ARTICLE IN PRESS

Journal of Hydrology (2006) **xxx**, xxx–xxx

available at www.sciencedirect.com

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

Intercomparison of lumped versus distributed hydrologic model ensemble simulations on operational forecast scales

Theresa M. Carpenter *, Konstantine P. Georgakakos

Hydrologic Research Center, 12780 High Bluff Drive, Suite 250, San Diego, CA 92130, United States
Scripps Institution of Oceanography, UCSD, La Jolla, CA 92093-0224, United States

Received 6 November 2004; received in revised form 14 December 2005; accepted 9 February 2006

KEYWORDS
Distributed hydrologic model;
Streamflow simulation;
Model comparison;
Uncertainty;
Ensemble simulation

Summary Distributed hydrologic models, with the capability to incorporate a variety of spatially-varying land characteristics and precipitation forcing data, are thought to have great potential for improving hydrologic forecasting. However, uncertainty in the high resolution estimates of precipitation and model parameters may diminish potential gains in prediction accuracy achieved by accounting for the inherent spatial variability. This paper develops a probabilistic methodology for comparing ensemble streamflow simulations from hydrologic models with high- and low-spatial resolution under uncertainty in both precipitation input and model parameters. The methodology produces ensemble streamflow simulations using well calibrated hydrologic models, and evaluates the distinctiveness of the ensembles from the high- and low-resolution models for the same simulation point. The study watersheds are of the scale for which operational streamflow forecasts are issued (order of a few 1000 km^2), and the models employed are adaptations of operational models used by the US National Weather Service. A high-resolution (i.e., spatially distributed) model and a low-resolution (i.e., spatially lumped) model were used to simulate selected events for each of two study watersheds located in the southern Central Plains of the US using operational-quality data to drive the models. Ensemble streamflow simulations were generated within a Monte Carlo framework using models for uncertainty in radar-based precipitation estimates and in the hydrologic soil model parameters. The Kolmogorov–Smirnov test was then employed to assess whether the ensemble flow simulations at the time of observed peak flow from the high- and low-resolution models can be distinguished with high confidence. Further assessment evaluated the model performance in terms of reproducing the observed peak flow magnitude and timing. Most of the selected events showed the high- and low-resolution models produced statistically different flow ensembles

* Corresponding author. Tel.: +1 858 794 2726; fax: +1 858 792 2519.
E-mail address: tcarpenter@hrc-lab.org (T.M. Carpenter).

0022-1694/$ - see front matter
doi:10.1016/j.jhydrol.2006.02.013

for the peak flow. Furthermore, the high-resolution model ensemble simulations consistently had higher frequency of occurrence within specified bounds of the observed peak flow magnitude and timing.

Introduction

Distributed hydrologic models feature the capability to incorporate a variety of spatially varying data from a proliferating set of databases on land use, land and soil characteristics, and high resolution precipitation, temperature, and other forcing input. In addition to facilitating simulations and prediction with higher resolution than lumped models, this feature offers the potential to improve hydrologic predictions on current operational scales by accounting for the inherent spatial variability that has historically been lumped into watershed average characteristics (e.g., Smith et al., 2004; Boyle et al., 2001; Beven, 1992, 2002). Central to considering such potential improvement in prediction is the question of balancing model complexity and resolution with the uncertainty in estimating distributed model parameters and input (e.g., Michaud and Sorooshian, 1994).

This work addresses this concern within a probabilistic or ensemble streamflow prediction framework (e.g., Georgakakos and Krzysztofowicz, 2001). Specifically, we address the question: Under present-day operational parametric and radar-rainfall uncertainty, can we distinguish ensemble peak flow simulations produced on scales of order 1000 km^2 by a hydrologic model with spatially aggregate parameters and input from those produced on the same scales by a model with spatially distributed parameters and input? And, if so, how can we characterize the model performance under these uncertainties? The focus of this paper is the development of a methodology for answering these research questions and to illustrate its application for two watersheds in the south-central United States.

A variety of earlier studies have inter compared distributed versus lumped model simulations (e.g., Koren et al., 2004; Boyle et al., 2001; Zhang et al., 2004; Refsgaard and Knudsen, 1996; Shah et al., 1996). The present study generalizes the approach of intercomparison by allowing for uncertainty in the input and parameters of the distributed and lumped models. This is a critical enhancement, as results of limited applicability may be obtained from intercomparisons that ignore the substantial uncertainty in distributed radar rainfall estimates and model parameters.

The ensemble streamflow simulations that incorporate uncertainty are generated for selected event periods from well-calibrated hydrologic models with both spatially-lumped and spatially-distributed resolutions and for each of the case-study watersheds. The Sacramento soil moisture accounting model is employed for runoff generation in both model resolutions, with the distributed model additionally utilizing a kinematic channel routing component. Both model resolutions reproduce observed streamflow well over the calibration period (1993–1999) for the study watersheds. Selected event periods are simulated with the two model resolutions, incorporating uncertainty in both model parameters and in radar-based precipitation input with prescribed error distributions developed in prior work (Carpenter and Georgakakos, 2006).

The method of intercomparison for the ensemble simulations involves a statistical test of whether the distributions of simulated flow from the two different model resolutions are distinguishable. The test (Kolmogorov–Smirnov) does not require the ensembles be drawn from a specified distribution, but rather applies generally to any continuous distribution. The K–S test is applied at the time of the observed peak flow for selected events identified for each watershed to assess whether the peak flow ensembles for the two model resolutions are different. Further, model performance criteria were then examined for these events to assess which model better represented the observed hydrograph in terms of peak flow magnitude and peak timing.

The case study watersheds are discussed in the following section along with descriptions of the hydrologic models, data and calibration. The methodology for the comparison of ensemble simulations is presented in the section entitled ''Methodology for comparison of ensemble flows''. An overview of the uncertainty models for model parameters and radar-based precipitation input is given in ''Models of uncertainty'' section. The comparison results for the study watersheds are presented in the penultimate section. This section also includes a discussion of the model performance assessment. The final section summarizes the conclusions and addresses potential future research directions.

Study watersheds and hydrologic models

The study watersheds are two of four river basins selected for the US National Weather Service (NWS) sponsored Distributed Model Intercomparison Project (DMIP, Smith et al., 2004). The Blue River basin with outlet near Blue, in southern Oklahoma, encompasses an area of 1232 km^2 and is characterized by its relatively narrow, elongated shape. The 1589 km^2 Illinois River basin with outlet near Watts is located northeast of the Blue River basin, along the Arkansas–Oklahoma border (see Fig. 1). The region is characterized with a semiarid climate and receives much of its warm season rainfall from convective storms. The study watersheds have a history of distributed modeling research effort due in part to the record of NEXRAD radar precipitation in Oklahoma and activities such as DMIP (e.g., Georgakakos et al., 1996; Zhang et al., 2004; Smith et al., 2004, and others in that special issue).

