograph tail than windows defined for other basins. This means that storm volumes for selected events shown in Table 7 may not reflect all of the runoff associated with that particular event. Also, in a few cases, multiple flood peaks occurring close in time were treated as one event (e.g. Event 21 for Tahlequah and Watts) in one basin but as separate events for another basin (e.g. Events 22–24 for Eldon). These small differences in how event windows were defined for different basins have little impact on the conclusions of this paper.

3. Results and discussion

Overall statistics, event statistics, and event improvement statistics will be presented and discussed. Mathematical definitions of the statistics used here are provided by Smith et al. (2004b). The event improvement statistics (flood runoff improvement, peak flow improvement, and peak time improvement) are used to measure the improvement from distributed models relative to lumped models and the improvement from calibrated models relative to uncalibrated models.

3.1. Overall Statistics

Fig. 2a and b show the cumulative simulation errors for models applied to the Watts and Blue River basins. The vertical gray line in these figures indicates the end of the calibration period. The trends in these graphs reflect known historical bias characteristics in the radar rainfall archives. At several times during the 1990's, there were improvements to the algorithms used to produce multi-sensor precipitation grids at RFCs, and therefore the statistical characteristics of multi-sensor precipitation grids archived at

Table 5
Selected events for Kansas

Event	Start time		End time		Peak ($m^3 s^{-1}$)	Volume (mm)
1	1/13/1995	0:00	1/18/1995	23:00	60	30.7
2	3/6/1995	0:00	3/10/1995	23:00	22	12.8
3	5/6/1995	0:00	5/12/1995	23:00	94	47.7
4	6/8/1995	0:00	6/15/1995	23:00	27	40.2
5	5/10/1996	17:00	5/14/1996	23:00	14	6.99
6	9/26/1996	0:00	9/29/1996	23:00	79	17.2
7	11/6/1996	0:00	11/12/1996	23:00	27	16.4
8	11/24/1996	2:00	12/4/1996	23:00	45	46.4
9	2/20/1997	0:00	2/25/1997	23:00	272	53.9
10	8/17/1997	0:00	8/21/1997	23:00	5	3.92
11	1/4/1998	0:00	1/14/1998	23:00	72	61.3
12	3/16/1998	0:00	3/24/1998	23:00	37	38
13	10/5/1998	0:00	10/11/1998	23:00	27	13.8
14	2/7/1999	0:00	2/11/1999	23:00	85	26.4
15	4/4/1999	0:00	4/9/1999	23:00	8	9.35
16	5/4/1999	0:00	5/9/1999	23:00	89	39.5
17	6/24/1999	0:00	7/6/1999	23:00	162	57.3
18	1/3/2000	0:00	1/7/2000	23:00	6	4.37
19	5/27/2000	0:00	5/30/2000	23:00	9	4.61
20	6/16/2000	0:00	7/4/2000	23:00	538	207

the ABRFC have changed over time (Young et al., 2000; 'About the StageIII Data', http://www.nws.noaa.gov/oh/hrl/dmip/stageiii_info.htm). In the earlier years of multi-sensor precipitation processing, gridded products tended to underestimate the amount of rainfall relative to gauge-only rainfall estimates. The underestimation of simulated flows in the early years seen in Fig. 2 is consistent with this known trend. In the latter part of the total simulation period (June 1999–July 2000), the fact that the slopes of the cumulative error curves tend to level off for several of the models is a positive indicator that issues of rainfall bias are being dealt with in the multi-sensor rainfall processing procedures; however, a longer

Table 6
Selected events for Savoy

Event	Start time		End time		Peak ($m^3 s^{-1}$)	Volume (mm)
1	5/10/1996	16:00	5/13/1996	13:00	190	24.7
2	9/26/1996	0:00	10/4/1996	23:00	26	10.5
3	11/5/1996	13:00	11/14/1996	23:00	313	55.4
4	11/24/1996	2:00	12/4/1996	9:00	202	86.6
5	2/20/1997	2:00	2/25/1997	23:00	274	47.4
6	8/17/1997	0:00	8/20/1997	23:00	10	1.5
7	1/4/1998	0:00	1/16/1998	23:00	823	135
8	3/16/1998	0:00	3/24/1998	23:00	137	47.1
9	10/5/1998	0:00	10/10/1998	23:00	166	24.9
10	2/7/1999	0:00	2/13/1999	23:00	150	24.1
11	4/3/1999	0:00	4/8/1999	23:00	93	22.9
12	5/4/1999	0:00	5/8/1999	23:00	184	24.5
13	6/29/1999	0:00	7/5/1999	23:00	350	45.3
14	1/2/2000	0:00	1/5/2000	23:00	25	4.1
15	5/26/2000	0:00	5/31/2000	23:00	145	19.9
16	6/16/2000	13:00	7/8/2000	23:00	651	204

Table 7
Selected events for Eldon and Christie

Event	Start time		End time		Eldon peak ($m^3 s^{-1}$)	Eldon volume (mm)	Christie peak ($m^3 s^{-1}$)	Christie volume (mm)
1	11/4/1994	14:00	11/8/1994	24:00	152	27	9	20.4
2	1/13/1995	6:00	1/17/1995	23:00	289	43.6	9	24.9
3	4/20/1995	1:00	4/22/1995	23:00	205	19.8	4	11.8
4	5/6/1995	18:00	5/11/1995	23:00	532	62.8	26	42.9
5	6/9/1995	1:00	6/12/1995	23:00	133	28.7	3	0.6
6	1/18/1996	13:00	1/20/1996	23:00	217	14.3	1	2.1
7	4/22/1996	1:00	4/23/1996	4:00	221	9.42	6	3.2
8	5/10/1996	23:00	5/13/1996	12:00	189	15.6	2	5.4
9	9/26/1996	5:00	9/29/1996	23:00	874	62.8	53	48.4
10	11/7/1996	1:00	11/10/1996	23:00	429	38.3	7	20.1
11	11/16/1996	22:00	11/18/1996	23:00	129	11.9	4	8.0
12	11/24/1996	1:00	11/25/1996	15:00	347	28.2	10	14.7
13	2/20/1997	14:00	2/24/1997	23:00	893	62.3	51	43.3
14	1/4/1998	1:00	1/7/1998	23:00	894	75.7	62	41.7
15	1/8/1998	1:00	1/11/1998	18:00	197	39.3	7	21.6
16	3/15/1998	20:00	3/22/1998	23:00	217	54.4	9	33.6
17	10/5/1998	15:00	10/8/1998	23:00	274	20.8	4	6.6
18	3/12/1999	19:00	3/16/1999	23:00	187	32.8	8	23
19	5/4/1999	3:00	5/7/1999	23:00	351	30.1	12	18.6
20	6/30/1999	1:00	7/2/1999	23:00	100	10.2	1	2.5
21	5/26/2000	1:00	5/29/2000	23:00	260	20.8	2	5.5
22	6/17/2000	1:00	6/20/2000	18:00	303	31.7	9	18.6
23	6/20/2000	19:00	6/24/2000	23:00	1549	106	136	86.2
24	6/28/2000	1:00	7/1/2000	23:00	407	38.9	40	58.8

period of record will be required to confirm this observation. For future hydrologic studies with multi-sensor precipitation grids, OHD plans to do reanalysis of archived multi-sensor precipitation grids to remove biases and other errors; however it was not possible to do this analysis prior to DMIP.

Fig. 2 shows that not all modelers placed priority on minimizing simulation bias during the calibration period as a criterion for calibration. NWS calibration strategies (Smith et al., 2003; Anderson, 2003), do emphasize producing a low cumulative simulation bias over the entire calibration period and this strategy is reflected in the lumped (LMP) model results. The cumulative error for the Watts LMP model at the end of the calibration period is about −97 mm or 4.1% and the cumulative error for the Blue LMP model is about −21 mm or 1.5%. As one might expect, several of the calibrated distributed models (ARS, LMP, ARZ, OHD, and HRC) also produce relatively small cumulative errors over the calibration period. Models that do achieve a small bias over the calibration period tend to underestimate flows more in earlier years (to about mid-1997), reflecting low rainfall estimates, and overestimate flows in the later years up to the end of the calibration period, in an attempt maintain a small simulation bias over the whole period.

In the DMIP modeling instructions, a distinct calibration period from June 1, 1993, to May 31, 1999, and validation period from June 1, 1999, to July 31, 2000 were defined. However, many of the statistics presented in this paper are computed over a single time period that overlaps both the original calibration and validation periods: April 1, 1994, to July 31, 2000. There are several reasons for this. One reason that the validation statistics are not presented separately in most graphs and tables is that the original validation period is relatively short and contains only a few or no significant storm events (no significant events on the Blue River). Early on in DMIP the intention was to have a longer validation period (i.e. through July, 2001) but the energy forcing data required for some of the models was

Table 8
Selected events for Blue

Event	Start time		End time		Peak ($m^3 s^{-1}$)	Volume (mm)
1	4/25/1994	0:00	5/8/1994	23:00	224	59.1
2	11/12/1994	0:00	11/27/1994	23:00	215	43.8
3	12/7/1994	0:00	12/13/1994	23:00	142	22
4	3/12/1995	0:00	3/20/1995	23:00	148	30.2
5	5/6/1995	0:00	5/21/1995	23:00	289	71.8
6	9/17/1995	0:00	9/24/1995	23:00	47	5.1
7	9/26/1996	0:00	10/11/1996	23:00	156	10.6
8	10/19/1996	0:00	11/3/1996	23:00	253	37.4
9	11/6/1996	0:00	11/21/1996	23:00	483	48.4
10	11/23/1996	0:00	12/6/1996	23:00	230	62.3
11	2/18/1997	0:00	3/5/1997	23:00	194	44.9
12	3/25/1997	0:00	3/30/1997	23:00	60	6.1
13	6/9/1997	0:00	6/16/1997	23:00	130	8.2
14	12/20/1997	0:00	12/28/1997	23:00	120	22
15	1/3/1998	0:00	1/14/1998	23:00	176	59.3
16	3/6/1998	0:00	3/13/1998	23:00	118	15.8
17	3/14/1998	0:00	3/29/1998	23:00	204	51.6
18	1/28/1999	0:00	2/2/1999	23:00	25	3.6
19	3/27/1999	0:00	4/7/1999	23:00	172	17
20	6/22/1999	0:00	7/6/1999	23:00	29	5.7
21	9/8/1999	0:00	9/24/1999	23:00	17	3.4
22	12/9/1999	0:00	12/19/1999	23:00	26	3.0
23	2/22/2000	0:00	3/2/2000	23:00	11	2.6
24	4/29/2000	0:00	5/11/2000	23:00	23	4.8

only available through July 31, 2000, and therefore the validation period duration was shortened. We feel that for most graphs and tables, separately presenting numerous statistical results for a distinct, but short, validation period will not strengthen the conclusions of this paper, but rather, would add unnecessary length and detail. The starting date for the April, 1994–July, 2000 statistical analysis period (10 months after the June 1993 calibration start date) allows for a model warm-up period to minimize the effects of initial conditions on results. Unless otherwise noted, this analysis period is used for all statistics presented.

Fig. 3a and b show the overall Nash-Sutcliffe efficiency (Nash and Sutcliffe, 1970) for uncalibrated and calibrated models respectively for all basins while Fig. 4a and b show the overall modified correlation coefficients, r_{mod} (McCuen and Snyder, 1975; Smith et al., 2004b). Tables 9 and 10 list the overall statistics used to produce Figs. 3 and 4. It is desirable to have both Nash-Sutcliffe and r_{mod} values close to one. In Figs. 3a and 4a, dashed lines indicate the arithmetic average of uncalibrated results. In Figs. 3b and 4b, dashed lines for both the average of uncalibrated and calibrated results are shown (each point used to draw these lines is the average of all model results for a given basin). These lines show an across the board improvement in average model performance after calibration.

Note that the results labeled 'Watts4' and 'Savoy4' shown in Figs. 3 and 4 correspond to modeling instruction number 4 described by Smith et al. (2004b), which specifies calibration at Watts rather than at Tahlequah. Results for 'Watts5' and 'Savoy5' from calibration at Tahlequah are similar to 'Watts4' and 'Savoy4' (see discussion below), and therefore are not included on these graphs.

