A Method for Evaluating the Accuracy of Quantitative Precipitation Estimates from a Hydrologic Modeling Perspective

JONATHAN J. GOURLEY

Cooperative Institute for Mesoscale Meteorological Studies, University of Oklahoma, Norman, Oklahoma

BAXTER E. VIEUX

Department of Civil Engineering and Environmental Science, University of Oklahoma, Norman, Oklahoma

(Manuscript received 6 May 2004, in final form 22 September 2004)

ABSTRACT

A major goal in quantitative precipitation estimation and forecasting is the ability to provide accurate initial conditions for the purposes of hydrologic modeling. The accuracy of a streamflow prediction system is dependent upon how well the initial hydrometeorological states are characterized. A methodology is developed to objectively and quantitatively evaluate the skill of several different precipitation algorithms at the scale of application—a watershed. Thousands of hydrologic simulations are performed in an ensemble fashion, enabling an exploration of the model parameter space. Probabilistic statistics are then utilized to compare the relative skill of hydrologic simulations produced from the different precipitation inputs to the observed streamflow. The primary focus of this study is to demonstrate a methodology to evaluate precipitation algorithms that can be used to supplement traditional radar–rain gauge analyses. This approach is appropriate for the evaluation of precipitation estimates or forecasts that are intended to serve as inputs to hydrologic models.

1. Introduction

It is well acknowledged that the accuracy of streamflow predictions from a hydrologic model is heavily dependent on the accuracy of the precipitation inputs. In addition, predictions from hydrologic models have been observed to be quite sensitive and nonlinear to perturbations in the initial conditions (Droegemeier et al. 2000). Entekahbi et al. (2002) identify the precipitation inputs as one of the major limitations to improved hydrologic predictability. There exists a need to systematically monitor the progress of improvements in precipitation algorithms in a statistically and hydrologically relevant way. A probabilistic approach is presented herein to evaluate different quantitative precipitation algorithms from the perspective of hydrologic model simulations.

Initial conditions describing the current or future rainfall states are supplied through quantitative precipitation estimation (QPE)/quantitative precipitation forecasting (QPF). These hydrologic model inputs can be derived from many meteorological models and observational sensors, such as rain gauges, satellites, weather radars, numerical weather forecasts, and combinations thereof (i.e., multisensor estimates). In the last two decades, many studies have demonstrated the capability of quantitatively estimating rainfall using multiple parameters that are available with polarization-diverse radar measurements (e.g., Doviak and Zrnic 1993; Zrnic and Ryzhkov 1999; Straka et al. 2000; Bringi and Chandrasekar 2001). As these QPE algorithms are being formulated, it is vital to the developers to know the error characteristics associated with the estimates.

Traditionally, improvements in rainfall algorithms have been measured by comparing the remotely sensed QPEs to rain gauges at collocated grid points. This verification methodology may not be sufficient for all of the intended applications of QPE/QPF. For example, many operational precipitation algorithms (e.g., Seo and Breidenbach 2002) incorporate rain gauge data in their QPE schemes. Thus, it becomes a challenge to utilize truly independent data sources for verification purposes (Young et al. 2000). This situation could potentially be remedied if some gauges were withheld from the estimation scheme and used for verification instead. This evaluation design potentially reduces the accuracy of the precipitation estimates and relies on a dense network of rain gauges.

Corresponding author address: J. J. Gourley, National Severe Storms Laboratory, 1313 Halley Circle, Norman, OK 73069.
E-mail: gourley@ou.edu

Several references site the lack of accuracy with rain gauge point measurements (e.g., Zawadzki 1975; Wilson and Brandes 1979; Marselek 1981; Legates and DeLiberty 1993; Nystuen 1999 and others). Underestimations are common with increasing wind speeds, and the gauges may only represent the rainfall that occurs in close proximity to the gauge. Highly variable rainfall may not be represented very accurately by a rain gauge network (Goodrich et al. 1995). Moreover, there are issues regarding the scales of measurement between rain gauges and remotely sensed rainfall. It has been noted that the sampling sizes between a typical radar pixel and a rain gauge orifice differ by about eight orders of magnitude (Droegemeier et al. 2000).

Evaluation of the hydrologic response to differing rainfall algorithm inputs (observed and simulated) for flash flood events has recently been explored (Peters and Easton 1996; Frank et al. 1999; Ogden et al. 2000; Warner et al. 2000a; Warner et al. 2000b; Yates et al. 2000; Sanchez-Diezma et al. 2001; Westrick and Mass 2001). The modeling system approaches to evaluation often recognize the nonlinearity in the rainfall–runoff transformation (Droegemeier et al. 2000); and, moreover, they note the complexity of the propagation of seemingly small errors in rainfall fields to much larger errors in predicted discharges (Faures et al. 1995; Frank et al. 1999; Ogden et al. 2000). A potential drawback in using such an approach is the introduction of model uncertainty, which is present in the physical representations of most environmental systems. The nonlinear nature of the physical transformation of rainfall to runoff may be exacerbated by uncertainties in the modeling process itself, further complicating the evaluation capabilities of a hydrologic prediction system.

A hydrologic approach to QPE evaluation may also become complicated because model parameters can be judiciously adjusted or calibrated to account for errors in model inputs (e.g., QPEs). Systematic biases, which are originally present in the model inputs, can be mitigated or corrected in order to yield accurate streamflow forecasts. In short, globally optimized model parameters can obscure uncertainties that were initially present in the model inputs. Probabilistic calibration methods exist, such as the generalized likelihood uncertainty estimation (GLUE; Beven and Binley 1992), to compute the probability that a given parameter set adequately simulates the observed system behavior. Furthermore, it is suggested in Freer et al. (1996) that the GLUE technique be expanded to include the uncertainties associated with different rainfall inputs. Extension of the GLUE methodology to account for differing rainfall inputs provides a consistent methodology to evaluate the hydrologic response to each input *independently*. At this time, no known studies have evaluated QPE algorithms using a hydrologic model in a probabilistic way that is independent of the calibration dataset.