For this effort, the watersheds were modeled using both low-resolution (lumped parameter and input) and high-resolution (distributed parameter and input) hydrologic models to provide streamflow simulations at the two watershed outlets. Fig. 1 shows the discretization of subcatchments

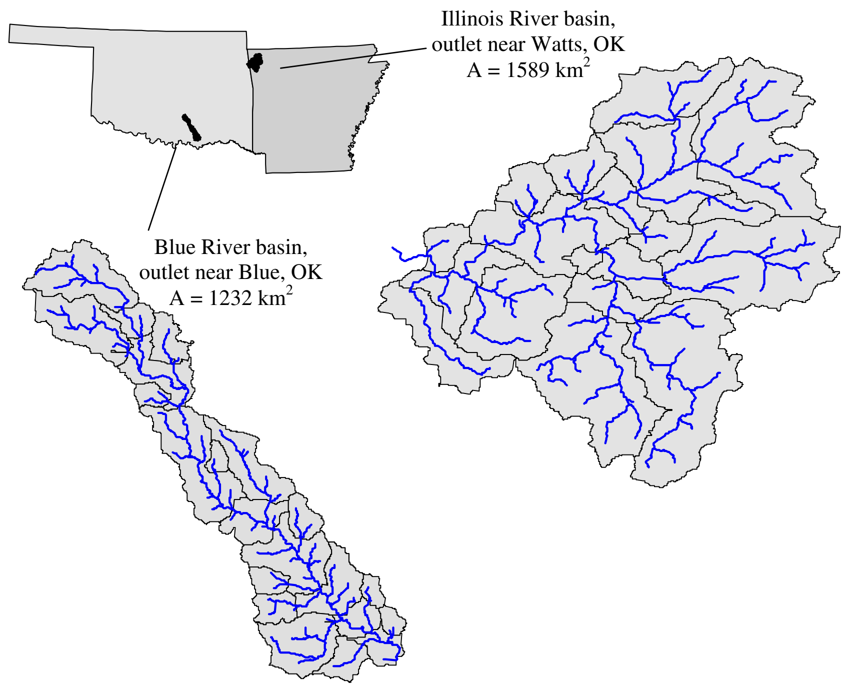

Figure 1 Location of study watersheds. The large watershed images indicate the delineated subcatchments used in the distributed modeling (black outlines), along with the stream network (thick lines).

for the distributed modeling of the two study watersheds. For the Blue River, a total of 21 subcatchments were delineated, yielding an average subcatchment size of 59 km^2. A total of 19 subcatchments were delineated for the Illinois River at Watts, with an average subcatchment size of 84 km^2. The lumped model considers each watershed as a single aggregate unit with model parameters and input applied to the entire area of each watershed.

The lumped model is based on the Sacramento soil moisture accounting model (Burnash et al., 1973), which is used operationally by the US NWS for streamflow forecasts. The model conceptualizes the watershed as a vertical column of soil with upper and lower soil layers, representing the upper several centimeters and the deeper meter or so of soil depth. The Sacramento model accepts soil moisture input through precipitation, simulates extraction of soil moisture by evapotranspiration, and estimates the water fluxes at the surface, between soil layers, and to groundwater. The output of the model is catchment outflow (or inflow to the receiving channel). A set of model parameters describes the moisture storage capacities, withdrawal rates, percolation to the lower soil layer, and within-catchment time delay between precipitation forcing and appearance of flow at the watershed outlet.

The distributed model employed, HRCDHM, is a subcatchment based model which uses the Sacramento model for runoff generation within each subcatchment. It also includes components for temporal distribution of runoff at the subcatchment outlet and kinematic channel routing between subcatchments and to the watershed outlet (Carpenter et al., 2001; Carpenter and Georgakakos, 2004a). The structure of HRCDHM allows for distributed model parameters and precipitation input at the scale of the subcatchments delineated. Parameters include the Sacramento model parameters (as described for the lumped model) for each subcatchment, along with estimates of channel characteristics required for routing.

Estimates of the model parameters for both catchments and models were derived by calibration using a 6-year record of historical hydrometeorological data (5/1993–7/1999). The hourly data used consisted of discharge, NEXRAD multisensor precipitation estimates, and energy forcing provided by DMIP (used to estimate potential evaporation for the distributed model application, see Carpenter and Georgakakos, 2004a). For the lumped model, calibration involved deriving estimates for the set of Sacramento model parameters, while for the distributed model, parameters were derived for the study watersheds from the application of HRCDHM in DMIP. The spatial variation in model parameters is based on the variation in selected soil properties derived from the STATSGO database (NCRS, 1994). Calibration

Table 1 Statistics of simulated discharge at the watershed outlets over the calibration period (5/1993–7/1999) for both lumped and distributed models

	Avg. *Q* (mm/h)	SD *Q* (mm/h)	% Bias	Correlation
Illinois R, Watts				
Observed	0.047	0.081		
Lumped	0.05	0.09	−0.9	0.87
Distributed	0.046	0.082	−2.47	0.87
Blue R, Blue				
Observed	0.028	0.09		
Lumped	0.034	0.093	9.6	0.86
Distributed	0.029	0.07	1.03	0.83

of the Sacramento model parameters involved first estimating a set of ''nominal'' model parameters applied to all subcatchments and representing the watershed average parameters, and then allowing for spatial variation of the model parameters between subcatchments based on the STATSGO data. Detail on the use of the STATSGO data is provided by Carpenter and Georgakakos (2004a).

Table 1 presents performance statistics of the lumped and distributed models over the calibration period for the study watersheds. The performance of the two models is comparable and suggests that both models reproduce the historical streamflow at the watershed outlet well, with less than 10% bias and with high hourly correlation of observed and simulated flows. Fig. 2 presents the annual cycle in observed and simulated flows. Both models are able to reproduce the observed annual cycle quite well. For the Blue River, the lumped model shows an overestimation in the summer months (July–August) when flows are quite low. However, the events selected for further analysis in the intercomparison of ensemble flows are outside of these months; thus, for this analysis we consider the performance of the distributed and lumped models comparable from hourly to monthly scales.

As part of DMIP, Reed et al. (2004) inter compared the streamflow simulations from 12 different distributed models along with a calibrated operationally-based lumped model