The basins in Figs. 3 and 4 are listed from left to right in order of increasing drainage area. A noteworthy trend is that both the Nash–Sutcliffe efficiency and correlation coefficient are poorer (on average) for the smaller interior points (particularly for Christie and Kansas). A primary contributing factor to this may be that smaller basins have less

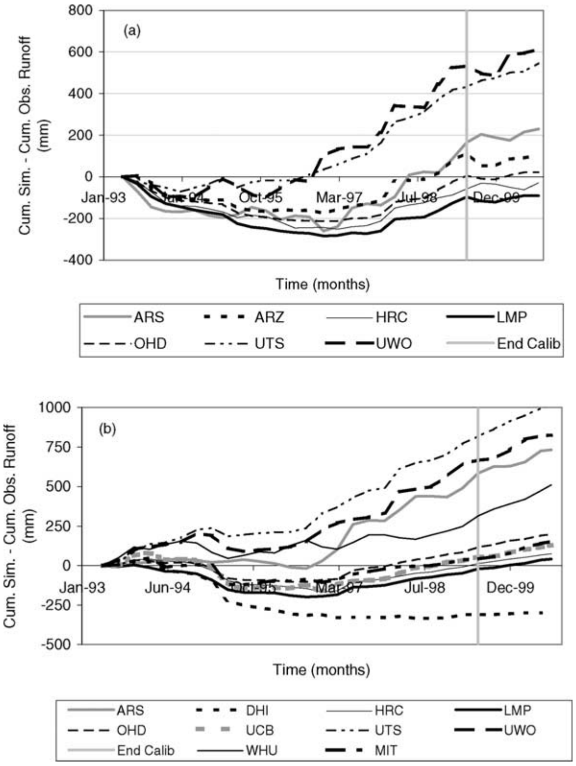

Fig. 2. Cumulative simulation errors for calibrated models: (a) Watts and (b) Blue.

capacity to dampen out inputs and corresponding input errors. Fig. 5 shows that observed streamflows in small basins do in fact exhibit more variability than streamflows on larger basins, making accurate simulation more difficult. There is also more uncertainty in the spatially averaged rainfall estimates for smaller basins. Another possible contributing factor to this trend for the calibrated results is that simulations for Christie, Kansas, and Savoy used parameters calibrated for the parent basin only, without the use of streamflow data from the Christie, Kansas, or Savoy gauges. However, this cannot be the only factor since the trend exists for both calibrated and uncalibrated results.

The fact that calibrated models have improved statistics on average over uncalibrated models agrees with the consensus in the literature cited in Section 1 that some type of calibration is beneficial when estimating distributed model parameters from physical data. The improvements from calibration are also evident in Section 3.2 discussing event statistics (Fig. 17). Since uncalibrated models do not have the benefit of accounting for the known biases in the rainfall archives over the calibration period and the calibrated models do, one could question whether or not the calibrated models would outperform uncalibrated models in the absence of these biases. Overall r_{mod} statistics computed separately for the validation period (average lines for all calibrated and uncalibrated models are shown in Fig. 6) indicate that on average, the calibrated models still outperform uncalibrated models in the validation period, during which the calibration adjustments cannot account for any rainfall biases.

3.2. Event statistics

The event statistics percent absolute runoff error and percent absolute peak error for different basins are shown in Figs. 7–14. Figs. 7a and 8a, etc. show uncalibrated results and Figs. 7b and 8b, etc. show calibrated results. The best results with the lowest event runoff and peak errors are located nearest the lower left corner in these graphs. Data used to produce these graphs are summarized in Tables 11 and 12.

Looking collectively at the calibrated results in Figs. 7–14, a calibrated model that performs relatively well in one basin typically has about the same relative performance in other basins with the notable exception of the smallest basin (Christie). For Christie (Fig. 7b), the UTS model produces by far the best percent absolute event runoff error and percent absolute peak error results; however, the UTS model does not perform as well in the larger basins. Although not a physical explanation, an examination of the event runoff bias statistics shown in Table 13 can offer some understanding as to why this reversal of performance occurs. The UTS model tends to underestimate event runoff for all basins except Blue and Christie. For Christie, although the UTS model overestimates event runoff, it is a less extreme overestimation than some of the other models. This suggests that the UTS model's tendency to simulate relatively lower flood runoff serves it well statistically in Christie where several other models significantly overestimate flood runoff. Further study is needed to understand the reason for the tendency of most models to overestimate peaks in Christie. The performance of the MIT and UWO models is also improved for

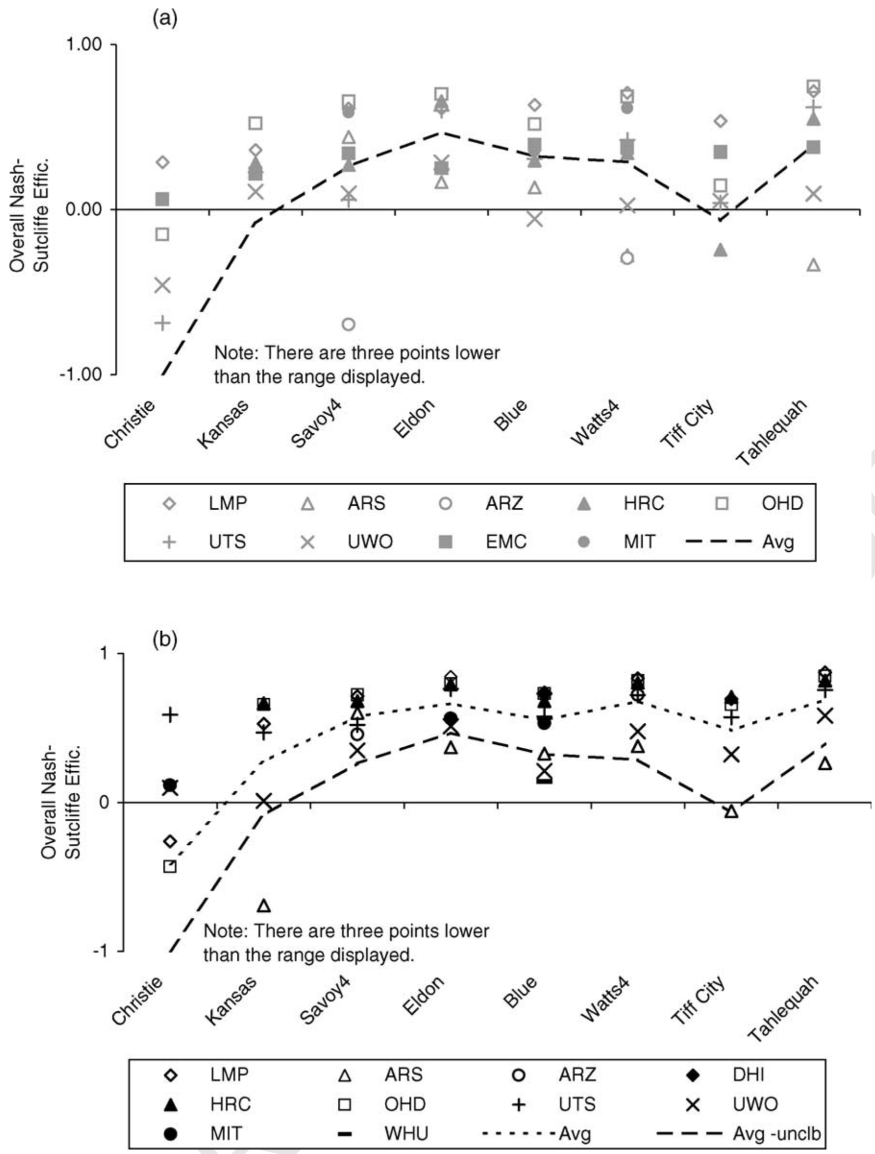

Fig. 3. Overall Nash-Sutcliffe efficiency for April 1994–July 2000: (a) uncalibrated models and (b) calibrated models.

Christie relative to the performance of these models in the parent basin for Christie (Eldon, Fig. 10b).

For the calibrated results, the three models that consistently exhibit the best performance on basins other than Christie (LMP, OHD, and HRC) all use the SAC-SMA model for soil moisture accounting. The OHD and HRC distributed modeling approaches both combine features of conceptual lumped models for rainfall–runoff calculations and physically based routing models. Although only available for the Blue River, the DHI submission showed comparable performance to these three models. Similar to the OHD and HRC models, the DHI modeling approach for the results presented here was to subdivide the Blue River into smaller units (eight subbasins supplied by OHD), apply conceptual rainfall–runoff modeling methods to those smaller units (again, methods like those used in lumped models),

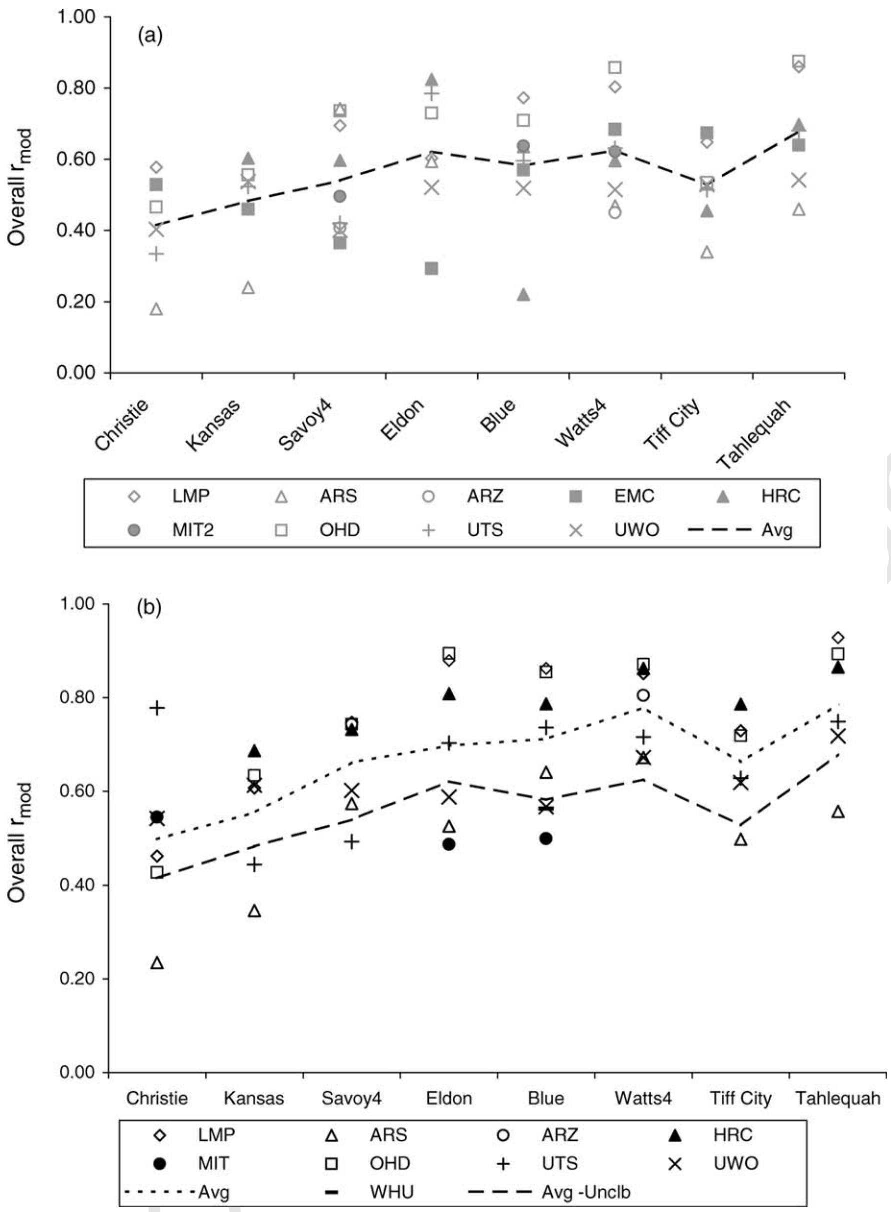

Fig. 4. Overall r_{mod} for April 1994–July 2000: (a) uncalibrated models and (b) calibrated models.

and then use a physically based method to route the water to the outlet (DHI used a fully dynamic solution of the St. Venant equation). The same eight subbasins used by DHI were also used in the earlier modeling studies by Boyle et al. (2001) and Zhang et al. (2003).