A methodology is proposed herein that provides the framework to evaluate QPE algorithms at the hydrologic scale of application. The developed evaluation methodology assesses the skill of each rainfall algorithm after they have been submitted to the same degrees of model parametric uncertainty. Probabilistic measures are used to describe the accuracy of the ensemble of model simulations that were created using differing inputs. The stochastic approach proposed herein ensures that the probabilistic measures are computed independently of the inputs that comprised the calibration dataset and is, thus, completely objective. While the method is applied here to a single model on one basin, it is believed that this hydrologic analysis will serve as a tool for supplemental QPE algorithm assessment and refinement using additional operational rainfall–runoff models.

Section 2 provides a description of the quantitative precipitation estimates, the hydrologic model, and the test basin that is used for this study. The methodology of evaluating the QPE algorithms using a hydrologic model is presented in section 3. Three cases are used to demonstrate the developed approach in section 4. A summary of the case results and a discussion follows in section 5.

2. Background

Precipitation estimates generated from a multisensor algorithm under development at the National Severe Storms Laboratory are evaluated in this study. This algorithm, called Quantitative Precipitation and Estimation and Segregation Using Multiple Sensors (QPE SUMS; Gourley et al. 2001) ingests raw radar reflectivity data, satellite infrared imagery, numerical model fields, and rain gauge observations to produce an ensemble of precipitation products. Each of these products has various levels of complexity. It is of particular interest to algorithm developers to understand the error characteristics of each product and to determine the amount of improvement yielded by the addition of sensors to relatively simple gauge-based products. For instance, do the rain gauge data alone provide an initial rainfall state that is detailed enough for accurate hydrologic simulation? Precipitation estimates from each member comprising the QPE SUMS ensemble are evaluated herein; thus, a brief algorithm overview follows.

Rain gauge observations are incorporated on an hourly basis from the Oklahoma Mesonetwork (Mesonet) to produce a gauge-only (GAG) rainfall product. The point estimates are analyzed on a 1 km $\times$ 1 km common grid using a Barnes objective analysis scheme (Barnes 1964). The GAG product serves as a benchmark for comparing and improving products that rely on remote sensors. Radar reflectivity data from the Twin Lakes, Oklahoma (KTLX), and Fort Worth, Texas (KFWS), Next Generation Weather Radar (NEXRAD) sites are resampled on a common Carte-

sian grid using the reflectivity that was measured closest to the ground (O'Bannon 1997). Next, the mosaicked reflectivity data are converted to rainfall rates using the default, empirically derived convective relationship from Woodley et al. (1975),

$$Z = 300R^{1.4}, \quad (1)$$

where Z is in mm^6 mm^{-3} and R is in mm h^{-1}. The rainfall estimates are aggregated in time to yield hourly rainfall accumulations that are available at the top of each hour. This particular rainfall member relies almost solely on radar data and is, thus, referred to as the radar-only (RAD) rainfall product.

A more complex precipitation product incorporates information from radar, numerical models, and infrared satellite data. This multisensor (MS) algorithm begins by automatically searching for the melting layer (Gourley and Calvert 2003), and then discarding the contaminated reflectivity values that were measured in this region. This zone of elevated reflectivity, or radar bright band (Austin and Bemis 1950), can cause precipitation overestimation up to a factor of 10 if it is not accounted for (Smith 1986). The multisensor algorithm builds upon the infrared satellite and radar regression technique that was prototyped in Gourley et al. (2002) to supply meaningful precipitation rates to stratiform pixels at the intermediate and far range where radar estimates of rainfall are known to be inaccurate (see, e.g., Joss and Waldvogel 1990; Fabry et al. 1992; Kitchen and Jackson 1993; Smith et al. 1996; Seo et al. 2000). A goal in this study is to analyze the accuracy of the multisensor approach to QPE on hydrologic simulations.

Two different gauge-adjustment strategies are implemented on the remotely sensed rainfall estimates (i.e., RAD and MS products). The biasing techniques are designed to 1) apply a mean field bias to the rainfall fields (Wilson and Brandes 1979), while maintaining the spatial details obtained by the remotely sensed data, or 2) adjust the RAD and MS field by applying a spatially nonuniform bias (Seo and Breidenbach 2002), which may result in a smoother rainfall field. The former, mean field bias approach (-G hereafter) is implemented on both RAD and MS products to yield the RAD-G and MS-G products. The latter method, referred to as a local bias adjustment (-LG hereafter), can be designed to produce perfect agreement with rain gauge observations. However, the spatial details present in the original radar data may be lost, and can have an impact on hydrologic simulations. Products utilizing local bias adjustments are referred to as RAD-LG and MS-LG. Determining the impact of different gauge adjustment strategies on hydrologic simulations is another objective in this study.

Hourly precipitation estimates from all seven QPE SUMS products are input to a hydrologic model to simulate river discharges. The hydrologic model used in this study is a commercial model called Vflo (Vieux and Vieux 2002). The Vflo model treats parameters and inputs in a spatially distributed fashion, and flow simulations are allowed to vary with time, that is, be unsteady. The governing equations are derived from conservation principles while the model parameters are calibrated by scalars using the ordered physics-based parameter adjustment (OPPA) method described by Vieux and Moreda (2003). Documentation of the model formulation is discussed in Vieux et al. (2004), and a summary is provided here in appendix B.

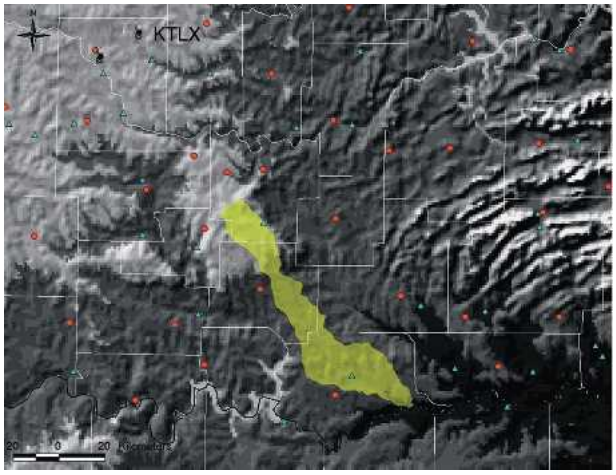

FIG. 1. The study area, including the Blue River basin (shaded in yellow), Weather Surveillance Radar-1988 Doppler (WSR-88D), Mesonet rain gauge locations (red filled circles), USGS gauging sites (light blue filled triangles), and a shaded relief map of Oklahoma.