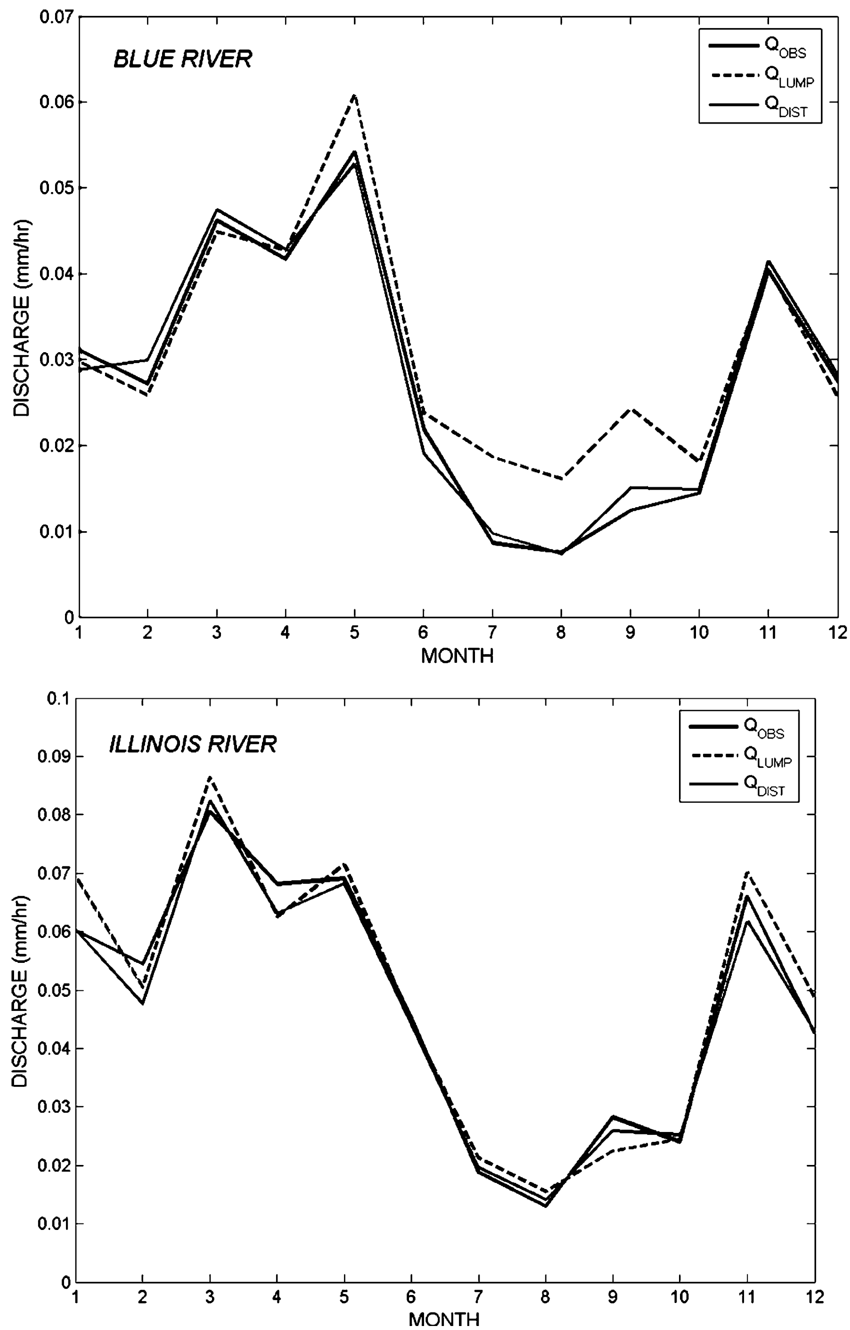

Figure 2 Annual cycle of observed and simulated discharge (both lumped and distributed models) at the study watershed outlet locations.

for the four DMIP study watersheds. They find that while some models have comparable performance to the lumped model, the calibrated lumped model still outperforms the distributed models. The exception to this appears to relate to basin shape, orientation, and soil characteristics, as in the Blue River basin. The comparison of Reed et al. (2004) did not consider uncertainty due to input, model structure, or calibration adequacy, but rather single, deterministic simulations from each model. Georgakakos et al. (2004) and Butts et al. (2004) begin to look at simulated flow uncertainty through multi-model ensembles for the DMIP study watersheds. In both studies, their conclusions suggest the importance of considering uncertainty for operational streamflow forecasting and the utility of multiple model ensembles to consider model parametric and structural uncertainty.

Methodology for comparison of ensemble flows

The focus herein is on the development of a methodology to intercompare ensemble streamflow simulations from hydrologic models with high- and low-spatial resolutions. The use of ensemble simulations allows for consideration of uncertainty due to erroneous model parameters values and noisy radar rainfall input present in the operational streamflow forecasting environment. The basis for this analysis presumes the following:

(a) parametric and radar-rainfall uncertainties are significant under operational streamflow forecasting;
(b) ensemble streamflow simulations provide a more useful and complete representation of model response under uncertainty than a single ''nominal'' simulation;
(c) uncertainty in model parameters may be characterized by variation in available soil properties or/and by the hydrologist degree-of-belief error estimates;
(d) radar rainfall pixel-scale uncertainty is spatially correlated and may be upscaled to various levels of model resolution;
(e) scales of the order of 1000 km^2 are significant for operational flow forecasting;
(f) peak flows are significant for operational flow forecasting.

The general approach is to generate ensemble streamflow simulations of selected events from both distributed and lumped hydrologic models incorporating uncertainty due to potentially erroneous model parameters values and noisy radar rainfall input, and to perform a statistical comparison of the simulated flow at a given time during each event to assess whether or not the sample distributions of simulated flows from the two model resolutions belong to the same population. The ensembles are generated using the calibrated hydrologic models for the study watersheds. The models are run in a Monte Carlo fashion with sampling from defined models of uncertainty for parametric and precipitation input. These uncertainty models were put forth by Carpenter and Georgakakos (2006) and are summarized in the section ''Models of uncertainty''.

To evaluate the pair of simulation ensembles for each event, a Kolmogorov–Smirnov (K–S) test was applied. In this application, the K–S test examines the distributions of simulation ensembles from the distributed and lumped models at a selected time during the event period, and assesses whether the two sample distributions come from the same parent population. The K–S test is independent of the form of input distributions, but it does require that the distributions be continuous. The K–S test statistic, D_o, is given by:

$$D_o = \max(P_{\text{distrib}}(Q_p) - P_{\text{lumped}}(Q_p)) \tag{1}$$

with

$$\text{Prob}(D > D_o) = F(N_e, D_o). \tag{2}$$

D_o represents the maximum separation between the sample cumulative frequency distribution functions, $P_{\text{distrib}}(Q_p)$ and $P_{\text{lumped}}(Q_p)$, of the ensemble flows (Q_p) at a selected time (t_p) during each event for the distributed and lumped models, respectively. This time, t_p corresponds to the timing of the observed peak flow during each event. N_e is the ratio of the product of the number of ensemble members in the lumped and distributed cases divided by their sum ($N_e = N_D * N_L/[N_D + N_L]$). The number of ensemble members was set to 100 for both distributed and lumped model runs (N_D and N_L, respectively). F represents the probability of having a separation greater than D_o for two ensembles from the same distribution given D_o and the number of ensemble members represented by N_e. This value can be found in standard statistical tables. The K–S test is effective for measuring shifts in distributions as it is most sensitive around the median value of distributions. The null hypothesis in this application is that the two sets of flow ensembles are drawn from the same distribution, and therefore are indistinguishable. The significance level used for accepting the null hypothesis was 1%. This comparison evaluates whether the ensemble flow simulations at the time of observed peak flow are statistically different with high confidence. This K–S test was performed for each event of the study watersheds.

Models of uncertainty

In hydrologic simulations, uncertainty exists in model structure, model parameters, and forcing input. In this analysis, the uncertainty in model parameters and radar rainfall estimates is considered. The uncertainty models have been developed in Carpenter et al. (2001), Georgakakos and Carpenter (2003), Carpenter and Georgakakos (2004b, 2006). The uncertainty models used in this analysis are (a) the model for soil model parameter uncertainty and (b) the spatially-correlated radar rainfall uncertainty model presented in Carpenter and Georgakakos (2006). The uncertainty models and applications are summarized in the following subsections.