For the better performing models, the percent absolute peak errors shown in Figs. 7–14 are noticeably higher for the three smallest basins, while the percent absolute runoff errors appear to be less sensitive to basin size.

Improvement indices quantifying the benefits of calibration on event statistics are described in Section 3.3, but comparing uncalibrated and calibrated graphs in Figs. 7–14 also provides a sense of the gains that were made from calibration for various models. The scales for uncalibrated and calibrated graph pairs are

Table 9
Overall Nash–Sutcliffe efficiencies for Fig. 3

	Christie	Kansas	Savoy4	Eldon	Blue	Watts4	Tiff City	Tahlequah
Uncalibrated								
LMP	0.29	0.36	0.61	0.61	0.63	0.71	0.54	0.72
ARS	−5.03	−2.29	0.44	0.17	0.14	−0.28	−1.35	−0.33
ARZ			−0.70			−0.29		
EMC	0.06	0.22	0.34	0.25	0.40	0.37	0.35	0.38
HRC		0.28	0.27	0.66	0.30	0.34	−0.24	0.55
MIT			0.59		0.36	0.61		
OHD	−0.15	0.52	0.66	0.70	0.52	0.69	0.15	0.75
UTS	−0.69	0.23	0.06	0.60	0.31	0.42	0.04	0.62
UWO	−0.46	0.11	0.10	0.29	−0.06	0.03	0.05	0.10
Calibrated								
LMP	−0.26	0.53	0.71	0.85	0.72	0.83	0.69	0.87
ARS	−2.58	−0.69	0.60	0.37	0.33	0.38	−0.06	0.27
ARZ			0.46			0.72		
DHI					0.73			
HRC		0.67	0.68	0.79	0.68	0.81	0.71	0.82
MIT	0.12			0.57	0.53			
OHD	−0.43	0.66	0.72	0.80	0.73	0.82	0.66	0.85
UTS	0.59	0.47	0.52	0.76	0.58	0.72	0.57	0.76
UWO	0.10	0.01	0.35	0.51	0.21	0.48	0.32	0.58
WHU					0.14			

Table 10
Overall modified correlation coefficients (r_{mod}) for Fig. 4

	Christie	Kansas	Savoy4	Eldon	Blue	Watts4	Tiff City	Tahlequah
Uncalibrated								
LMP	0.58	0.46	0.70	0.60	0.77	0.80	0.65	0.86
ARS	0.18	0.24	0.74	0.59	0.64	0.47	0.34	0.46
ARZ			0.41			0.45		
EMC	0.53	0.46	0.37	0.29	0.57	0.68	0.67	0.64
HRC		0.60	0.60	0.82	0.22	0.60	0.46	0.70
MIT			0.50		0.64	0.62		
OHD	0.47	0.56	0.74	0.73	0.71	0.86	0.54	0.88
UTS	0.33	0.52	0.42	0.79	0.60	0.63	0.51	0.68
UWO	0.40	0.54	0.40	0.52	0.52	0.52	0.53	0.54
Calibrated								
LMP	0.46	0.61	0.75	0.88	0.86	0.85	0.73	0.93
ARS	0.24	0.35	0.57	0.53	0.64	0.67	0.50	0.56
ARZ			0.74			0.81		
DHI					0.78			
HRC		0.69	0.73	0.81	0.79	0.86	0.79	0.87
MIT	0.55			0.49	0.50			
OHD	0.43	0.63	0.74	0.89	0.86	0.87	0.72	0.89
UTS	0.78	0.44	0.49	0.70	0.74	0.72	0.63	0.75
UWO	0.54	0.61	0.60	0.59	0.57	0.67	0.62	0.72
WHU					0.56			

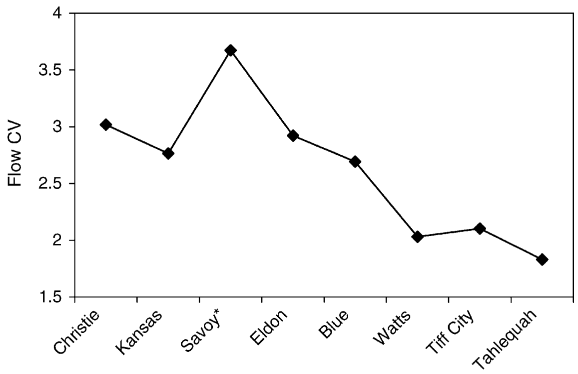

Fig. 5. Coefficients of Variation (CV) for hourly streamflow, April 1994–July 2000 (*Savoy period is October 1995–July 2000).

the same, and in general, the uncalibrated results are more scattered, dictating the domain and range required for the graph pairs presented. A big improvement from an uncalibrated to a calibrated result for an individual model does not necessarily indicate better calibration techniques were used for that model. It could mean that the scheme used with that model to estimate initial (uncalibrated) model parameters is less effective and therefore the potential gain from calibration is greater.

Not all participants in DMIP defined calibration in the same way, and varying levels of emphasis were placed on calibration. For example, EMC submitted only uncalibrated results. Among uncalibrated models, the relative performance of the EMC model is interesting because it varies quite a bit among different basins. It is surprising that the relatively

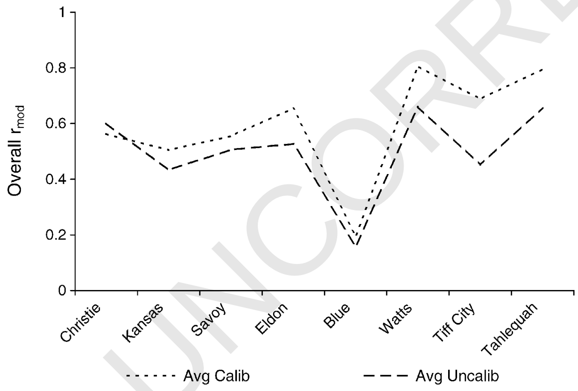

Fig. 6. Overall r_{mod}: Averaged values for calibrated and uncalibrated models during the validation period (June 1999–July 2000).

coarse resolution EMC model (1/8 degree grid boxes) does relatively well in terms of the percent peak error statistics for Christie (similar performance to the calibrated UTS model). Visual examination of event hydrographs reveals that the EMC model predicts relatively good flood volume and peak flow estimates for Christie. However, as might be expected with such a coarse resolution, the shapes of hydrographs are rather poor (wide at the top with steep recessions).

Some caution is warranted in interpreting the results for Christie given that some of the distributed Christie submissions were generated by models with a relatively coarse computational resolution compared to the size of the basin (e.g. EMC and OHD). These models would not satisfy the criterion suggested by Kouwen and Garland (1989) that at least five subdivisions are required to provide a meaningful representation of a basin's area and drainage pattern with a distributed model. Numerical experiments run in OHD using multi-sensor precipitation data in and around the DMIP basins suggest a similar criterion. These experiments showed that representing a basin using ten or more elements significantly reduces the error dependency on the scale of rainfall averaging.

3.3. Event improvement statistics

Fig. 15a–c show flood runoff, peak flow, and peak time improvement for calibrated distributed models relative to the 'standard' calibrated lumped model. There are 51 points (model-basin combinations) shown in each of Fig. 15a–c. To prevent outliers in small basins from dominating the graphing ranges for all basins, different plotting scales are used for the three smallest basins (Christie, Kansas, and Savoy). There are more cases when the lumped model outperforms a distributed model (negative improvement) than when a distributed model outperforms the lumped model. Only 14% of cases show flood runoff improvement greater than zero, 33% show peak flow improvement greater than zero, and 22% show peak time improvement greater than zero. The percentages of cases with flood runoff and peak flow improvement statistics greater than −5% are 43 and 51%, respectively, and in 33% of cases, peak time improvements are greater than −1 h. Therefore, although there are many cases where certain calibrated distributed models cannot outperform the calibrated lumped model, there are also

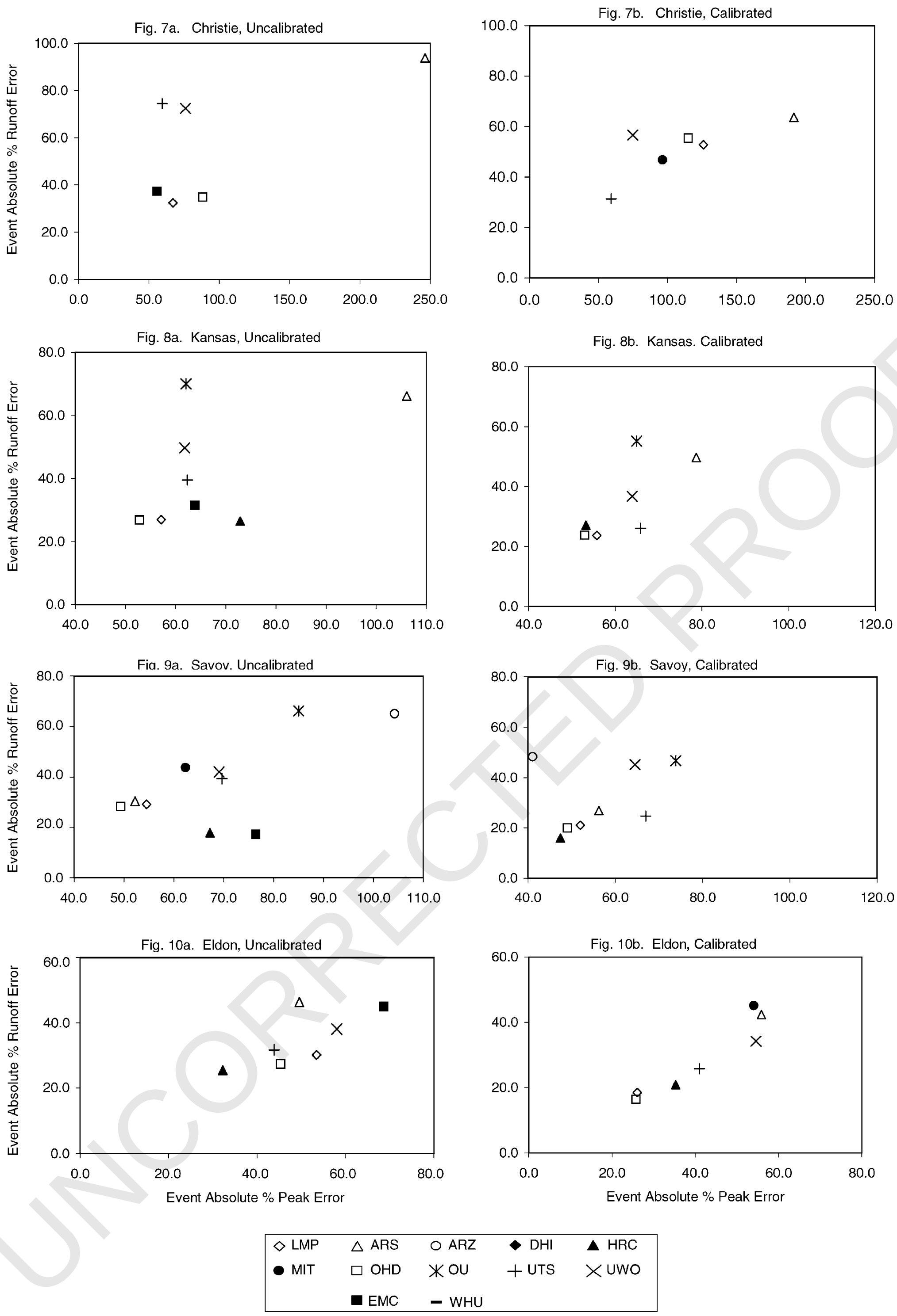

Fig. 7–14. Event percent absolute runoff error versus event percent absolute peak error for (a) uncalibrated and (b) calibrated cases.