Model simulations are performed on the Blue River basin, which is an existing natural outdoor laboratory for hydrologic research. The Blue River basin drains about 1200 km^2. The headwaters of the basin are approximately 80 km away from the nearest weather radar, KTLX, while the basin outlet is over 200 km away from KTLX (see Fig. 1). Radar-based rainfall estimates at a far range may have errors due to beams overshooting shallow profiles of reflectivity. The long distance from radars makes this basin a good candidate for evaluating QPE algorithms from a moderate to far range. The Blue River basin is also attractive for hydrologic research due to the natural characteristics of the basin. There are no reservoirs in the basin, and there are very few known diversions.

Hydrologic forecasts are evaluated using U.S. Geological Survey (USGS) hourly discharge observations on the Blue River near Blue, Oklahoma (site number 07332500). The hydrologic evaluation focuses on three cases that resulted in significant flow on the Blue River. Details about each case, including data availability, are summarized in Table 1.

3. Hydrologic evaluation methodology

The first step in setting up a model parameter ensemble involves assigning the ranges and distributions

TABLE 1. A summary of the hydrologic events studied. The observed time refers to the time of peak discharge, the observed peak is the magnitude of the peak discharge, and the observed volume is the time-integrated discharge normalized by the basin area. See section 2 for a complete description of the rainfall products.

Case start time	Case end time	Observed time	Observed peak (cms)	Observed volume (mm)	Available rainfall products
0000 UTC 23 Oct 2002	0000 UTC 30 Oct 2002	0000 UTC 26 Oct 2002	56.7	6.1	GAG, RAD, RAD-G, RAD-LG, MS, MS-G, MS-LG
0000 UTC 28 Oct 2002	0000 UTC 1 Nov 2002	0600 UTC 30 Oct 2002	16.4	2.3	GAG, MS, MS-G, MS-LG
0000 UTC 3 Oct 2002	0000 UTC 10 Dec 2002	1600 UTC 5 Dec 02	13.4	2.5	GAG, RAD, RAD-G, RAD-LG

of parameter values. If no information is known a priori about the parameter distributions, then a uniform distribution can and will be assumed. The Vflo model utilizes maps of the saturated hydraulic conductivity (K), initial fractional water content of the soils (θ), soil suction at wetting front (ψ), bed slope (S_o), flow direction, flow accumulation, and Manning roughness coefficient (n) that are distributed spatially (see appendix B). The digital elevation model (DEM)-derived parameters, that is, S_o, flow direction, and accumulation, are well-defined, and variations of these parameters are related to scale issues that are beyond the scope of this study. The ψ parameter is inversely proportional to K, as revealed in Table 4.3.1 of Chow et al. (1988). Varying the ψ parameter in addition to the K parameter increases the dimensionality of the parameter space, thus, increasing the number of simulations by an order of magnitude. For these reasons, the K, θ, and n parameters are assigned ranges, while the other parameters remain fixed at their predefined values. Soil properties may be inferred from State Soil Geographic (STATSGO) survey maps, while depictions of the Manning roughness coefficient may be inferred from readily available land use/land cover maps.

Channel characteristics such as the cross-sectional area, geometry, channel side slope, and hydraulic roughness must be specified for each stream. Channel cross sections were obtained from site surveys that were conducted as part of the Distributed Model Intercomparison Project (DMIP; Smith et al. 2004) and are shown in Vieux et al. (2004). Uncertainty and representativeness of these measurements are not known, but are assumed to be negligible as compared to uncertainty in the model parameters and inputs. Investigating the contribution of uncertainty from channel hydraulic parameters is an area inviting future research.

The K, θ, and n parameters are distributed in space and are derived from ancillary data. The inference of K and n from soil types and land use/land cover data is associated with some uncertainty because they are not measured directly. For this reason, they must be perturbed within their physical bounds. Perturbing spatially distributed parameter maps can be a daunting task if one were to consider perturbing the values from cell to cell. The parameter space would increase beyond bounds that are not practical to sample and perform model simulations. The ordered physics-based parameter adjustment method described by Vieux and Moreda (2003) employs scalars to adjust parameter maps so that the magnitudes change while the spatial variation is preserved. This method maintains spatial variability that is commensurate with distributed parameter modeling, while minimizing the dimensionality of the parameter space. The scalars used to multiply the K and n parameter maps are defined as follows:

$$N_i = \frac{1}{8}(2 + 3i)|_{i=0,1,2,3,4}, \tag{2}$$

where N_i is the adjustment factor. In essence, (2) employs scalars that adjust parameters from 25% to 175% of their given values. Perturbing the Manning roughness coefficients results in smooth (rough) surfaces that have the affect of reducing (increasing) the amount of time it takes water moving overland to reach the basin outlet. Increasing (decreasing) the saturated hydraulic conductivity has the affect of increasing (decreasing) potential infiltration rates. In this case, more (less) water is lost to the soils and a lower (higher) volume of water reaching the basin outlet results. The initial soil saturation parameter is varied from 20% (dry) to 100% (saturated) in increments of 20%. The Vflo model is run on an event-based mode; thus, there is no known information about spatially distributed antecedent soil

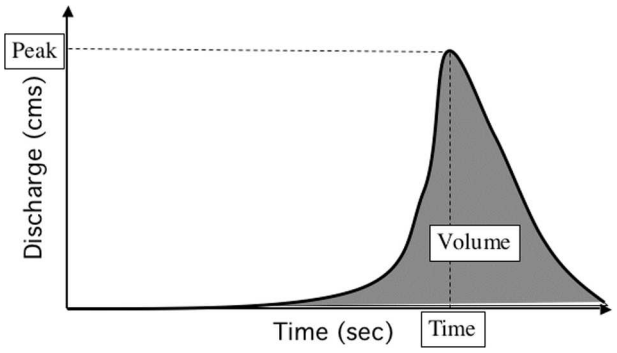

FIG. 2. A schematic of a typical hydrograph and the derived variables that are used in this study. The time (s) variable is the time at which the maximum discharge occurs, peak (cms) is the magnitude of the maximum discharge, and volume is the time-integrated flow under the hydrograph that has been normalized by the basin area (mm).