Model for parameter uncertainty

Parametric uncertainty involved multiple parameters of the Sacramento model runoff generation component. The parameters of the upper layer soil moisture capacities, interflow rate, and percolation function were considered simultaneously. Uncertainty in the lower soil layer parameters was not considered in this analysis due to the focus on selected high flow events, which in model simulations are primarily the result of the upper soil zone activity. The generic formulation for parametric uncertainty is:

$$a = \mu_a + \varepsilon_a, \tag{3}$$

where a is a model parameter, μ_a is the mean value of that parameter, ε_a is an random error selected from a uniform distribution in the range $(-a_L, +a_U)$. This range in error was derived based on the attributes of spatial soils data for the distributed model and on 'degree of belief' estimates of the hydrologist who performed the calibration for the lumped model. The STATSGO database was again used to derive ranges in subcatchment soil properties that define the limits of the error within the distributed modeling. For the soil property of available water content, the range for the Blue River watershed subcatchments was ±17% to 22% and for the Illinois River watershed, this range was ±20% to 25%. For the permeability soil properties, the range was approximately ±50% for both study watersheds. For the lumped model, the error range was used was set to ±25% of the parameter value for the upper zone capacity. Georgakakos and Carpenter (2003) show that for the Blue River, this range resulted in model performance criteria deemed as adequate for calibration with multiple objectives. This range was retained for both the Blue River and Illinois River watersheds.

Model for radar-rainfall uncertainty

Precipitation input to the hydrologic models is derived from NEXRAD multisensor (quality-controlled) radar-rainfall estimates, which have a spatial resolution of approximately $3.5 \times 3.5\ \mathrm{km}^2$. This input is averaged for the scale of model resolution (i.e., as mean areal precipitation (MAP) estimates over the entire watershed or for each subcatchment of a given watershed). The objective in the uncertainty model formulation was to upscale the uncertainty in radar pixel-scale rainfall estimates to the spatial resolutions of the hydrologic models. The methodology and relevant formulation is presented in Carpenter and Georgakakos (2006) and we only summarize its salient features here for easy reference.

The radar-rainfall pixel-scale error model was adopted from previous studies (Georgakakos and Carpenter, 2003; Carpenter and Georgakakos, 2004b, 2006), and includes a magnitude dependent error relationship such that low magnitude rainfall rates are associated with higher relative errors. A linear relationship between the standard deviation of rainfall rate error and rainfall rate magnitude is assumed for rainfall rates less than 25 mm/h, with the ratio of standard deviation to rainfall rate magnitude ($2\sigma/R$) ranging from a value of 1.0 at a rainfall rate of 0.0 to a value of 0.5 at R = 25 mm/h. For rates greater than 25 mm/h, a con-

Table 2 Peak observed flows of selected event periods for the study watersheds

#	Illinois river dates	Q_{peak} (mm/h)	Blue river dates	Q_{peak} (mm/h)
1	9/12–19/1993	0.682	2/27–3/07/1994	0.388
2	–	–	4/25–5/10/1994	0.655
3	11/12–20/1993	0.849	10/19–31/1994	0.098
4	3/06–18/1994	0.309	11/12–19/1994	0.628
5	3/25–4/1/1994	0.358	3/11–20/1995	0.433
6	4/10–18/1994	0.330	4/02–09/1995	0.432
7	4/28–5/7/1994	0.235	5/06–13/1995	0.843
8	11/03–13/1994	0.447	11/05–12/1996	1.411
9	12/07–14/1994	0.240	11/15–23/1996	0.274
10	1/12–18/1995	0.756	11/22–12/5/1996	0.672
11	5/06–13/1995	1.171	2/18–25/1997	0.568
12	6/08–16/1995	0.899	3/11–18/1997	0.155
13	9/24–10/1/1996	1.293	3/23–29/1997	0.175
14	11/05–12/1996	1.149	4/02–10/997	0.290
15	11/15–23/1996	0.364	4/09–17/1997	0.183
16	11/22–12/5/1996	0.982	5/07–14/1997	0.123
17	2/19–25/1997	1.174	6/08–16/1997	0.380
18	3/08–21/1997	0.317	12/18–29/1997	0.350
19	4/07–16/1997	0.218	1/03–14/1998	0.514
20	–	–	1/20–31/1998	0.205
21	1/03–14/1998	1.591	3/05–14/1998	0.346
22	3/05–15/1998	0.339	3/14–27/1998	0.595
23	3/14–27/1998	0.691	12/02–10/1998	0.117
24	10/3–12/1998	0.392	4/01–09/1999	0.501
25	2/05–12/1999	0.509	5/08–16/1999	0.177
26	3/10–19/1999	0.466		
27	4/02–13/1999	0.330		
28	5/02–10/1999	0.751		
29	5/09–17/1999	0.344		
30	12/19/97–1/02/98	0.250		

stant relative error ($2\sigma/R = 0.5$) is used. Subcatchment size-dependent relationships between the subcatchment MAP uncertainty and MAP magnitude were derived by applying the pixel-scale error model to the historical record of hourly radar precipitation in a Monte Carlo sampling framework and computing subcatchment and catchment-scale MAP uncertainty. Utilizing the formulation of Carpenter and Georgakakos (2006), the application allowed for spatial correlation of the radar pixel error with the nearest two pixel locations in each of the north, south, east and west directions. The resulting subcatchment and catchment-scale MAP uncertainty relationships, expressed through the ratio of MAP standard deviation to MAP rainfall rate ($2\sigma_{MAP}/R_{MAP}$), were summarized by linear relationships between the $2\sigma_{MAP}/R_{MAP}$ and R_{MAP} for MAP rainfall rates less than 25 mm/h, and by a constant value of $2\sigma_{MAP}/R_{MAP}$ for R_{MAP} greater than 25 mm/h. Furthermore, these relationships varied with the size of the subcatchments, due to the number of radar pixels included in the MAP computation. For the subcatchments of the Blue River watershed, this ranged from 1 to 9 pixels, and from 1 to 13 pixels for the Illinois River. The linear relationships for relative error decreased in both magnitude and slope as the size of the subcatchment increased. For a subcatchment containing 2 pixels, the ratio ($2\sigma_{MAP}/R_{MAP}$) varies from 0.94 at $R_{MAP} = 0$ mm/h to 0.47 at $R_{MAP} = 25$ mm/h where the ratio is held constant beyond $R = 25$ mm/h. For the Blue River and for the catchment containing 9 radar pixels, the ratio was estimated to vary from 0.79 at $R_{MAP} = 0$ mm/h to 0.41 at $R_{MAP} = 25$ mm/h and beyond. These MAP uncertainty relationships were also applied for the entire catchment area of the study watersheds to produce the radar-based rainfall forcing uncertainty for the lumped model.