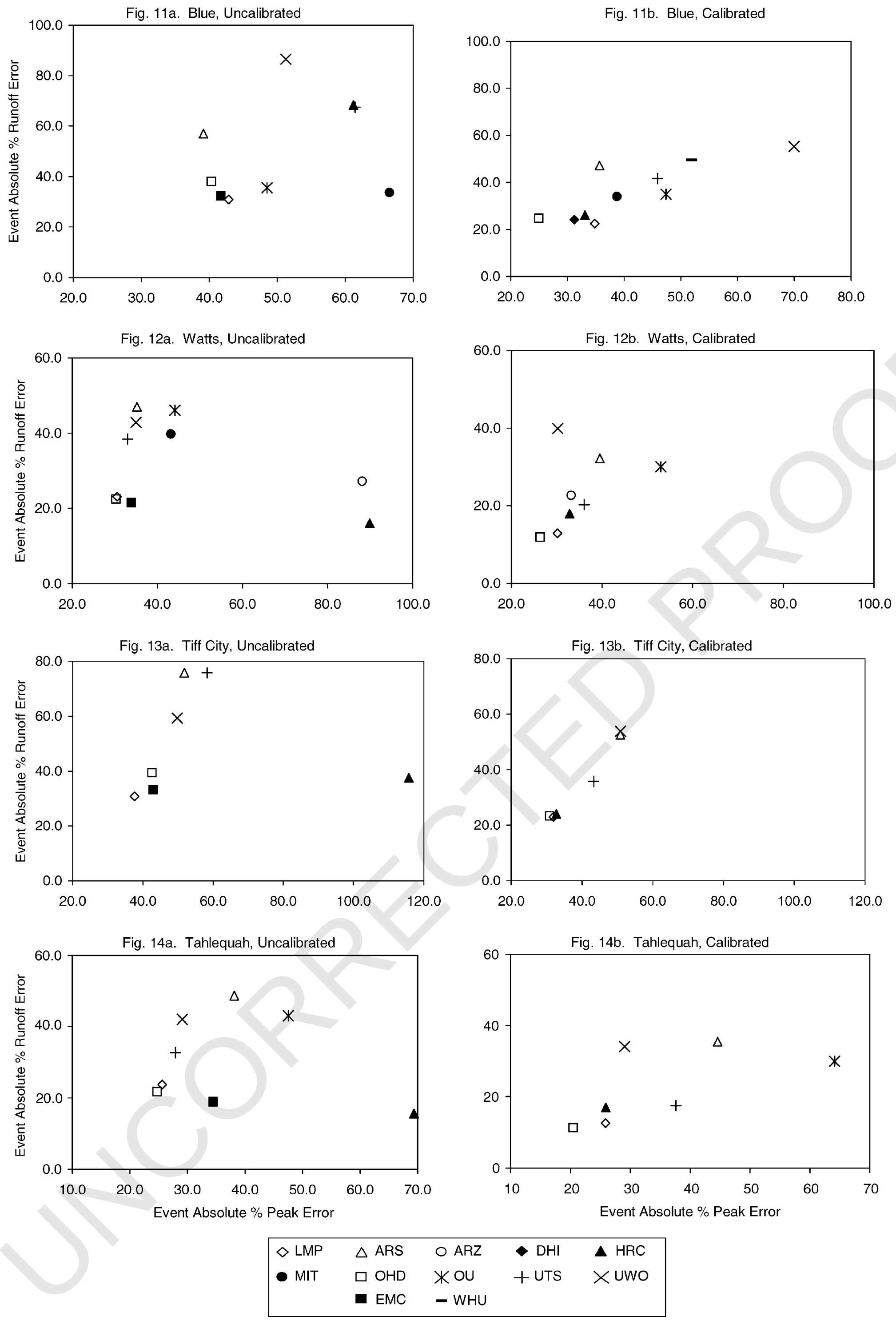

Fig. 7–14. (*continued*)

Table 11
Event percent absolute runoff error used for Figs. 6–13

	Christie	Kansas	Savoy4	Eldon	Blue	Watts4	Tiff City	Tahlequah
Uncalibrated								
LMP	32.4	26.9	29.1	30.2	30.9	23.1	30.8	23.7
ARS	93.8	66.1	30.4	46.3	57.0	47.0	75.8	48.7
ARZ			65.0			27.2		
EMC	37.3	31.5	17.1	45.0	32.3	21.5	33.1	18.8
HRC		26.5	17.9	25.5	68.3	16.1	37.5	15.6
MIT			43.7		33.7	39.8		
OHD	34.8	26.8	28.3	27.4	38.1	22.5	39.4	21.7
OU		70.0			35.5			43.0
UTS	74.5	39.5	39.3	31.7	67.5	38.4	75.8	32.7
UWO	72.5	49.7	42.0	38.1	86.5	42.9	59.3	42.0
Calibrated								
LMP	52.8	23.7	21.1	18.5	22.5	12.9	22.9	12.6
ARS	63.7	49.7	26.9	42.3	47.2	32.2	52.6	35.4
ARZ			48.2			22.7		
DHI					24.2			
HRC		27.1	16.0	20.9	26.1	18.0	24.0	17.0
MIT	46.8			45.1	34.0			
OHD	55.4	23.8	19.9	16.4	24.7	11.9	23.3	11.3
OU		55.2			35.0			29.9
UTS	31.4	26.1	24.7	25.8	41.6	20.3	35.7	17.5
UWO	56.6	36.8	45.1	34.2	55.3	39.9	53.8	34.1
WHU					49.5			

a significant number of cases when distributed models perform at a level close to or better than the lumped model.

Among calibrated models applied to multiple basins, no one model was able to produce positive improvements for all types of statistics (flood runoff, peak flow, and peak time) in all basins; however, the OHD model exhibited positive improvements in peak flow for all basins. The largest percentage gains and the most numerous cases with gains from distributed models are in predicting the peak flows for the Blue River and Christie (Fig. 15b). Three models (OHD, DHI, and HRC) showed peak flow improvement for the Blue River and four models (UTS, UWO, OHD, and MIT) showed peak flow improvement for Christie. Among the parent basins in DMIP, the Blue River has distinguishable shape, orientation, and soil characteristics (See Smith et al. 2004b; Zhang et al., 2003). One possible explanation for the improved calibrated, peak flow results in Christie is that the lumped 'calibrated' model parameters (from the parent basin calibration) are scale dependent and will not outperform parameters that account for spatial variability in the basin if transferred directly from a parent basin to interior points without adjustment.

Fig. 16a–c show flood runoff, peak flow, and peak time improvement for uncalibrated distributed models relative to the uncalibrated lumped model. As with the calibrated models, there are more model-basin combinations when a lumped model outperforms a distributed model (negative improvement) than when a distributed model outperforms a lumped model. There are 56 model-basin cases plotted in each of Fig. 16a–c. Flood runoff improvement is positive in 22% of cases, peak flow improvement positive in 25% of cases, and peak time improvement positive in 24% of cases. The percent of cases with improvement statistics greater than or equal to -5% is 40% for flood runoff and 45% for peak flow, and in 25% of cases, peak time improvements are greater than -1 h. The percentage of cases in which improvement is seen from uncalibrated lumped to uncalibrated distributed models is similar to the percentage of cases where improvement was seen from calibrated lumped to

Table 12
Event percent absolute peak error used for Figs. 6–13

	Christie	Kansas	Savoy4	Eldon	Blue	Watts4	Tiff City	Tahlequah
Uncalibrated								
LMP	67.1	57.1	54.5	53.4	42.8	30.5	37.6	25.6
ARS	246.3	106.1	52.2	49.6	39.2	35.2	51.8	38.1
ARZ			104.3			88.2		
EMC	55.9	63.9	76.4	68.6	41.7	33.9	43.0	34.5
HRC		72.9	67.2	32.2	61.2	89.9	115.8	69.3
MIT			62.4		66.5	43.2		
OHD	88.3	52.8	49.4	45.3	40.3	30.3	42.6	24.7
OU		62.1			48.5			47.5
UTS	59.4	62.3	69.7	43.9	61.4	33.1	58.3	27.9
UWO	75.9	61.8	69.1	58.0	51.2	35.0	49.8	29.1
Calibrated								
LMP	126.0	55.8	52.0	26.0	34.8	30.2	31.9	25.8
ARS	191.5	78.7	56.2	55.9	35.7	39.5	50.9	44.6
ARZ			41.1			33.2		
DHI					31.2			
HRC		53.2	47.4	35.3	33.1	32.9	32.8	25.9
MIT	96.4			54.1	38.7			
OHD	115.0	53.0	49.0	25.8	25.0	26.4	30.8	20.5
OU		64.9			47.4			64.1
UTS	59.0	65.9	67.0	41.0	45.9	36.1	43.3	37.6
UWO	74.9	63.9	64.5	54.6	70.0	30.2	50.8	29.0
WHU					51.9			

calibrated distributed. Note that the performance of the uncalibrated lumped model (and the OHD uncalibrated model) is governed in a large part by the a-priori SAC-SMA parameter estimation procedures defined by Koren et al. (2003b).

An interesting trend in the peak time improvement for both calibrated and uncalibrated results compared to lumped results (Figs. 15c and 16c) is that less improvement is achieved in larger basins (basins are listed from left to right in order of increasing drainage area on the *x*-axis). In fact, none of the distributed models outperform the lumped models in predicting peak time for the three largest basins. Although a definitive reason for this cannot be identified from the analyses done for this paper, one causative factor to consider from our experience in running the OHD distributed model is that the predicted peak time from a physically based routing scheme (with velocities dependent on flow rate) is more sensitive to errors in runoff depth estimation from soil moisture accounting than a linear (e.g. unit hydrograph) routing scheme with constant velocities at all flow levels. Therefore, if runoff is overestimated, the distributed model would tend to predict an earlier peak and if the volume is underestimated the distributed model would tend to predict a later peak, while the unit hydrograph would predict the same peak time regardless of runoff depth. This factor would likely have a greater impact in larger basins.

Fig. 17a–c summarize the improvements gained from calibration. Fig. 17a shows flood runoff improvement gained by calibration for each model in each basin, Fig. 17b shows the peak flow improvement, and Fig. 17c shows the peak time improvement. There are 53 points (model-basin combinations) shown in each of Fig. 17a–c. The majority of points show gains from calibration. Positive flood runoff improvement is seen for 91% of the cases shown, positive peak flow improvement is attained in 66% of the cases, and positive peak time improvement is seen in 70% of the cases.

An interesting note about the OHD results shown in Fig. 17a–c is that this distributed model showed, in some cases, comparable or greater improvements due

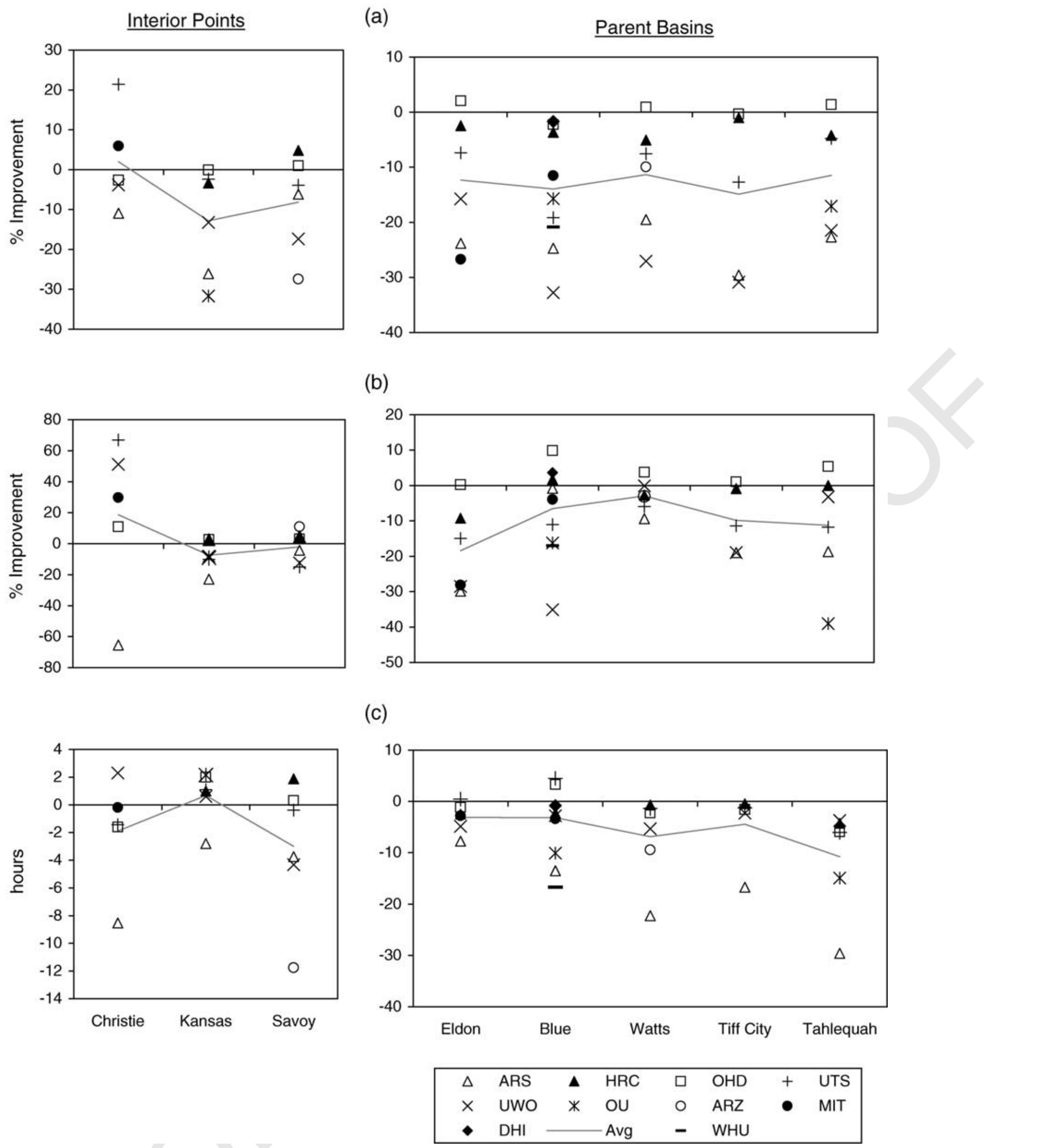

Fig. 15. Distributed results compared to lumped results for calibrated models. (a) Flood runoff improvement, (b) flood peak improvement, and (c) peak time improvement.