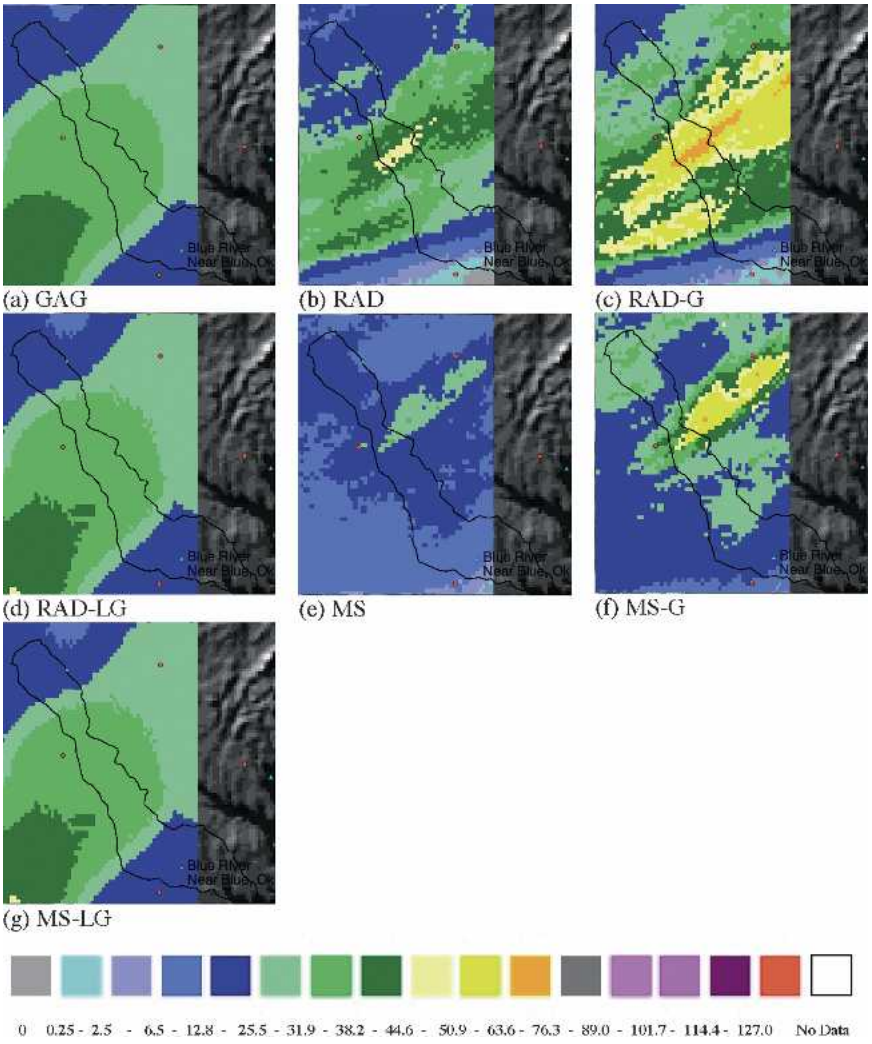

FIG. 3. Storm total precipitation plots for the 23 Oct 2002 case from the QPE SUMS products: (a) GAG, (b) RAD, (c) RAD-G, (d) RAD-LG, (e) MS, (f) MS-G, and (g) MS-LG. A complete description of each algorithm is provided in section 2. Values are in mm. Red dots correspond to Mesonet rain gauging locations and cyan triangles are USGS stream gauges.

moisture conditions. The parameter is essentially varied from dry to moist conditions.

The next procedure in setting up a model parameter ensemble involves running multiple hydrologic simulations using combinations of possible parameter settings for each QPE input. If there were information available regarding the initial parameter distributions, then an algorithmic procedure would be devised to sample the parameter space strategically to avoid excessive computational expense. A uniform distribution is assumed here, and the parameter space is sampled thoroughly such that no particular parameter setting is favored and given more weight in computing the final probabilities.

The first simulation uses rainfall inputs from the GAG product for the 23 October 2002 event (Table 1). Scalars used to multiply the n and K maps are determined by (2), while the first parameter setting of θ is set to 20%. Results from this simulation are stored in a file. This procedure is repeated over and over until all user-specified parameter combinations have been utilized in the model. Because a total of five different scalars are used to adjust each of the three parameters, a total of 125 simulations are performed for the GAG rainfall input for the first event. Next, the RAD product is input to the Vflo model in lieu of the GAG product. The exact same parameters are iterated through the model, with the only difference being the model inputs. Again, model simulations using the RAD inputs are stored in a separate file. This procedure is repeated for all available rainfall inputs for the 23 October 2002, 28 October 2002, and 3 December 2002 events (see Table 1).

TABLE 2. Statistical evaluation of hydrologic simulations for the 23 Oct 2002 case. Detailed definitions of the statistics employed are provided in appendix A. Numbers in boldface indicate the best agreement with observations.