Results

A set of events were selected from the 6-year historical record for each watershed to be used for the intercomparison of ensemble simulated flows. The selection of events was based on the observance of a distinct peak in observed discharge record at the watershed outlet. A total of 25 events were selected for the Blue River basin, and 28 events were used for the Illinois River at Watts. The dates of the events, along with the observed peak flow, are listed in Table 2. Each event period was defined from approximately two days prior to the observed peak and through the event until baseflow conditions returned. In several instances, multiple peaks were included during the event period. In such cases only the highest peak of the event was considered for the analysis described herein. For the lumped and distributed models, each event was simulated in a Monte Carlo fashion, sampling from both the parameter and precipitation input uncertainty models outlined in ''Models of uncertainty'' section and yielding an ensemble of simulated stream flows with 100 ensemble members. Examples of the ensemble simulated flows for a few events are presented in Fig. 3. Although the figure presents the ensemble simulations and observed flows for given event periods only, each ensemble was generated sampling from uncertainty starting approximately two months prior to the event period so that a stable

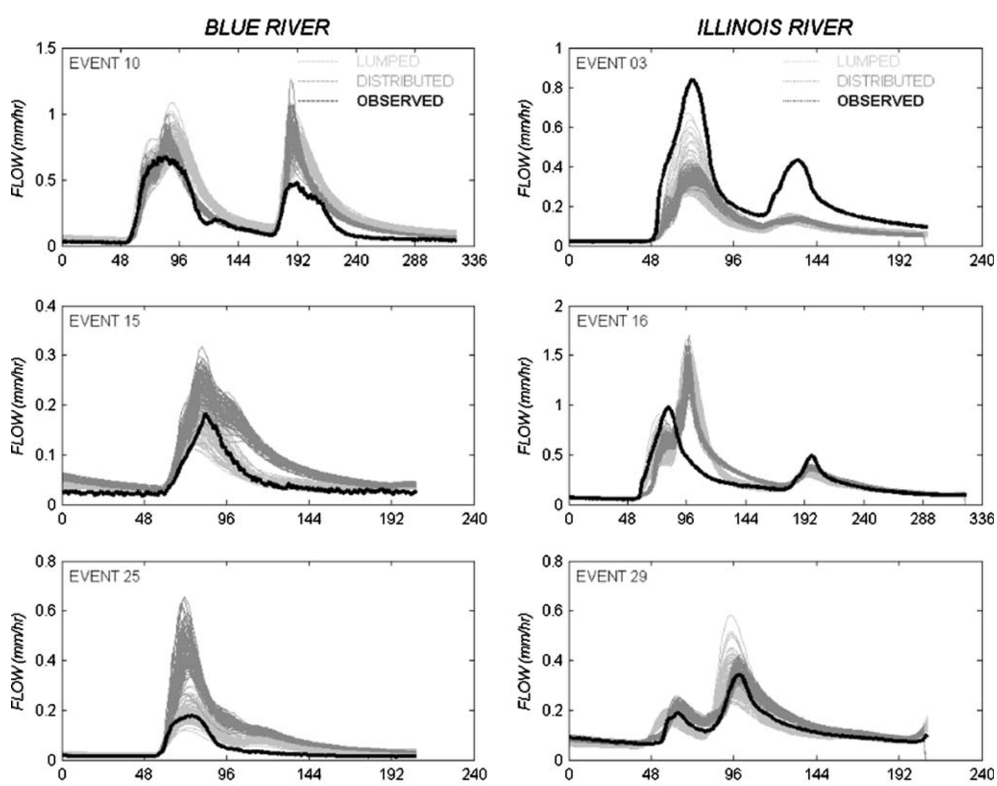

Figure 3 Representative ensemble flow simulations from both the distributed (dark gray) and lumped (light gray) models, along with the observed event hydrograph.

distribution of soil moisture conditions was established at the beginning of each event.

The Kolmogorov–Smirnov test described in ''Methodology for comparison of ensemble flows'' section was applied to compare the ensemble distributions of simulated flows from the lumped and distributed models at the time of the observed peak flow of each event. The K–S test results indicate that only one event for the Blue River basin and two events for the Illinois River basin failed to reject the null hypothesis that the ensembles are drawn from the

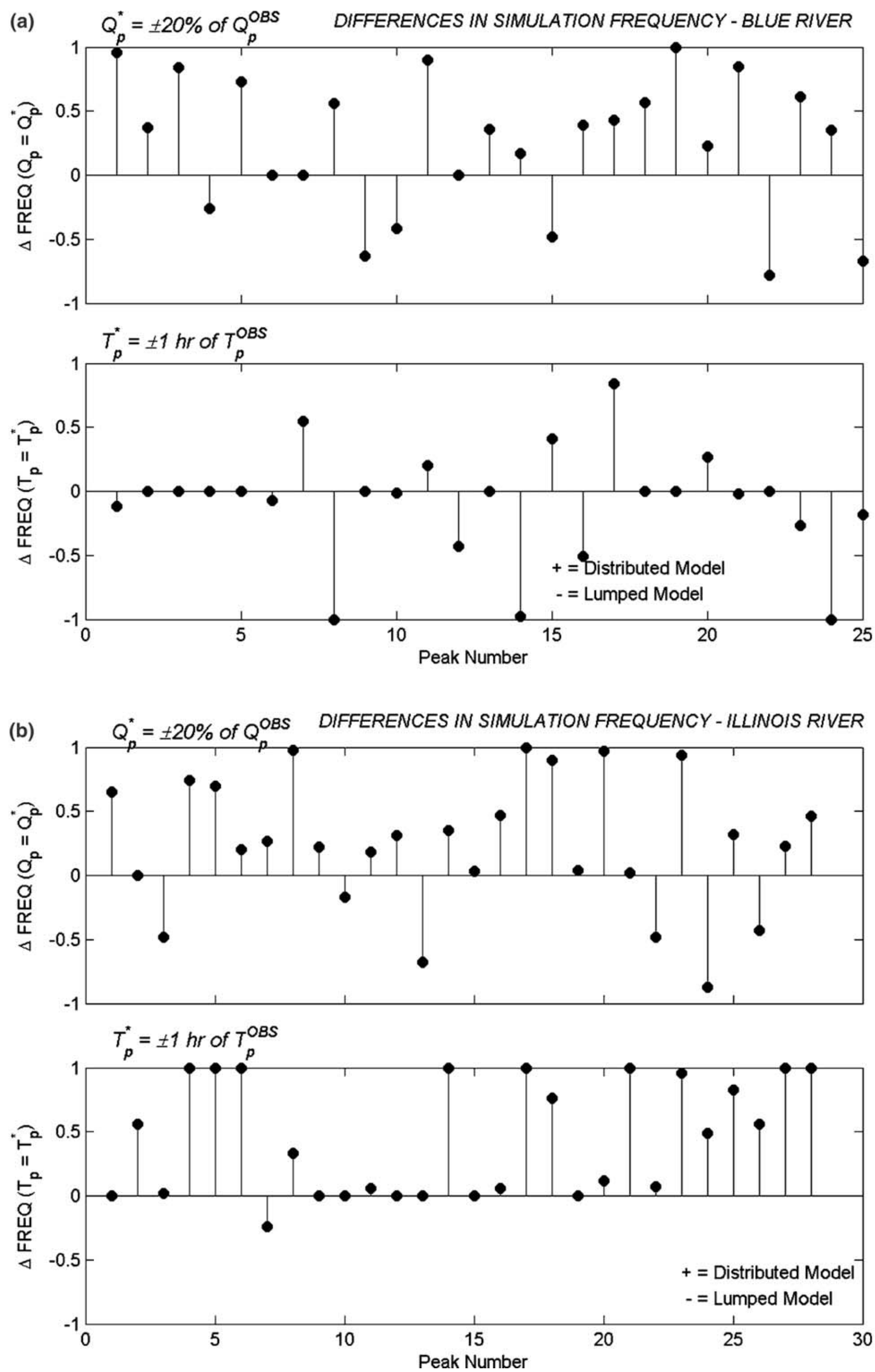

Figure 4 Differences in the ensemble frequencies of producing peak flow simulations within 20% of observed peak magnitude and within ±1 h of observed peak flow timing. Positive values indicate higher frequency for the distributed model and negative values mean the lumped model produced higher frequency.