to calibration compared with the lumped model. This occurs even though calibration procedures for distributed models are not as well defined and significantly less effort was put into the OHD distributed model calibrations than the lumped model calibrations for DMIP. Although other distributed models also show greater improvement after calibration than the lumped model, this may be due to large differences in uncalibrated parameter estimation procedures. The comparison is more pertinent for the OHD model because the OHD and lumped models use the same rainfall–runoff algorithm (SAC-SMA) and the same estimation scheme for the uncalibrated SAC-SMA parameters.

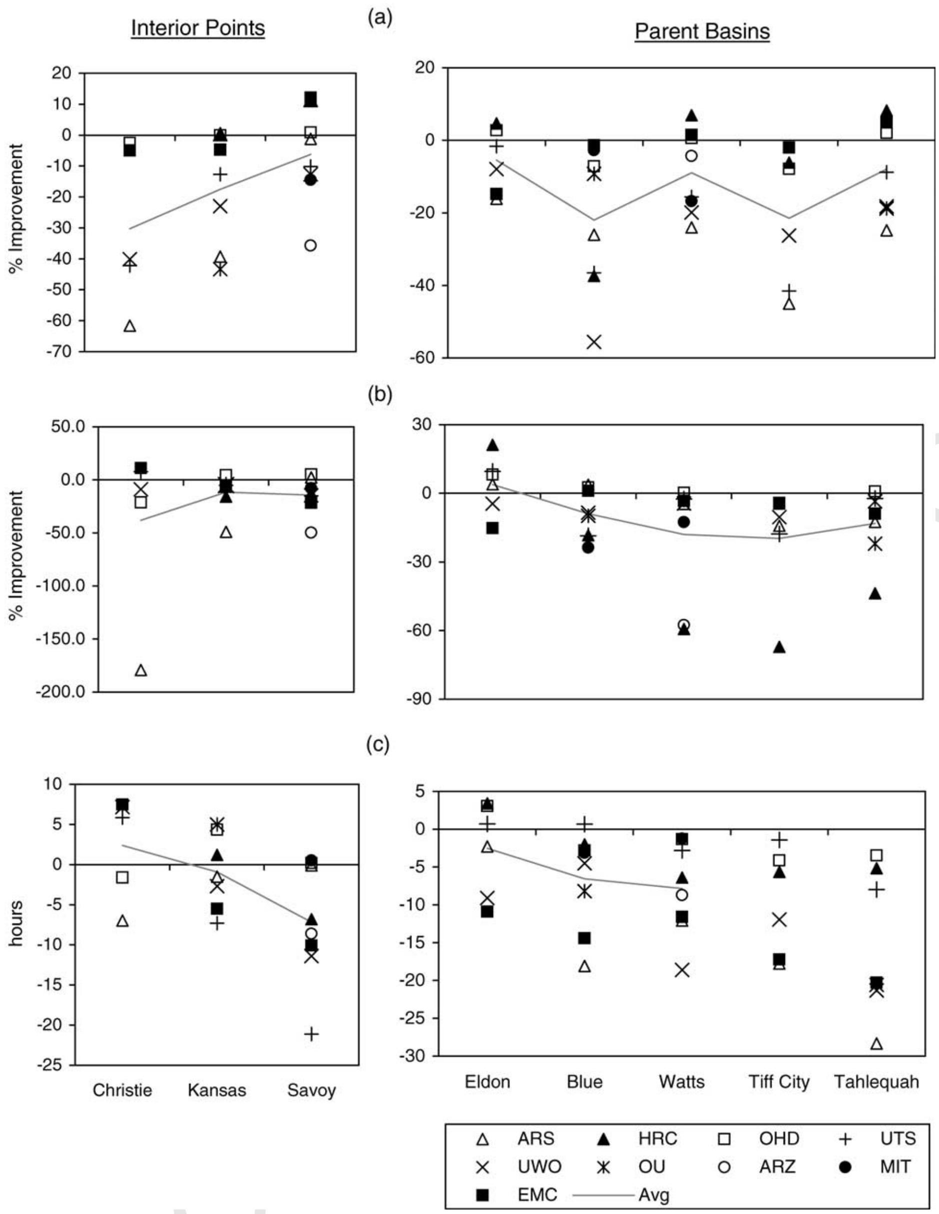

Fig. 16. Distributed results compared to lumped results for uncalibrated models. (a) Flood runoff improvement, (b) flood peak improvement, and (c) peak time improvement.

Each data point shown in Figs. 15–17 is an aggregate measure of the performance of a specific model in a specific basin for many events. Data used to produce Figs. 15–17 are summarized in Tables 14–16. Plotting all of the statistical results for all the events, all basins, and all models would be too lengthy for this paper. However, a few plots showing results for individual events are included here to illustrate the significant scatter in model performance on different events.

Fig. 18a (uncalibrated) and b (calibrated), plots of the peak flow errors from the distributed model versus the peak flow errors from the lumped model for the Eldon basin, show significant scatter. Each point

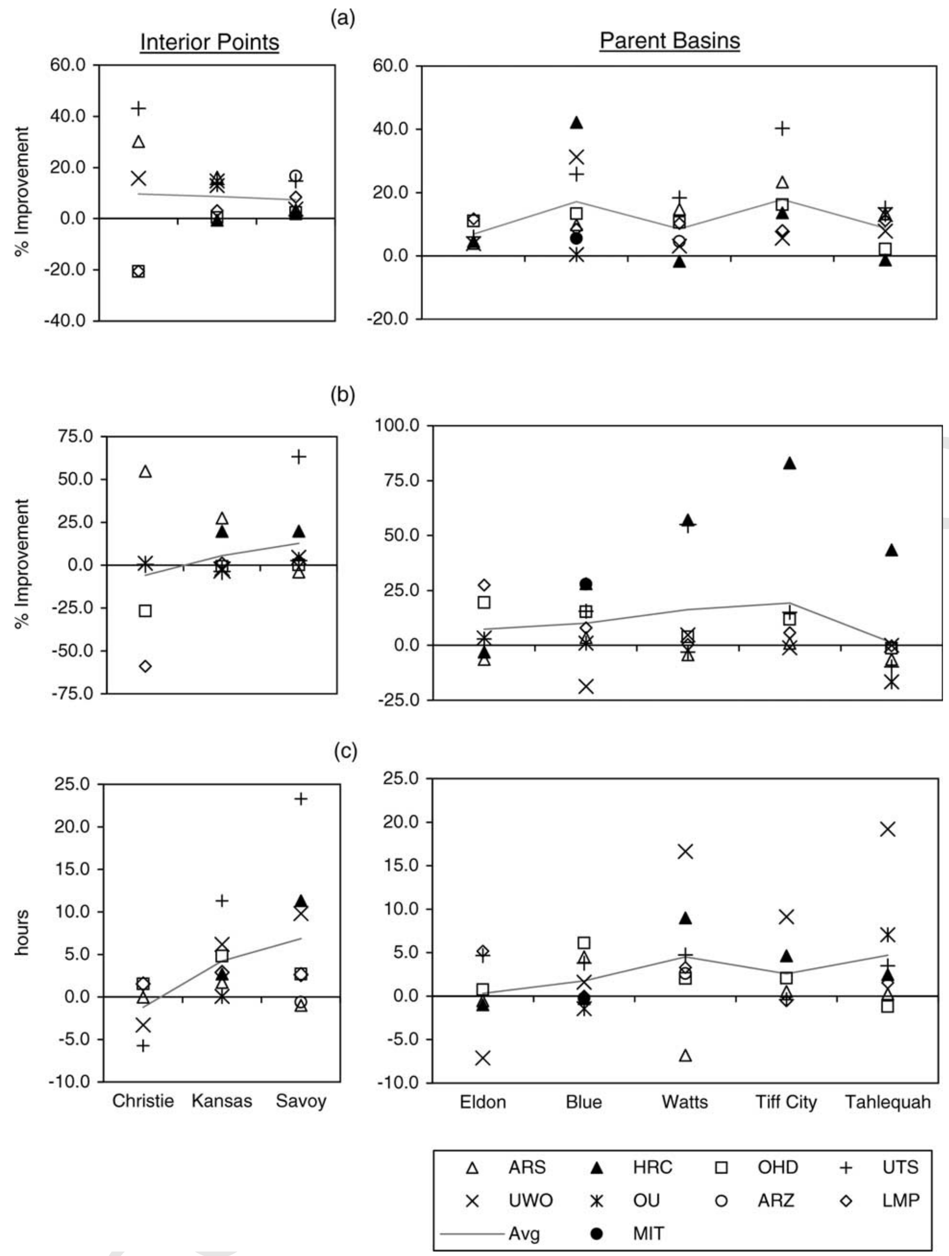

Fig. 17. Calibrated results compared to uncalibrated results. (a) Flood runoff improvement, (b) flood peak improvement, and (c) peak time improvement.

represents a result for a single model and a single event. For points below the 45 degree line, the distributed model outperforms the lumped model. For Eldon, it is interesting to see more cases with gains going from uncalibrated lumped to uncalibrated distributed than going from calibrated lumped to calibrated distributed. Eldon is somewhat unusual in this regard, as indicated by the results in Figs. 15b and 16b. Perhaps in the case of Eldon spatial variability is an important factor in runoff generation but less

Table 13
Event percent runoff bias

	Christie	Kansas	Savoy4	Eldon	Blue	Watts4	Tiff City	Tahlequah
Calibrated								
LMP	49.1	−0.5	−10.5	−2.1	7.3	−0.8	11.4	−2.1
ARS	35.3	0.1	24.1	−18.0	35.1	−8.1	10.7	−11.5
ARZ	–	–	33.7	–	–	1.2	–	–
DHI	–	–	–	–	−10.8	–	–	–
HRC		13.3	−1.4	−7.1	6.0	4.8	11.2	9.5
MIT	−4.6			−37.9	−23.0			
OHD	52.7	1.2	−8.7	0.3	14.6	1.5	14.3	−0.6
OU		−36.8			−20.6			−8.5
UTS	21.6	−11.0	−2.3	−14.1	28.0	−6.9	−9.7	−5.8
UWO	53.7	27.5	12.3	−6.7	49.2	21.3	33.1	18.8
WHU					11.4			

Table 14
Event improvement statistics: distributed results compared to lumped results for calibrated models