Time	RO coef	Bias	MAE (min)	Rmse (min)	RPS
GAG	0.21	0.95	0.39	0.43	1.10
RAD	0.23	0.95	0.39	0.43	1.02
RAD-G	0.18	**0.98**	**0.25**	**0.33**	**0.32**
RAD-LG	0.16	0.94	0.36	0.44	0.99
MS	0.39	0.93	0.59	0.59	2.14
MS-G	0.22	0.94	0.63	0.63	1.79
MS-LG	0.19	0.94	0.37	0.42	1.04
Peak	RO coef	Bias	MAE (cms)	Rmse (cms)	RPS
GAG	0.21	1.31	23.67	36.65	0.44
RAD	0.23	1.27	**21.39**	31.02	**0.43**
RAD-G	0.18	1.87	49.17	63.56	1.87
RAD-LG	0.16	2.17	66.12	88.24	2.42
MS	0.39	0.58	26.45	**29.62**	1.21
MS-G	0.22	**1.15**	21.96	30.39	**0.43**
MS-LG	0.19	1.54	31.52	49.57	0.78
Volume	RO coef	Bias	MAE (mm)	Rmse (mm)	RPS
GAG	0.21	2.03	6.27	8.86	1.86
RAD	0.23	1.96	5.87	7.97	1.88
RAD-G	0.18	2.79	10.89	12.98	4.11
RAD-LG	0.16	2.84	11.22	13.53	4.14
MS	0.39	**1.15**	**2.31**	**3.42**	**0.51**
MS-G	0.22	1.96	5.87	8.41	1.65
MS-LG	0.19	2.29	7.87	10.47	2.58

Results from each simulation are compared to observed streamflow in the verification step. There are several ways in which simulations of streamflow can be evaluated based on comparison to observations. Typically, a particular model's goodness of fit is determined by using objective functions that rely on the sum of squared errors (see, e.g., Beven and Binley 1992; Romanowicz et al. 1994). An individual error results from the difference between discharge simulations and observations at each observational time step. The error variance is then computed throughout an event or during an entire season. The objective functions described above have useful statistical properties, but can result in a loss of information about a given hydrograph. Consider, for example, a simulated hydrograph that mirrors the observed hydrograph perfectly, but is offset in time by a few hours. The errors computed at each observational time step would be significant in this case, as would be the error variance throughout the event. The objective functions described above would have little capability in differentiating this kind of a simulated hydrograph versus another that has a significantly different shape. Similar magnitudes in error variances can be achieved from simulated hydrographs that have much different shapes and behavior. This study utilizes three objective functions to describe the degree to which each model matches the streamflow observations. These objectives provide more information about hydrograph shapes than statistics based on error variances.

The following variables are computed from the observed discharge and each simulation of streamflow: integrated discharge throughout the storm event normalized by the basin area (volume hereafter), maximum discharge throughout the event (peak hereafter), and the time at which the maximum discharge occurred (time hereafter). Figure 2 shows how each of these variables are derived from a given hydrograph. For the purposes of QPE development, the volume is the most informative variable. Recall that the outlet of the Blue

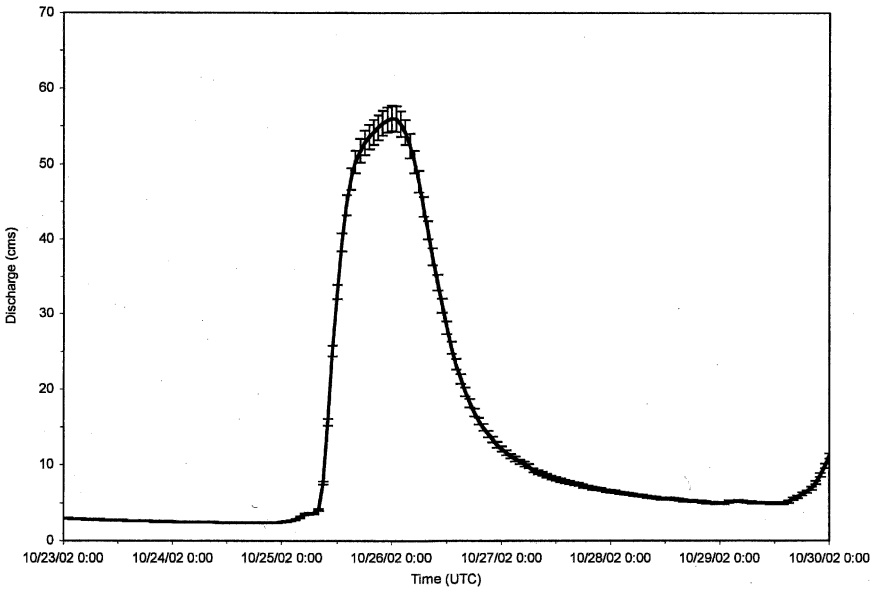

FIG. 4. Observed streamflow at the Blue River, near Blue, OK, gauging site (USGS number 07332500) for the 23 Oct 2002 case. Error bars indicate ± 3% uncertainty associated with streamflow estimates.

River basin is 200 km from a nearby radar, KTLX (see Fig. 1). The extent to which each product may possess biases, especially at medium to far ranges from the radar, is important in assessing the relative strengths and weaknesses of a particular QPE. These biases will be revealed most explicitly upon analysis of the volume results.

The mean time, peak, and volume of each ensemble is computed. It is recognized that observations of river discharge, including the recording times, are associated with some uncertainty (Sauer and Meyer 1992). In fact, a correction factor was added to the recording times to convert from local time to UTC. Discharges are inferred from measurements of stage height through the use of rating curves. This technique has been shown to have greater uncertainty with high flows (Bradley and Potter 1992). For this application, it is quite probable that uncertainties in the model inputs and parameters overshadow the streamflow measurement uncertainties. Nonetheless, streamflow measurements and derived quantities (e.g., time, peak, and volume) cannot be considered perfectly accurate. The following statistics are computed in order to compare the ensemble averages to observations: *bias*, *mean absolute error* (*MAE*), and *root-mean-squar error* (*rmse*) (see appendix A). Observed rainfall is also compared to observed streamflow using a runoff coefficient (RO coef). This quantity is the discharge volume divided by the basin-averaged precipitation that is summed throughout each event.