same distribution at the 1% confidence level. This implies that the distributions of ensemble simulated flows from the lumped and distributed models at the time of the observed peak flow are indeed statistically distinguishable with high confidence. The events for which the K–S test does not distinguish the two distributions of ensemble flows are event 15 on the Blue River and events 5 and 6 on the Illinois River.

Given that the distributed and lumped model simulations of event peak flow under parametric and input uncertainty are statistically different, the question arises as to which model has better performance. To make an assessment regarding model performance, the assessment metrics must explicitly take into consideration that the flow simulations are in the form of ensembles due to the parametric and input uncertainties. We do so by considering the frequency with which the ensemble simulations produced flows that were (a) within ±20% of the observed peak flow magnitude and (b) within ±1 h of the timing of the observed peak. There were 100 ensemble 'samples' of simulated flows considered for each event in this assessment. The differences in these frequencies are illustrated in Fig. 4 for all events. The difference in frequency is defined as:

$$\Delta FREQ = FREQ_{\text{distrib}} - FREQ_{\text{lumped}}, \tag{4}$$

where $\Delta FREQ$ signifies the difference in sample frequencies for the distributed model ($FREQ_{\text{distrib}}$) and for the lumped model ($FREQ_{\text{lumped}}$). Positive values of $\Delta FREQ$ indicate that the distributed model had a larger number of ensembles within the peak flow or timing bounds and, thus, better performance. Negative values indicated better performance for the lumped model. As the assessment metric is dependent on the number of samples, we consider only $\Delta FREQ$ values outside the interval (-0.2, $+0.2$) as indicative of better performance.

With respect to the peak flow magnitude, this comparison shows that in 60% of the events for the Blue River and 57% for the Illinois River, the distributed model performs better in simulating the peak flow within 20% of the observed peak magnitude. Less than 25% of the events indicated better performance by the lumped model in simulating the observed peak. With respect to the timing of the peak flow, Fig. 4 shows that for the Illinois River the distributed model consistently performed better than the lumped model. In 15 of the 28 events, the distributed model shows a higher frequency of simulating the peak flow within a 3-h time period centered on the observed time of peak. In only one event on the Illinois River was the frequency of accurate lumped model timing greater than that for the distributed model. In contrast, the comparison results for the Blue River were nearly even, with 20% of the events indicating better timing performance for the distributed model compared to 24% of the events indicating better timing performance by the lumped model.

Fig. 5 presents a comparison of the peak flow model performance indicators with respect to the magnitude of the observed peak flow. The distributed model performance appears to be consistently better on the Blue River for flows with medium-range peaks, i.e., in the flow range from of 0.3–0.6 mm/h, as indicated by the large, positive $\Delta FREQ$ values within this peak flow range. For flows outside this range, there is no clearly defined pattern of performance between the distributed and lumped models. For the Illinois River, the results show that the distributed model most consistently performs better for observed peak flows less than approximately 0.5 mm/h.

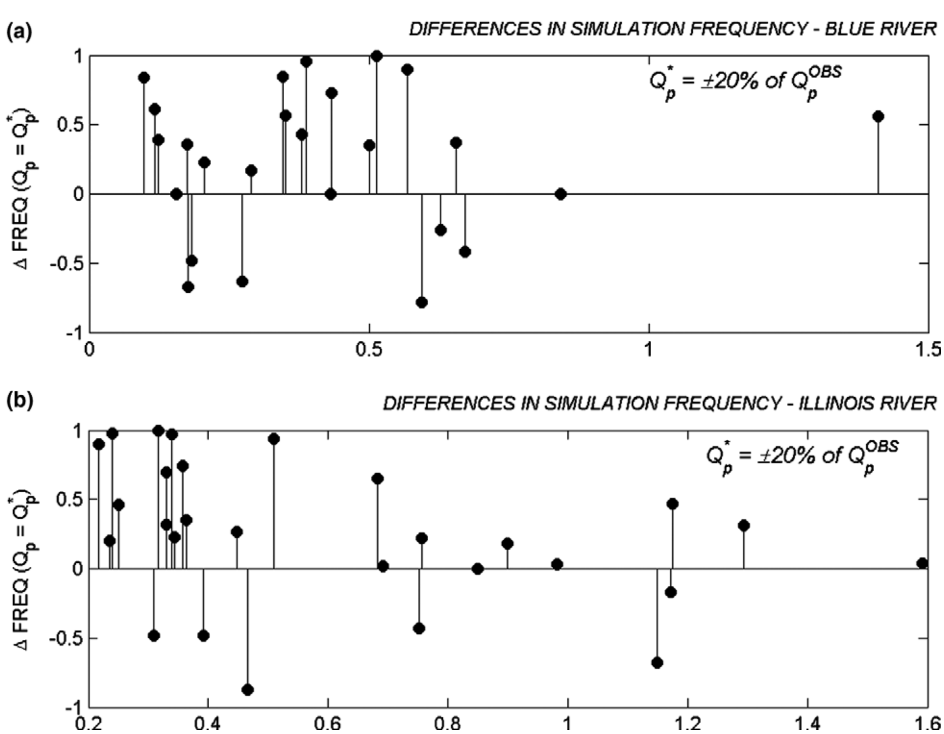

Figure 5 Differences in the ensemble peak flow simulation frequencies as a function of peak flow magnitude. Again, positive values indicate higher frequency for the distributed model and negative values mean the lumped model produced higher frequency.

Finally, we examined differences in the spatial distribution of certain model states and output between the distributed and lumped models. These differences were at first examined for those events identified as the lumped model performing better with respect to peak flow magnitude and secondly for those events for which the distributed model performed better. This analysis used detailed model output from the results of Carpenter and Georgakakos (2006) for the Blue River at Blue and Illinois River at Tahlequah, of which the Illinois River at Watts is a subcatchment. This output included the initial soil water content at the beginning of each event period, event total precipitation and measures of the simulated runoff and streamflow variability for each subcatchment in the distributed model at resolutions corresponding to the subcatchment delineation used in this paper.