	ARS	HRC	OHD	UTS	UWO	OU	ARZ	MIT	DHI	WHU
Flood runoff										
Christie	−10.9		−2.6	21.4	−3.9			6.0		
Kansas	−26.1	−3.4	−0.1	−2.4	−13.2	−31.7				
Savoy	−6.2	4.8	1.0	−3.9	−17.4		−27.4			
Eldon	−23.8	−2.5	2.1	−7.4	−15.7			−26.7		
Blue	−24.7	−3.6	−2.3	−19.2	−32.8	−15.8		−11.5	−1.7	−20.9
Watts	−19.5	−5.1	0.9	−7.5	−27.1		−9.9			
Tiff City	−29.6	−1.0	−0.3	−12.7	−30.9					
Tahlequah	−22.7	−4.2	1.4	−4.8	−21.4	−17.1				
Flood Peak										
Christie	−65.4		11.0	67.0	51.1			29.7		
Kansas	−22.9	2.6	2.8	−10.1	−8.1	−9.2				
Savoy	−4.2	4.6	3.0	−15.0	−12.5		10.9			
Eldon	−29.9	−9.3	0.3	−15.0	−28.6			−28.1		
Blue	−0.8	1.7	9.9	−11.1	−35.1	−16.2		−3.9	3.6	−13.6
Watts	−9.4	−2.7	3.8	−5.9	−0.1		−3.1			
Tiff City	−19.0	−0.9	1.1	−11.4	−18.9					
Tahlequah	−18.7	0.0	5.4	−11.8	−3.2	−39.0				
Peak time										
Christie	−8.5		−1.6	−1.4	2.3			−0.2		
Kansas	−2.8	1.0	2.0	1.1	0.6	2.2				
Savoy	−3.8	1.9	0.3	−0.4	−4.3		−11.8			
Eldon	−7.8	−2.5	−1.1	0.5	−4.8			−2.8		
Blue	−13.5	−2.3	3.3	4.5	−2.8	−10.1		−3.4	−0.8	−16.7
Watts	−22.3	−0.7	−2.2	−1.4	−5.3		−9.4			
Tiff City	−16.7	−0.5	−1.5	−1.3	−2.3					
Tahlequah	−29.6	−4.2	−5.9	−6.0	−3.7	−15.0				

Table 15
Event improvement statistics: distributed results compared to lumped results for uncalibrated models

	ARS	HRC	OHD	UTS	UWO	OU	ARZ	MIT	EMC
Christie	−61.6		−2.5	−42.2	−40.1				−5.0
Kansas	−39.3	0.3	−0.1	−12.7	−23.0	−43.3			−4.7
Savoy	−1.2	11.3	0.9	−10.2	−12.7		−35.7	−14.5	12.1
Eldon	−16.1	4.7	2.8	−1.6	−7.9				−14.9
Blue	−26.1	−37.4	−7.1	−36.6	−55.6	−9.8		−2.8	−1.3
Watts	−24.0	6.9	0.6	−15.6	−19.9		−4.3	−16.8	1.5
Tiff City	−45.0	−6.1	−7.9	−41.5	−26.3				−2.0
Tahlequah	−24.8	8.2	2.0	−8.8	−18.2	−18.8			4.9
Christie	−179.2		−21.2	7.7	−8.8				11.2
Kansas	−49.0	−15.8	4.3	−5.2	−4.7	−5.0			−6.8
Savoy	2.3	−12.7	5.1	−15.2	−14.6		−49.8	−7.8	−21.9
Eldon	3.9	21.2	8.1	9.5	−4.6				−15.2
Blue	3.7	−18.4	2.5	−18.6	−8.4	−10.4		−23.7	1.1
Watts	−4.7	−59.4	0.3	−2.5	−4.4		−57.6	−12.6	−3.3
Tiff City	−14.2	−67.2	−4.3	−17.7	−10.4				−4.6
Tahlequah	−12.5	−43.7	0.9	−2.3	−3.5	−22.0			−8.9
Christie	−7.0		−1.6	5.9	7.1				7.5
Kansas	−1.5	1.2	4.4	−7.3	−2.7	5.0			−5.5
Savoy	−0.1	−6.8	0.2	−21.1	−11.4		−8.6	0.5	−10.1
Eldon	−2.3	3.4	3.0	0.7	−9.1				−10.9
Blue	−18.1	−2.0	−2.8	0.7	−4.5	−8.2		−3.1	−14.4
Watts	−12.1	−6.4	−1.3	−2.8	−18.6		−8.7	−1.2	−11.6
Tiff City	−17.8	−5.6	−4.1	−1.4	−11.9				−17.2
Tahlequah	−28.3	−5.2	−3.5	−8.0	−21.3	−20.6			−20.3

important in affecting hydrograph shape so the lumped calibration is able to account for the spatially variable runoff generation, leaving less potential for gains from distributed runoff and routing in the calibrated case.

We infer based on DMIP results and other results reported in the literature (Zhang et al., 2003; Koren et al., 2003a; Smith et al., 2004a) that spatially variability of rainfall does have a big impact on hydrograph shape in the Blue River and this is why noticeable gains are achieved by running a distributed model. Similar to Fig. 18a and b; Fig. 19a (uncalibrated) and 19b (calibrated) show the peak flow errors from distributed models versus the peak flow errors from the lumped model, but for the Blue basin. However, to remove some of the scatter and emphasize the significant improvements possible for the Blue river basin, only results from the three best performing models (in terms of event peak flows for Blue) are plotted.

To force the same domain and range for plotting in Figs. 18 and 19, the plotting range is defined by the range of errors that existed in the lumped model simulations. Since the maximum errors for distributed models are greater than the maximum errors for lumped models, some data points are not seen in Figs. 18 and 19.

3.4. Additional analysis for interior points

One of the big benefits of using distributed models is that they are able to produce simulations at interior points; however, studies are needed to quantify the accuracy and uncertainty of interior point simulations. Streamflow data from a limited number of interior points were provided in DMIP. These interior points include Watts (given calibration at Tahlequah), Savoy, Kansas, and Christie. Based on the presentation and discussion of overall and event-based statistics above, it is seen that some models are able to

Table 16
Event improvement statistics: calibrated results compared to uncalibrated results

	ARS	HRC	OHD	UTS	UWO	OU	ARZ	LMP	MIT
Flood runoff									
Christie	30.2		−20.6	43.1	15.8			−20.5	
Kansas	16.3	−0.6	0.5	13.4	12.9	14.8		3.1	
Savoy	3.4	2.0	2.4	14.7	3.8		16.7	8.4	
Eldon	4.0	4.5	11.0	6.0	3.9			11.7	
Blue	9.8	42.2	13.3	25.8	31.2	0.5		8.4	5.5
Watts	14.7	−1.7	10.5	18.3	3.1		4.5	10.2	
Tiff City	23.3	13.6	16.2	40.3	5.6			7.9	
Tahlequah	13.3	−1.3	2.1	15.1	7.8	13.1		11.1	
Flood peak									
Christie	54.8		−26.7	0.4	1.0			−58.9	
Kansas	27.5	19.7	−0.7	−3.6	−2.1	−2.9		1.3	
Savoy	−4.0	19.8	0.1	2.7	4.6		63.2	2.5	
Eldon	−6.3	−3.1	19.5	2.9	3.4			27.4	
Blue	3.5	28.1	15.4	15.5	−18.7	1.1		8.0	27.8
Watts	−4.3	57.1	3.9	−3.0	4.8		54.9	0.4	
Tiff City	1.0	83.1	11.8	15.0	−1.0			5.7	
Tahlequah	54.8		−26.7	0.4	1.0			−58.9	
Peak time									
Christie	0.0		1.5	−5.8	−3.3			1.5	
Kansas	1.7	2.7	4.8	11.3	6.2	0.1		2.9	
Savoy	−1.0	11.3	2.7	23.3	9.8		−0.6	2.625	
Eldon	−0.5	−1.0	0.8	4.7	−7.1			5.2	
Blue	4.5	−0.3	6.1	3.8	1.6	−1.5		0.0	−0.3
Watts	−6.8	9.0	2.0	4.7	16.6		2.6	3.3	
Tiff City	0.53	4.65	2.06	−0.41	9.12			−0.53	
Tahlequah	0.2	2.5	−1.2	3.5	19.2	7.1		1.5	

produce reasonable simulations for these interior points, although errors are typically greater than for parent basins.

Another question that can be investigated with DMIP data is whether a model calibrated at a smaller basin (Watts) shows advantages in simulating flows at a common interior point with a model calibrated at a larger parent basin (Tahlequah). One of the tests requested in the DMIP modeling instructions (instruction 4) was for modelers to calibrate models at Watts and submit the resulting simulations for both Watts and two interior points (Savoy and an ungauged point) without using interior flow information. Modeling instruction 5 requested that the same be done for Tahlequah, with interior simulations generated at Watts, Savoy, and Kansas. For the common points (Watts and Savoy) from instructions 4 and 5, Figs. 20 and 21 compare the event percent absolute runoff error and percent absolute peak error statistics. Points above the 1:1 line indicate improvement after calibration at Watts. For the percent absolute runoff error results (Figs. 20a and 21a), none of the models showed significant improvement after calibration at Watts. This is perhaps not surprising considering the conclusion from the lumped calibration of Tahlequah and Watts that the same SAC-SMA parameter set produces reasonable results in both basins. For the peak flow error results, only the UTS model showed improvement.

Simulations were also requested at several ungauged interior points. One way to examine these results in the absence of observed streamflow data is to compare coefficients of variation (CVs) from different models. Simulated (calibrated) and observed CVs for flow are plotted against drainage area in Fig. 22a and b. The area range plotted in Fig. 22a encompasses all of

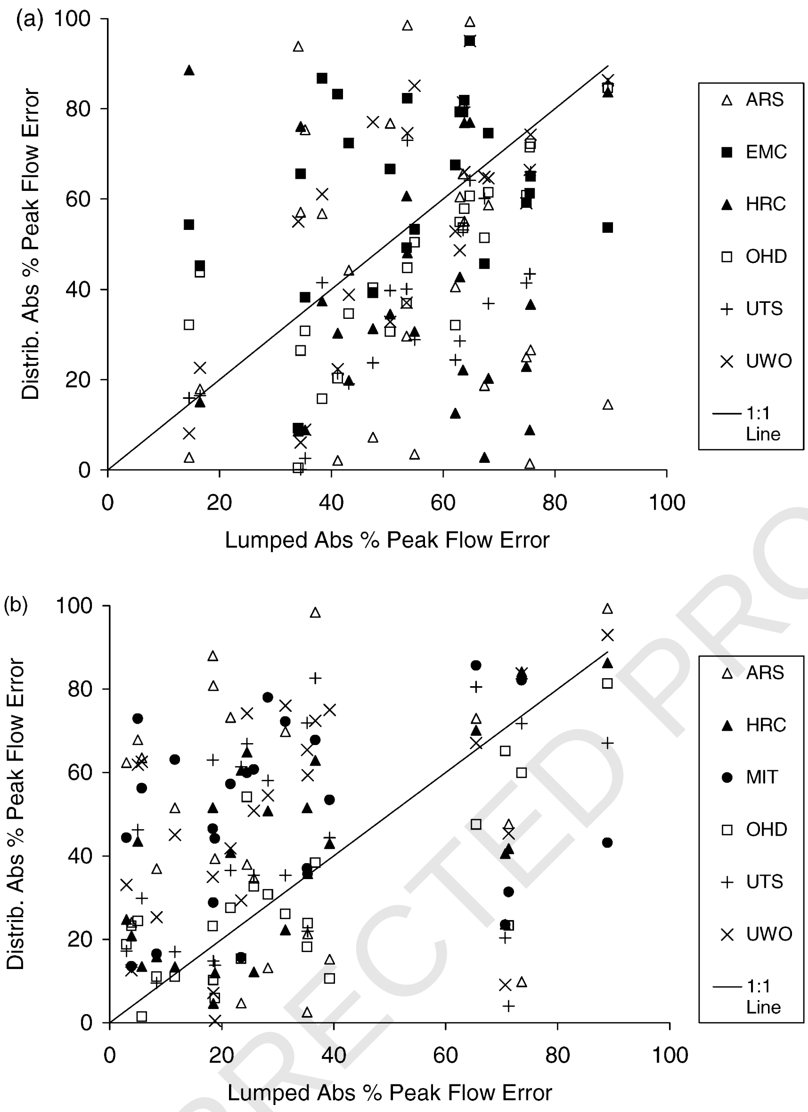

Fig. 18. Distributed percent absolute peak flow errors vs. lumped percent absolute peak flow errors for Eldon events: (a) uncalibrated and (b) calibrated models.

the DMIP basins while Fig. 22b provides a more detailed look at results for smaller basins. In Fig. 22a, the LMP, OHD, and HRC models reasonably approximate the trend of increasing CV with decreasing drainage area over the scales of most DMIP basins. It is not possible to infer much about the accuracy of simulated CV values for the range of scales shown in Fig. 22b because only one point with observed data (Christie at 65 km^2) is available. However, it is interesting that the UTS model, which had the best percent absolute runoff error and peak flow statistics for Christie among calibrated models, tends to underestimate the CV for Christie, as it does for the larger basins with observed data. It turns out that the standard deviation of flows predicted by the UTS model for Christie is close to that of the observed data but the mean flow predicted by the UTS model is too high, due primarily to high modeled base flows.