The statistics described above are simplistic and provide a first-order evaluation of the hydrologic variables as compared to observations. However, they only consider the ensemble mean and not the full spread and distribution of predictions. A forecast ensemble provides the *capability* to estimate the forecast probability distribution function (pdf) of the hydrologic variables. Probability distribution functions are computed for the time, peak, and volume using each a 125-member ensemble representing the different QPE inputs. In many applications, a histogram would be a sufficient qualitative manner in which to display the distribution of a given dataset. However, it must be noted that a judicious choice of the number of classes used and the class interval can affect the appearance and conclusions drawn from the computed histograms. Moreover, histograms are not always smooth or continuous, which reduces their utility if derivatives are needed. This study computes pdfs for each ensemble using Gaussian kernel density estimation (Silverman 1986) (see appendix A). In a simplified manner, this density estimation technique can be thought of as a smoothed histogram. The resulting function is smooth and continuous, which leads to better estimation of probability exceedence (i.e., 90% simulation limits) and calculation of derivatives.

Measures of central tendency, spread, and skill are computed from each pdf. A nonparametric measure of central tendency is the 50% simulation limit, or median. This is the value corresponding to 0.5 on the cumulative distribution function (cdf). Similarly, the spread can be quantified by determining the distance between the 5% and 95% simulation bounds. Similar to the median, these simulation limits are determined from the values corresponding to 0.05 and 0.95 on the cdf. This provides for measures of central tendency and uncertainty limits for each ensemble. Plots from these measures are produced for each ensemble, thus, enabling the subjective comparison of the simulation limits.

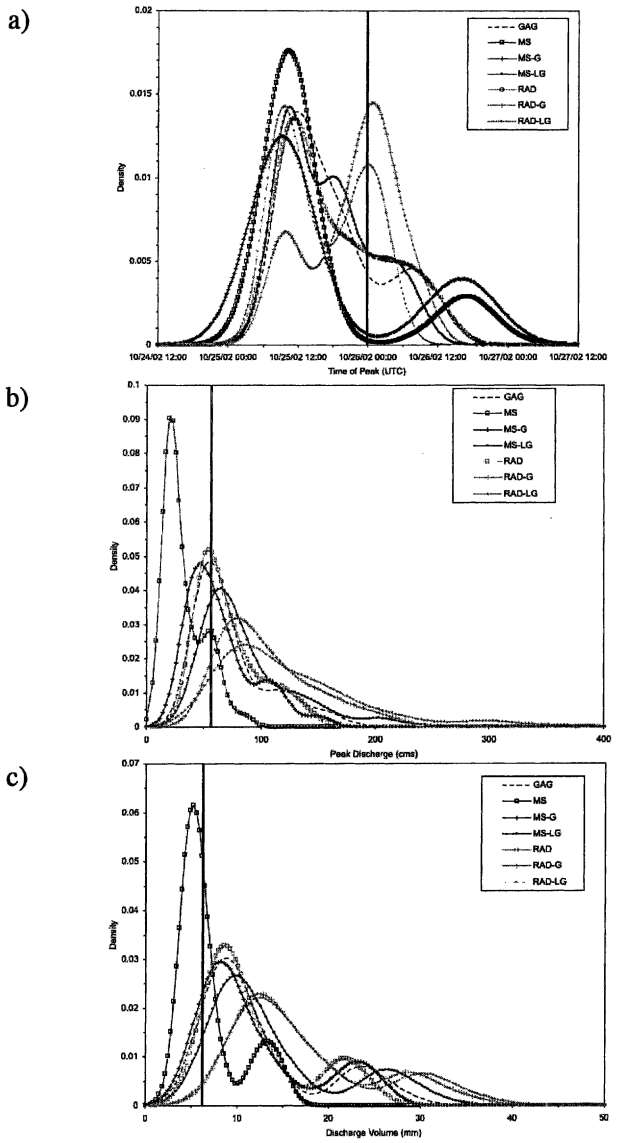

FIG. 5. Probability density functions of simulated (a) time, (b) peak, and (c) volume for the 23 Oct 2002 case. The coding of the curves (see the legend) corresponds to the different model inputs used to construct the model parameter ensembles. The vertical double line is the observed variable.

TABLE 3. Significance levels of RPS differences for the 23 Oct 2002 case. Significance levels are determined using a resampling test (see appendix A).

Time	GAG	RAD	RAD-G	RAD-LG	MS	MS-G	MS-LG
GAG	0.07	0.55	0.99	0.57	0.99	0.99	0.40
RAD		0.08	0.99	0.27	0.99	0.99	0.19
RAD-G			0.11	0.99	0.99	0.99	0.99
RAD-LG				0.04	0.99	0.99	0.38
MS					0.05	0.96	0.99
MS-G						0.06	0.99
MS-LG							0.06
Peak	GAG	RAD	RAD-G	RAD-LG	MS	MS-G	MS-LG
GAG	0.10	0.26	0.99	0.99	0.99	0.32	0.98
RAD		0.09	0.99	0.99	0.99	0.33	0.98
RAD-G			0.04	0.97	0.99	0.99	0.99
RAD-LG				0.04	0.99	0.99	0.99
MS					0.06	0.99	0.99
MS-G						0.13	0.99
MS-LG							0.06
Volume	GAG	RAD	RAD-G	RAD-LG	MS	MS-G	MS-LG
GAG	0.06	0.13	0.99	0.99	0.99	0.72	0.99
RAD		0.05	0.99	0.99	0.99	0.78	0.99
RAD-G			0.05	0.25	0.99	0.99	0.99
RAD-LG				0.04	0.99	0.99	0.99
MS					0.12	0.99	0.99
MS-G						0.05	0.99
MS-LG							0.05

It would be informative to the modeler to know the ensemble skill based on the pdfs. This approach is nonparametric and considers each member with equal weight. Statistics that are used to evaluate probability forecasts for multicategory events have been developed (Wilks 1995). A condition for a probability distribution is that the probability for a given event must be greater than or equal to zero and less than or equal to one. Second, the sum of probabilities for the event must be one. The pdfs computed from the ensembles may not be an accurate estimate of the true underlying pdf if there are uncertainties in the model structure, model inputs, or observations of model predictands. Thus, the pdfs computed here are treated as conditional probability distributions. The ensemble skill is assessed by using the *ranked probability score* (*RPS*; Wilks 1995) (see appendix A). This score essentially compares the entire pdf of the simulated variables to a single observation of time, peak, or volume. Simulations that are far removed from the observation are penalized more heavily than those falling into nearby categories. Last, the statistical significance of differences in RPS values is assessed using a resampling technique (see appendix A).