For only those events for which the lumped model was identified as having better performance in producing the simulated peak within 20% of the observed peak magnitude (this included a total of 6 events on the Blue River and 5 events on the Illinois River), the composite initial soil water content and composite event total precipitation were computed. The composites were computed as the average initial soil water content or event precipitation of each subcatchment over the selected events. Similar composites were computed for only those events for which the distributed model had better performance (this included 17 events for the Illinois River and 15 events for the Blue River basin). Fig. 6 presents the comparison for the Blue River basin, with the average initial soil water content presented as a fraction of total soil water capacity and the average event total precipitation given as a fraction of the upper zone soil water capacity. Spatial variability exists in both average initial soil water content and average event precipitation. For those events in which the lumped model showed better performance, the average initial soil water content is somewhat higher, but the spatial distribution in average soil water content between lumped model events and distributed model events is similar. There is, however, a clear difference in the spatial distribution of average event precipitation, for the Blue River basin. The lumped model events show a more uniform distribution of precipitation across

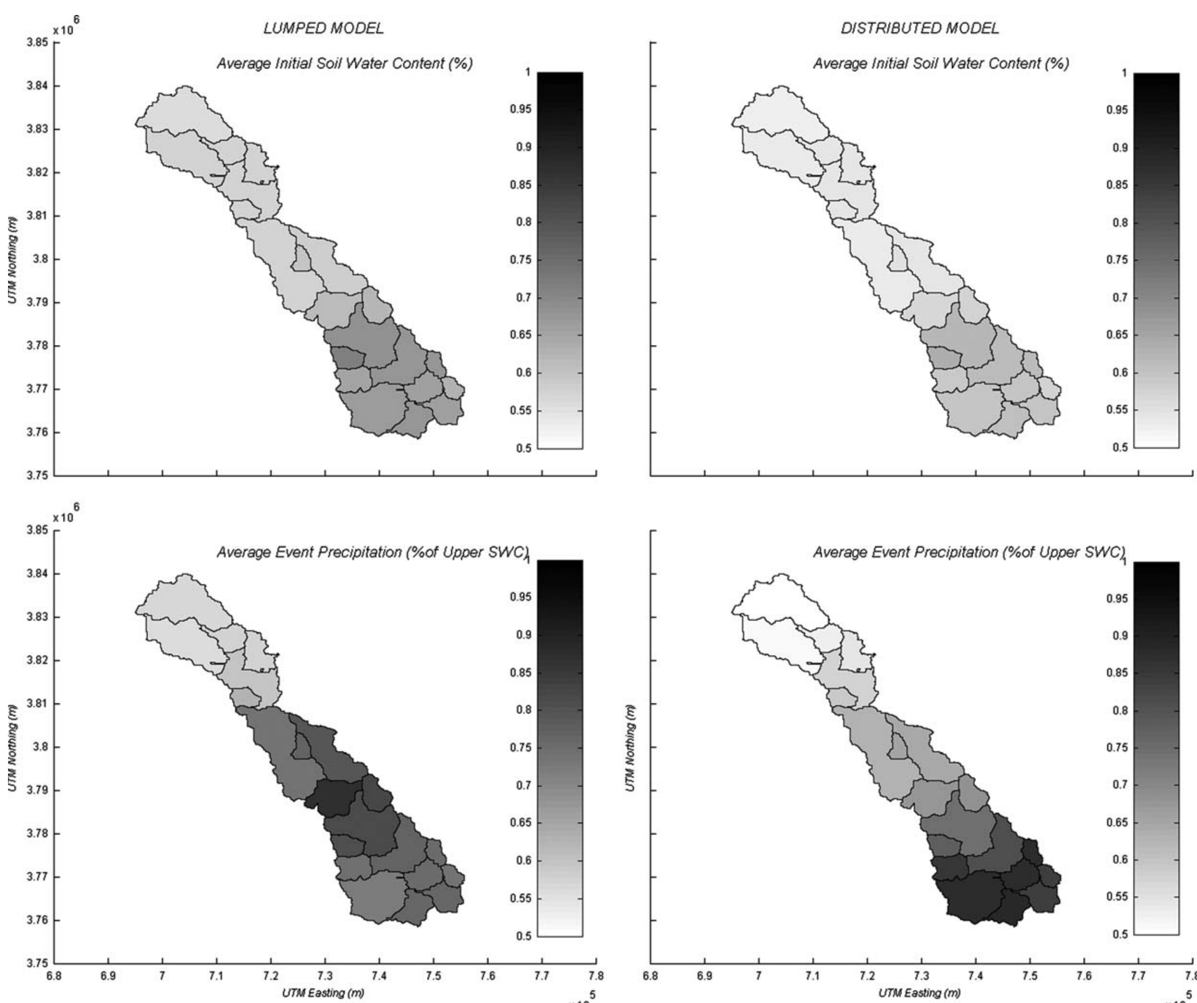

Figure 6 Difference in spatial distribution of initial soil water content (expressed as fraction of capacity) and event total precipitation (expressed as fraction of upper soil water capacity) averaged only for those events in which the lumped model showed better performance (left column) versus only those events in which the distributed model showed better performance (right column).

the watershed, with the upper third of the basin having lower precipitation. The highest average precipitation is located towards the centroid of the watershed. In contrast, the distributed model events show a wider range in average precipitation across the watershed and a gradual increase in average precipitation towards the watershed outlet. The highest average precipitation occurs near the watershed outlet for the distributed model performance events.

To ensure the above results were not prejudiced by the number of events used in the computation of the initial soil water and event total precipitation composites, this analysis was repeated using the same number of events for both cases (events that showed better performance with the lumped model versus events that showed better performance with the distributed model). As a greater number of events were identified for distributed model performance, this required random selection of a subset of these events such that the total number of events used for the distributed model event composite equaled that for the lumped model event composite. This sampling of the distributed model performance events was repeated a total of 10 times. The composites showed a pattern consistent with the previous remarks. For the initial soil moisture composite, the distributed model performance events showed a similar spatial distribution as the lumped model performance events with either similar or slightly lower average initial soil water content values. The composite event total precipitation generally shows a wider range in average total precipitation and highest precipitation towards the watershed outlet.

Similar analysis of composite initial soil water content and event precipitation was carried out for the Illinois River basin focusing only on the spatial distribution of both variables for the drainage basin up to the Watts subcatchment. The spatial distribution of average initial soil water content was quite uniform for both lumped model events and distributed model events. For the lumped model performance events, the average precipitation indicated lower values toward the basin outlet and with higher precipitation toward the northern and southern headwater reaches of the watershed. The composite analysis for various subsets of distributed model performance events indicated that the distributed model performed better for events where the highest precipitation occurred in the southern portion of the watershed. The magnitude of the composite event precipitation for the distributed model events tended to be lower that the composite event precipitation for the lumped model events. For both initial soil water content and event total precipitation, the spatial variation within the Illinois River watershed was much less pronounced than for the Blue River basin.

Conclusions

This paper presents a probabilistic methodology for assessing whether distributed model simulations are different from lumped model simulations under parametric and input uncertainties representative of present-day operational flow-forecasting conditions. It also discusses approaches for inter comparing performance under these uncertainties. For runoff generation, the models use the same soil water accounting component that is also used operationally in the US for real time streamflow prediction. The methodology involves the statistical comparison of the distributions of ensemble streamflow simulations from both spatially-distributed and spatially-lumped hydrologic models. The methodology is applied to two watersheds of the order of 1000 km^2 in size located in the south-central United States. The statistical comparison was made for selected events in the historical record from May 1993 to July 1999. Probabilistic measures of model performance are also developed to evaluate the models ability to reproduce peak flow magnitude and peak flow timing reliably. These measures examine the frequency with which the ensemble simulations produce flows within 20% of the observed peak flow magnitude and within a 3-h window of time centered on the observed time of peak flow.