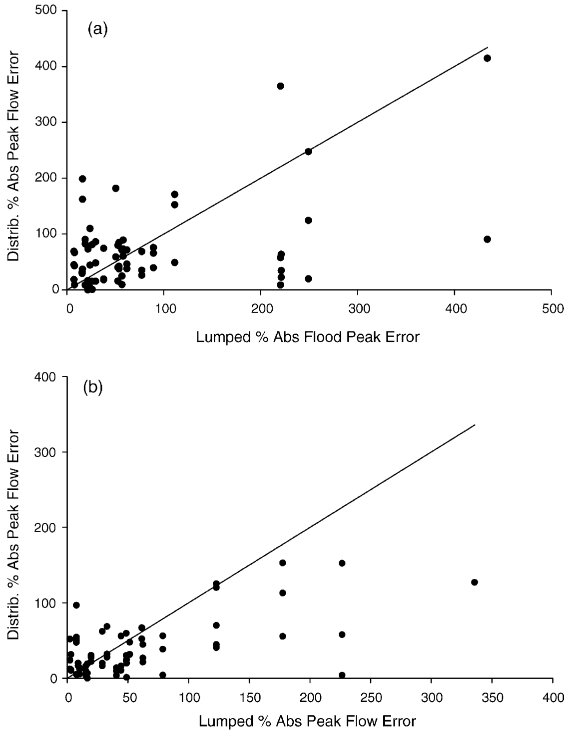

Fig. 19. Distributed percent absolute peak flow errors vs. lumped percent absolute peak flow errors for Blue events: (a) uncalibrated and (b) calibrated models. Data shown are for the three distributed models with the lowest average absolute peak flow simulation error for Blue.

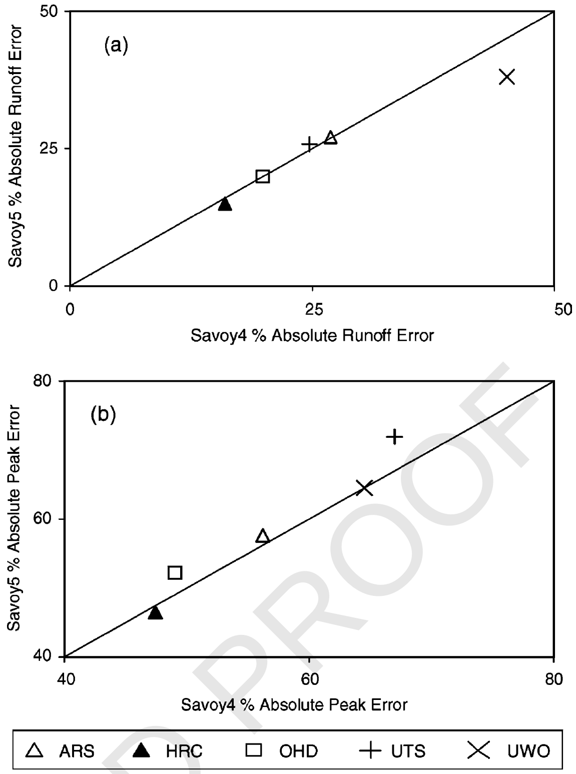

Fig. 20. Comparisons of results at Savoy from initial calibrations at Tahlequah (instruction 5) and Watts (instruction 4): (a) event percent absolute runoff error and (b) event percent absolute peak flow error.

4. Conclusions

A major goal of DMIP is to understand the capabilities of existing distributed modeling methods and identify promising directions for future research and development. The focus of this paper is to evaluate and intercompare streamflow simulations from existing distributed hydrologic models forced with operational NEXRAD-based precipitation data. A significant emphasis in the analysis is on comparisons of distributed models to lumped model simulations of the type currently used for operational forecasting at RFCs.

The key findings are as follows:

- Although the lumped model outperformed distributed models in more cases than distributed models outperformed the lumped model, some calibrated distributed models can perform at a level comparable to or better than a calibrated lumped model (the current operational standard). The wide range of accuracies among model results suggest that factors such as model formulation, parameterization, and the skill of the modeler can have a bigger impact on simulation accuracy than simply whether or not the model is lumped or distributed.
- Clear gains in distributed model performance can be achieved through some type of model calibration. On average, calibrated models outperformed uncalibrated models during both the calibration and validation periods.
- Gains in predicting peak flows for calibrated models (Fig. 15b) were most noticeable in the Blue and Christie basins. The Blue basin has distinguishable shape, orientation, and soil characteristics from other basins in the study. The Blue results are consistent with those of previous studies cited in Section 1 and indicate that the gains from

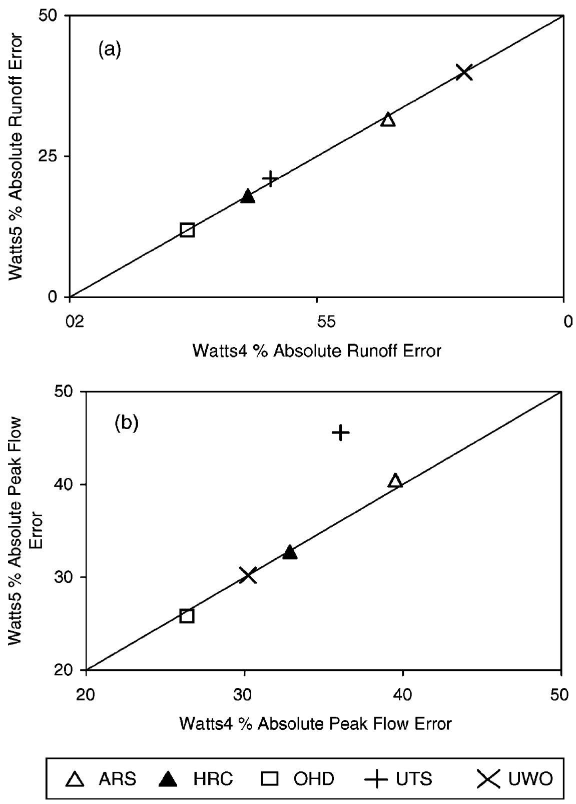

Fig. 21. Comparisons of results at Watts from initial calibrations at Tahlequah (instruction 5) and Watts (instruction 4): (a) event percent absolute runoff error and (b) event percent absolute peak flow error.

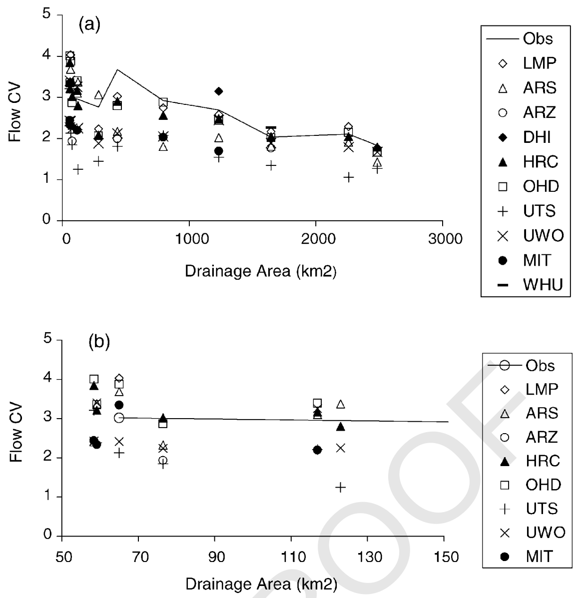

Fig. 22. Flow coefficients of variation for observed flows (solid line) and modeled flows (for both gaged and ungaged locations): (a) all basin sizes and (b) a closer look at the small basins.

applying a distributed simulation model at NWS forecast basin scales (on the order of 1000 km^2) will depend on the basin characteristics. Christie is distinguishable in this study because of its small size.

- Christie had distinguishable results from the larger basins studied, not just in overall statistics, but in relative inter-model performance compared with larger basins. One explanation offered for the improved calibrated, peak flow results (Fig. 15b) is that the lumped 'calibrated' model parameters (from the parent basin calibration, Eldon) are scale dependent and distributed model parameters that account for spatial variability within Eldon are less scale dependent. Some caution is advised in interpreting the results for Christie for model submissions with a relatively coarse cell resolution compared to the size of the basin (e.g. EMC and OHD). Since no other basins in DMIP are comparable in size to Christie, more studies on small, nested basins are needed to confirm and better understand these results.
- Among calibrated results, models that combine techniques of conceptual rainfall–runoff and physically based distributed routing consistently showed the best performance in all but the smallest basin. Gains from calibration indicate that determining reasonable a priori parameters directly from physical characteristics of a watershed is generally a more difficult problem than defining reasonable parameters for a conceptual lumped model through calibration.
- Simulations for smaller interior basins where no explicit calibration was done exhibited reasonable performance in many cases, although not as good statistically as results for larger, parent basins. The relatively degraded performance in smaller basins occurred both in cases when parent basins were calibrated and when they were uncalibrated, so the degraded performance was not simply a function of the fact that no explicit calibration at interior points was allowed.

- Distributed models designed for research can be applied successfully using operational quality data. Several models responded similarly to long term biases in archived multi-sensor precipitation grids. Ease of implementation could not be measured directly. However, an indirect indicator operational practicability is that several participants were able to submit a full set or nearly a full set of simulations (Table 2) with no financial support and in a relatively short time.

This study did not address the question of whether or not simulation model improvements will translate into operational forecast improvements. One important issue in operational forecasting is the use of forecast precipitation data. Because forecast precipitation data have a lower resolution and are much more uncertain than the observed precipitation used in this study, the benefits of distributed models may diminish for longer lead times that rely more heavily on forecast precipitation data. This assumption needs further study, but if true, greater benefits from distributed models would be expected for shorter lead times that are close to the response time of a basin. For example, analysis of several isolated storms in the Blue River indicates an average time between the end of rainfall and peak streamflow of about 9 h and an average time between the rainfall peak and the streamflow peak of about 18 h. Forecasts in this range of lead times could benefit without using any forecast precipitation.

5. Recommendations

The analyses in this paper addressed the following questions: Can distributed models exhibit simulation performance comparable to or better than existing lumped models used in the NWS? Are there differences in relative model performance when different distributed models are applied to different basins? Does calibration improve the performance of distributed models? The results also help to formulate useful questions that merit further investigation. For example: Why does one particular model perform relatively well in one basin but not as well in another basin? Because the widely varying structural components in participating models (e.g. different rainfall–runoff algorithms, routing algorithms, and model element sizes) have interacting and compensating effects, it is difficult to infer reasons for differences in model performance. More controlled studies in which only one model component is changed at a time will be required to answer questions related to causation.

Much work lies ahead to gain a clearer and deeper understanding of the results presented in this paper. Several other papers in this issue already begin to examine the underlying reasons for our results. Scale and uncertainty issues figure to be critical research topics that will require further study. An important potential benefit of using distributed models is the ability to produce simulations at small, ungauged locations. However, given uncertainty in available inputs, the spatial and temporal scales where explicit distributed modeling can provide the most useful products (and benefits relative to lumped modeling) is not clear. Forecasters will need guidance to define the confidence they should have in forecasts at various modeling scales. This is true for both lumped and distributed models. A recent NWS initiative to produce probabilistic quantitative precipitation estimates (PQPE) should help support this type of effort. Information about precipitation uncertainty can be incorporated into hydrologic forecasts through the use of ensemble simulations (e.g. Carpenter and Georgakakos, 2004).

Concurrent with future studies to improve our understanding, efforts are also needed to develop software that can test these techniques in an operational forecasting environment. All results presented in this paper were produced in an off-line simulation mode. Design for the forecasting environment raises a number of scientific and software issues that were not addressed directly in this paper. Issues such as model run-times, ease of use, and ease of parameterization are very important for successful operational implementation. Related issues to consider are capabilities to ingest both observed and forecast precipitation, update model states, and produce ensemble forecasts as necessary. A project to create and test an operational version of the OHD distributed model is currently in progress.