4. Application of the hydrologic evaluation on the Blue River basin

A unique methodology has been developed to evaluate the relative skill of hydrologic simulations using different QPE inputs. The results from this study reveal the precipitation estimator that is most likely to provide the best input to the Vflo model because it applies to the Blue River basin in Oklahoma. Conclusions stemming from these results may not be directly applicable to other basins or other hydrologic models. The focus of this study is on the methodology of evaluating QPEs by using a given hydrologic model. If the model parameter ranges are known a priori, then this methodology can be applied without prior modeling experience (i.e., calibration) on the basin of interest. Moreover, the ensemble approach is comprehensive, because it samples parameter settings within user-specified ranges, and is objective in that it is not designed to favor a particular model input.

a. 23 October 2002 case

A subjective comparison of the rainfall products is accomplished by simply aggregating the hourly accumulations over the duration of the entire storm. Storm total plots for the 23 October 2002 case are shown in Figs. 3a–g. Visual differences are apparent between the panels indicating the dependence of the rainfall patterns on the sensor(s) that are used to produce the estimates. For example, the gauge-only product (Fig. 3a) is quite smooth in comparison with the radar-only (Fig. 3b) product. The spatial and temporal resolution of the radar data is observably superior, however, the radar does not measure rainfall directly like the rain gauges. The multisensor (Fig. 3e) product bears some resemblance to the radar-only product in terms of the maximum rainfall. However, the MS product is noticeably smoother and reflects the use of satellite data in its scheme. Figures 3c, 3d, 3f, and 3g show how the different gauge-adjustment strategies impact the RAD and MS products. Mean field bias adjustments (-G) main-

tain the spatial structure in the original product while the magnitude has been scaled. In this case, the adjustment procedure results in heavier accumulations of rainfall in the RAD-G and MS-G products as compared to the unadjusted RAD and MS products. Local bias adjustments (-LG) to the RAD and MS products yield rainfall patterns that resemble the unadjusted, remotely sensed rainfall, yet ensure agreement with individual rain gauge amounts through application of a spatially nonuniform bias.

Table 2 shows a statistical comparison of hydrologic simulations from each input to observations of streamflow. The observed streamflow for this case is shown in Fig. 4. All statistics indicate that time is simulated most accurately using inputs from the RAD-G rainfall product. Peak values are simulated accurately using inputs from both RAD and MS-G products. All of the statistics regarding the volume hydrologic variable indicate the MS product *without any gauge adjustment* is the most accurate input after it has been run through the model using all of the possible parameter combinations.

The computed pdfs for time, peak, and volume are shown in Figs. 5a–c. The pdfs supply additional information about the distributions of the hydrologic simulations that are not as apparent in the statistical analysis. For example, Fig. 5a reveals that the simulations have a distinct bimodal shape. This double peak indicates the parameter settings that are used in the Vflo model to produce two different modes of behavior for all of the model inputs. Figures 5b–c also show that the pdfs have a bimodal shape. The MS pdf appears to be significantly different from the other pdfs for simulations of peak and volume. As it turns out, the statistical levels of the RPS differences between the MS input and all other inputs exceed 99% (see Table 3).

The final component of the hydrologic evaluation for this case involves deriving the region in the phase space where 90% of the simulations fall. Plots of the 5%, 50%, and 95% quantiles are shown in Figs. 6a–c. The observed hydrologic variables are shown in these plots as a double line. This convention is a reminder that observations of streamflow have their own uncertainty, but have been found to be only 3%–6% (Sauer and Meyer 1992). Figure 6a shows that all ensembles envelop the observed time regardless of the rainfall product that is being used as input. The central tendency of the RAD-G input indicates that this product is capable of producing the most accurate time predictions for this case. The capabilities of each ensemble to simulate the peak discharge are revealed in Fig. 6b. Similar to Fig. 6a, the simulation bounds from all ensembles encompass the observed peak. The relatively low uncertainty with the RAD inputs, despite a larger bias, result in nearly the same skill score as with the MS-G inputs (Table 2). Figure 6c shows that volume simulations using inputs from the RAD-G and RAD-LG rainfall products do not give the observed system behavior. The

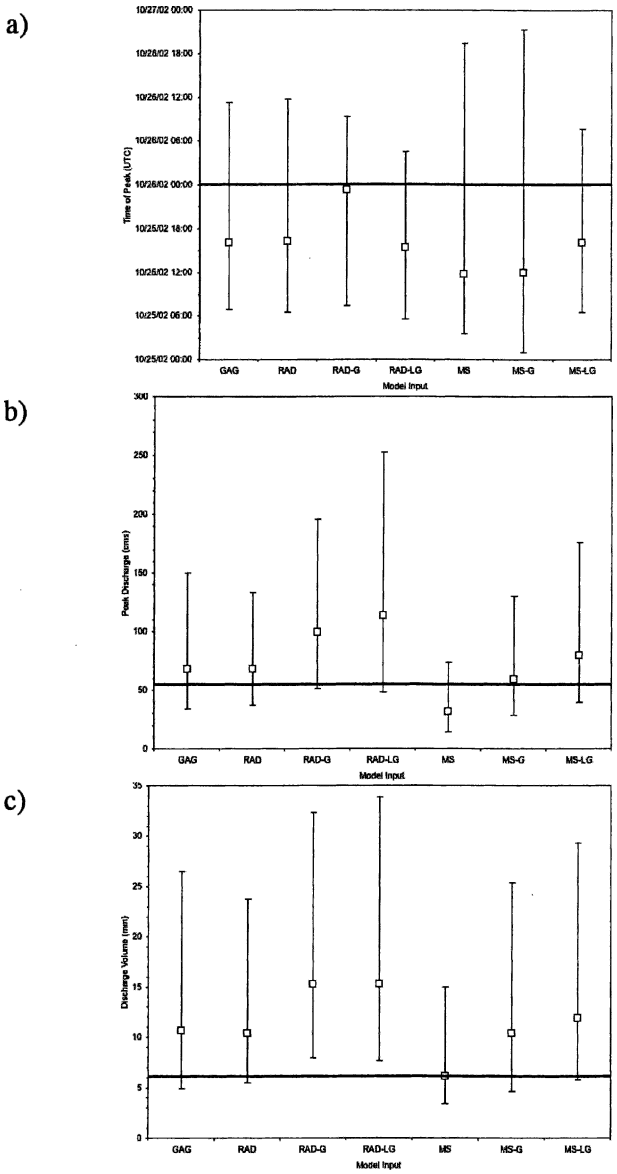

FIG. 6. Simulation bounds of the simulated (a) time, (b) peak, and (c) volume for the 23 Oct 2002 case. The open boxes refer to the 50% quantile (median), while the bars correspond to the 5% and 95% quantiles. The horizontal double line is the observed variable.