This intercomparison indicates that the distributed model ensemble simulations are statistically distinguishable from the lumped model ensemble simulations for both study watersheds with a high degree of confidence at the time of the observed peak flow for almost all of the events considered. The distributed model showed better performance with respect to peak flow magnitude in approximately 60% of the events for both study watersheds, whereas the lumped model showed better performance in less than 25% of the events. Better model performance for the distributed model was also shown with respect to peak flow timing on the Illinois River, while for the Blue River the models had comparable performance in that respect.

An analysis of the spatial distribution of initial soil water content and total event precipitation for events for which the distributed model was identified as having better performance versus those events for which the lumped model performed better suggests that the distribution of precipitation within the Blue River basin was more variable across the watershed, with higher total precipitation concentrated near the watershed outlet. Such spatial variability was not as pronounced for the Illinois River watershed.

The research reported herein is part of a number of analyses done on the assessment of distributed hydrologic models, and, from the authors' perspectives, on the impact of uncertainty in parameters and forcing on ensemble streamflow simulation. Our main conclusion is that even on the scales of current lumped operational forecasting models, distributed models offer clear performance advantages under present day parametric and input uncertainties, when used to produce ensemble streamflow simulations.

However, important areas of further research remain. Two are to assess whether: (a) similar behavior is found in different regions with different hydroclimatic and geomorphologic conditions; and (b) these results will hold true for the case of real-time streamflow prediction (rather than simulation). With respect to the former, the presence of snow and snow melt in mountainous areas of the western US will change not only the model response dependence on input but also the propagation of uncertainty from input to streamflow as the air temperature spatial distribution uncertainty will add significant noise to the model. With respect to the latter, the significant uncertainty associated with present day spatially resolved precipitation forecasts will undoubtedly dominate the analysis and may even alter the character of the results found herein.

Acknowledgments

The research effort presented herein has been supported by the US National Weather Service Office of Hydrologic Development under NOAA Order No. DG133W-03-SE-0904. Supplemental support was provided by the California Applications Project of the Scripps Institution of Oceanography, UCSD. The authors gratefully acknowledge the continued collaboration and suggestions of Drs. Michael Smith, Victor Koren and Seann Reed of the NWS Hydrology Laboratory. The comments from two anonymous reviewers were helpful in focusing and clarifying several issues in the manuscript. Their contributions are gratefully acknowledged.

References

Beven, K., 1992. Future of distributed modeling. Hydrologic Processes 6 (3), 253–254.
Beven, K., 2002. Towards an alternative blueprint for a physically based digitally simulated hydrologic response modeling system. Hydrologic Processes 16, 189–206.
Boyle, D.B., 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., Ferral, R.L., McGuire, R.A., 1973. A generalized streamflow simulation system: conceptual modeling for digital computers. Technical Report, Joint Federal-State River Forecast Center, U.S. National Weather Service and California Department of Water Resources, Sacramento, CA, 204pp.
Butts, M.B., Payne, J.T., Kristensen, M., Madsen, H., 2004. An evaluation of the impact of model structure on hydrological modeling uncertainty for streamflow simulation. Journal of Hydrology 298 (1–4), 242–266.
Carpenter, T.M., Georgakakos, K.P., 2004a. Continuous streamflow simulation with the HRCDHM distributed hydrologic model. Journal of Hydrology 298 (1–4), 61–79.
Carpenter, T.M., Georgakakos, K.P., 2004b. Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. Journal of Hydrology 298 (1–4), 202–221.
Carpenter, T.M., Georgakakos, K.P., 2006. Discretization scale dependencies of the ensemble flow range versus catchment area relationship in distributed hydrologic modeling. Journal of Hydrology, in press.
Carpenter, T.M., Georgakakos, K.P., Sperfslage, J.A., 2001. On the parametric and NEXRAD-radar sensitivities of a distributed hydrologic model suitable for operational use. Journal of Hydrology 254, 169–193.
Georgakakos, K.P., Sperfslage, J.A., Guetter, A.K., 1996. Operational GIS-based models for NEXRAD radar data in the U.S. Proceedings International Conference on Water Resources and Environmental Research: Towards the 21st Century, October 29–31, 1996, Kyoto, Japan, pp. 603–609.
Georgakakos, K.P., Seo, D.-J., Gupta, H., Schaake, J., Butts, M.B., 2004. Towards the characterizations of streamflow simulations uncertainty through multimodel ensembles. Journal of Hydrology 298 (1–4), 222–241.
Georgakakos, K.P., Carpenter, T.M., 2003. A methodology for assessing the utility of distributed model forecast applications in an operational environment. In: Tachikawa, Y., Vieux, B.E., Georgakakos, K.P., Nakakita, E. (Eds.), Weather Radar Information and Distributed Hydrological Modeling. IAHS Press, Wallingford, UK, pp. 85–92, IAHS Publ. No. 282.
Georgakakos, K.P., Krzysztofowicz, R., 2001. Probabilistic and ensemble forecasting. Journal of Hydrology 249 (1–4), 196.
Koren, V., Reed, S., Smith, M., Zhang, Z., Seo, D.-J., 2004. Hydrology Laboratory Research Modeling System (HL-RMS) of the U.S. National Weather Service. Journal of Hydrology 291 (3–4), 297–318.
Michaud, J., Sorooshian, S., 1994. Comparison of simple versus complex distributed runoff models on a mid-sized semiarid watershed. Water Resources Research 30 (3), 593–605.
NCRS (National Resource Conservation Service), 1994. State Soil Geographic (STATSGO) Database. Misc. Publ. 1492, U.S. Dept. of Agriculture, Fort Worth, TX, 37pp + Appendices.
Reed, S., Koren, V., Smith, M., Zhang, Z., Moreda, F., Seo, D.-J., 2004. DMIP participants, overall distributed model intercomparison project results. Journal of Hydrology 298 (1–4), 27–60.
Refsgaard, J.S., Knudsen, J., 1996. Operational validation and intercomparison of different types of hydrological models. Water Resources Research 32 (7), 2189–2202.
Shah, S.M.S., O'Connell, P.E., Hosking, J.R.M., 1996. Modelling the effects of spatial variability in rainfall on catchment response, 2, experiments with distributed and lumped models. Journal of Hydrology 175, 89–111.
Smith, M.B., Seo, D.-J., Koren, V.I., Reed, S.M., Zhang, Z., Duan, Q., Moreda, F., Cong, S., 2004. The distributed model intercomparison project (DMIP): motivation and experiment design. Journal of Hydrology 298 (1–4), 4–26.
Zhang, Z.Y., Koren, V., Smith, M., Reed, S., Wang, D., 2004. Use of next generation weather radar data and basin disaggregation to improve continuous hydrograph simulations. Journal of Hydrologic Engineering 9 (2), 103–115.