Finally, several ideas for future intercomparison work (e.g. DMIP Phase II) were suggested at the August 2002 DMIP workshop. These suggestions included defining a community-wide distributed modeling system, separating the comparisons of routing and rainfall runoff techniques, using synthetic simulations to complement work with real world data, doing more uncertainty analysis (e.g. ensemble simulations), looking in more detail at differences in model structures to improve our understanding of cause and effect, assessing the impact of model element size in a more systematic manner, identifying additional basins where scale issues can be studied effectively and where other processes such as snow modeling can be investigated, using additional sources of observed data for model verification (e.g. soil moisture), and using a longer verification period.

Appendix A

DMIP Participants: Jeff Arnold[b], Christina Bandaragoda[c], Allyson Bingeman[d], Rafael Bras[e], Michael Butts[f], Theresa Carpenter[g], Zhengtao Cui[h], Mauro Diluzio[i], Konstantine Georgakakos[g], Anubhav Gaur[h], Jianzhong Guo[l], Hoshin Gupta[j], Terri Hogue[j], Valeri Ivanov[e], Newsha Khodatalab[j], Li Lan[k], Xu Liang[l], Dag Lohmann[m], Ken Mitchell[m], Christa Peters-Lidard[m], Erasmo Rodriguez[d], Frank Seglenieks[d], Eylon Shamir[j], David Tarboton[c], Baxter Vieux[h], Enrique Vivoni[e], and Ross Woods[n]

1. Office of Hydrologic Development, NOAA/NWS, Silver Spring, Maryland
2. USDA-Agricultural Research Service, Temple, Texas
3. Utah State University, Logan, Utah
4. University of Waterloo, Ontario, Canada
5. Massachusetts Institute of Technology, Cambridge, Massachusetts
6. DHI Water and Environment, Horsholm, Denmark
7. Hydrologic Research Center, San Diego, California
8. University of Oklahoma, Norman, Oklahoma
9. TAES-Blacklands Research Center, Temple, Texas
10. University of Arizona, Tucson, Arizona
11. Wuhan University, Wuhan, China
12. University of California at Berkeley, Berkeley, California
13. NOAA/NCEP, Camp Springs, Maryland

References

Anderson, E., (2003). Calibration of Conceptual Hydrologic Models for Use XSC DC DC XAAAQQver Forecasting (copy available on request from: Hydrology Laboratory, Office of Hydrologic Development, NOAA/National Weather Service, (1325) East-West Highway, Silver Spring, MD 20910).

Andersen, J., Refsgaard, J.C., Jensen, H.J., 2001. Distributed hydrological modeling of the senegal river basin-model construction and validation. Journal of Hydrology 247, 200–214.

Bandaragoda, C., Tarboton, D., Woods, R., 2004. Application of topmodel in the distributed model intercomparison Project. Journal of Hydrology, xxthis issue.

Boyle, D.P., Gupta, H.V., Sorooshian, S., Koren, V., Zhang, Z., Smith, M., 2001. Toward Improved Streamflow Forecasts: Value of Semi-distributed Modeling. Water Resources Research 37(11), 2749–2759.

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

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

Butts, M.B., Payne, J.T., Kristensen, M., Madsen, H., 2004. An Evaluation of the impact of model structure and complexity on hydrologic modelling uncertainty for streamflow prediction. Journal of Hydrology this issue.

Carpenter, T.M., Georgakakos, K.P., 2004. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. Journal of Hydrology this issue.

Carpenter, T.M., Georgakakos, K.P., Spersflagea, J.A., 2001. On the parametric and NEXRAD-radar sensitivities of a distributed hydrologic model suitable for operational use. Journal of Hydrology 253, 169–193.

Christiaens, K., Feyen, J., 2002. Use of sensitivity and uncertainty measures in distributed hydrological modeling with an application to the MIKE SHE model. Water Resources Research 38(9), 1169.

Clark, C.O., 1945. Storage and the unit hydrograph. Transactions of the American Society of Civil Engineers 110, 1419–1446.

Di Luzio, M., Arnold, J., 2004. Gridded precipitation input toward the improvement of streamflow and water quality assessments. Journal of Hydrology this issue.

Finnerty, B.D., Smith, M.B., Seo, D.J., Koren, V., Moglen, G.E., 1997. Space-time scale sensitivity of the Sacramento model to radar-gage precipitation inputs. Journal of Hydrology 203, 21–38.

Fulton, R.A., Breidenbach, J.P., Seo, D.J., Miller, D.A., O'Bannon, T., 1998. The WSR-88D rainfall algorithm. Weather and Forecasting 13, 377–395.

Guo, J., Liang, X., Leung, L.R., 2004. Impacts of different precipitation data sources on water budget simulated by the VIC-3L hydrological model. Journal of Hydrology this issue.

Gupta, H.V., Sorooshian, S., Hogue, T.S., Boyle, D.P., 2003. In: Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A., Turcotte, R. (Eds.), Advances in Automatic Calibration of Watershed Models, Calibration of Watershed Models, Water Science and Application 6, American Geophysical Union, pp. 9–28.

Havno, K., Madsen, M.N., Dorge, J., 1995. Mike 11—A Generalized River Modelling Package. In: Singh, V.P., (Ed.), Computer Models of Watershed Hydrology, Water Resources Publications, Colorado, USA, pp. 733–782.

Ivanov, V.Y., Vivoni, E.R., Bras, R.L., Entekhabi, D., 2004. Preserving high-resolution surface and rainfall data in operational-scale basin hydrology: a fully-distributed physically-based approach. Journal of Hydrology this issue.

Johnson, D., Smith, M., Koren, V., Finnerty, B., 1999. Comparing mean areal precipitation estimates from NEXRAD and rain gauge networks. Journal of Hydrologic Engneering 4(2), 117–124.

Khodatalab, N., Gupta, H., Wagener, T., Sorooshian, S., 2004. Calibration of a semi-distributed hydrologic model for streamflow estimation along a river system. Journal of Hydrology this issue.

Koren, V, Schaake, J., Duan, Q., Smith, M., Cong, S., September (1998). PET Upgrades to NWSRFS—Project Plan, HRL Internal Report, (copy available on request from: Hydrology Laboratory, Office of Hydrologic Development, NOAA/ National Weather Service, 1325 East-West Highway, Silver Spring, MD 20910).

Koren, V.I., Finnerty, B.D., Schaake, J.C., Smith, M.B., Seo, D.J., Duan, Q.Y., 1999. Scale dependencies of hydrologic models to spatial variability of precipitation. Journal of Hydrology 217, 285–302.

Koren, V., Reed, S., Smith, M., Zhang, Z., Seo, D.J., 2003a. In review, Hydrology Laboratory Research Modeling System (HL-RMS) of the National Weather Service. Journal. of Hydrology.

Koren, V., Smith, M., Duan, Q., 2003b. Use of a priori parameter estimates in the derivation of spatially consistent parameter sets of rainfall–runoff models. In: Duan, Q., Sorooshian, S., Gupta, H., Rosseau, A., Turcotte, R. (Eds.), Advances in the Calibration of Watershed Models, AGU Water Science and Applications Series.

Kouwen, N., Garland, G., 1989. Resolution considerations in using radar rainfall data for flood forecasting. Canadian Journal of Civil Engineering 16, 279–289.

Kouwen, N., Soulis, E.D., Pietroniro, A., Donald, J., Harrington, R.A., 1993. Grouped Response units for distributed hydrologic modelling. Journal of Water Resources Planning and Management 119(3), 289–305.

Leavesley, G.H., Hay, L.E., Viger, R.J., Markstrom, S.L., 2003. Use of a priori parameter-estimation methods to constrain calibration of distributed-parameter models. In: Duan, Q., Sorooshian, S., Gupta, H., Rosseau, A., Turcotte, R. (Eds.), Advances in the Calibration of Watershed Models, AGU Water Science and Applications Series.

Liang, X., Xie, Z., 2001. A new surface runoff parameterization with subgrid-scale soil heterogeneity for land surface models. Advances in Water Resources 24, 1173–1193.

Liang, X., Lettenmaier, D.P., Wood, E.F., Burges, S.J., 1994. A simple hydrologically based model of land surface water and energy fluxes for general circulation models. Journal of Geophysical Research 99(D7), 14,415–14,428.

Madsen, H., 2003. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Advances in Water Resources 26, 205–216.

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

Nash, J.E., Sutcliffe, J.V., 1970. River flow forecasting through conceptual models part I— a discussion of principles. Journal of Hydrology 10, 282–290.

Neitsch, S.L., Arnold, J.G, Kiniry, J.R., Williams, J.R., King, K.W., 2000. Soil and Water Assessment Tool Theoretical Documentation, Version 2000, Texas Water Resources Institute (TWRI), Report TR-191, College Station, TX, 506pp.

Refsgaard, J.C., Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. Water Resources Research 32(7), 2189–2202.

Senarath, S.U.S., Ogden, F.L., Downer, C.W., Sharif, H.O., 2000. On the calibration and verification of two-dimensional, distributed, Hortonian, continuous watershed models. Water Resources Research 36(6), 1510–1595.

Seo, D.-J., Breidenbach, J.P., 2002. Real-time correction of spatially nonuniform bias in radar rainfall using rain gage measurements. J. Hydrometeorology 3, 93–111.

Seo, D.-J., Breidenbach, J.P., Johnson, E.R., 1999. Real-time estimation of mean field bias in radar rainfall data. Journal of Hydrology, 233.

Seo, D.-J., Breidenbach, J.P., Fulton, R.A., Miller, D.A., O'Bannon, T., 2000. Real-time adjustment of range-dependent biases in WSR-88D rainfall data due to nonuniform vertical profile of reflectivity. Journal of Hydrometeorology 1(3), 222–240.

Smith, M.B., Koren, V., Johnson, D., Finnerty, B.D., Seo, D.-J., 1999. Distributed Modeling: Phase 1 Results, NOAA Technical Report NWS 44, National Weather Service Hydrology Laboratory, 210 pp. Copies available upon request.

Smith, M.B., Laurine, D., Koren, V., Reed, S., Zhang, Z., 2003. Hydrologic model calibration in the National Weather Service. In: Duan, Q., Sorooshian, S., Gupta, H., Rosseau, A., Turcotte, R. (Eds.), Advances in the Calibration of Watershed Models, AGU Water Science and Applications Series.

Smith, M.B., Koren, V.I., Zhang, Z., Reed, S.M., Pan, J.-J., Moreda, F., Kuzmin, V., 2004aa. Runoff response to spatial variability in precipitation: an analysis of observed data. Journal of Hydrology this issue.

Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S., Zhang, Z., Duan, Q.-Y., Cong, S., Moreda, F., Anderson, R., 2004bb. The Distributed Model Intercomparison Project (DMIP): an overview. Journal of Hydrology this issue.

Sweeney, T.L., 1992. Modernized Areal Flash Flood Guidance, NOAA Technical Memorandum NWS Hydro 44, Silver Spring, MD.

Vieux, B.E., 2001. Distributed Hydrologic Modeling Using GIS, Water Science and Technology Series, vol. 38. Kluwer, Norwell, MA, 293 pp. ISBN 0-7923-7002-3.

Vieux, B.E., Moreda, F., 2003. Ordered Physics-Based Parameter Adjustment of a Distributed Model. In: Duan, Q., Sorooshian, S., Gupta, H., Rosseau, A., Turcotte, R. (Eds.), Advances in the Calibration of Watershed Models, AGU Water Science and Applications Series.

Wang, D., Smith, M.B., Zhang, Z., Reed, S., Koren, V., 2000. Statistical comparison of mean areal precipitation estimates from WSR-88D, operational and historical gage networks, 15th Conference on Hydrology, AMS, January 9–14, Long Beach, CA.

Young, C.B., Bradley, A.A., Krajewski, W.F., Kruger, A., 2000. Evaluating NEXRAD Multisensor precipitation estimates for operational hydrologic forecasting. Journal of Hydrometeorology 1, 241–254.

Zhang, Z., Koren, V., Smith, M., 2004. Comparison of continuous lumped and semi-distributed hydrologic modeling using NEXRAD data. Journal of Hydrologic Engneering in press.