rainfall amounts from these two algorithms are too high, which results in 90% of the volume predictions being overforecast. The median of the total discharge volume simulations using MS inputs agrees rather well with the observed volume. In addition, the uncertainty bounds are relatively small with this rainfall input. A better RPS skill score results.

The hydrologic evaluation for the 23 October 2002 case indicates that the determination of the most accurate precipitation input depends on the hydrologic objective (i.e., time, peak, or volume). The MS algorithm

as an input, without any rain gauge adjustments, results in the most accurate hydrologic predictions for volume. It should be noted, however, that conclusions based upon this analysis are conditioned on perfect model physics and observations of the streamflow. The perfect model physics assumption is tested in Gourley and Vieux (2004). Observational uncertainty with streamflow measurements are relatively small and can be neglected, but potential errors in the modeling process itself can be structured to favor a given QPE input that may or may not be biased. Nonetheless, the developed methodology informs the modeler of the preferred input for the most accurate hydrologic simulations for their model on a particular basin.

b. 28 October 2002 case

The next case used for QPE evaluation yielded approximately 60% less volume than from the 23 October 2002 case discussed above (refer to Table 1). Rainfall products from the RAD, RAD-G, and RAD-LG algorithms are not available for this hydrologic evaluation. The overall skill of the available algorithms is determined as well as the relative accuracy of different gauge adjustment strategies. The storm total precipitation amounts are shown in Figs. 7a–d, while the observed hydrograph is shown in Fig. 8.

Time is simulated most accurately using either the MS-G or MS-LG rainfall estimates as inputs to the model parameter ensembles (Table 4). Conclusions regarding peak simulations can differ when considering the mean of the ensemble predictions versus the entire pdf. The MS input has the best bias, MAE, and rmse, while the MS-G algorithm produces the best RPS score when the entire ensemble of peak simulations is considered. Volume simulations are most skillful using MS estimates as input to the model parameter ensemble. Statistics describing the ensemble mean and the entire pdf support this conclusion, while the products that rely more heavily on rain gauge estimates (GAG and MS-LG) result in worse RPS scores. Statistical significance of differences are provided in Table 5.

The derived pdfs from the ensembles reveal additional information about the hydrologic simulations produced with the different rainfall inputs (Figs. 9a–c). Similar to the analysis of the pdfs in the 23 October 2002 case, two distinct modes of behavior are noted with the simulations. The shapes of pdfs help explain the discrepancy in the statistics that rely on the ensemble mean versus the entire pdf for the peak simulations (Table 4). For example, the mean of the peak simulations from the MS ensemble is sensitive to the secondary relative maximum. Mean field bias adjustments increase the magnitude of MS peak simulations, resulting in better agreement with observations as indicated by the low RPS value (Table 4). Volume simulations using the MS algorithm are clustered near lower values relative to the other inputs (Fig. 9c), which is also the case with the previous event.

Figure 10a shows the 90% simulation bounds and

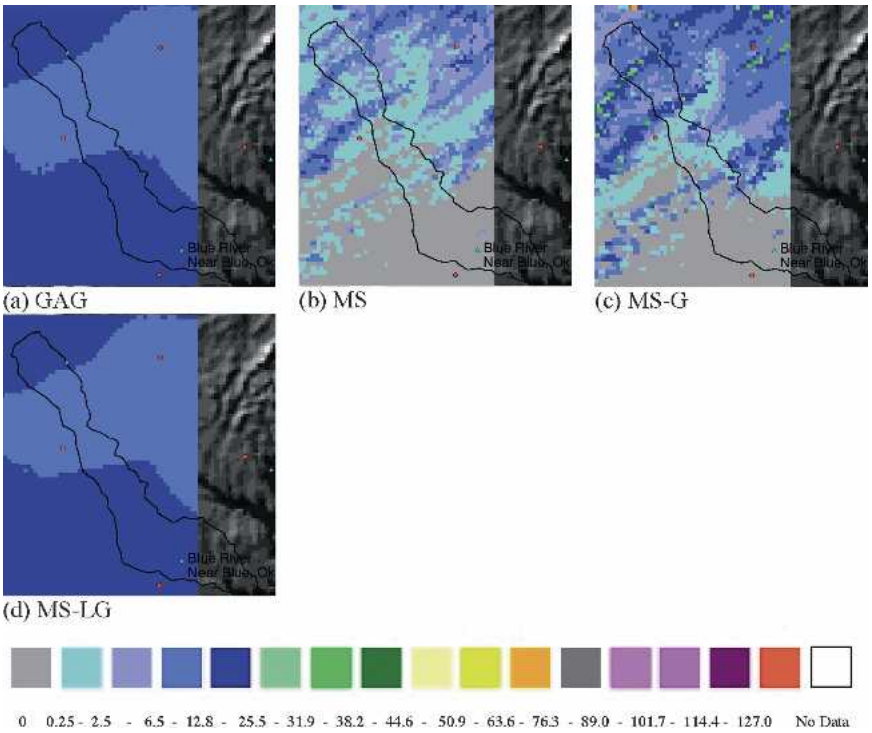

FIG. 7. As in Fig. 3, but for storm total precipitation for the 28 Oct 2002 case.

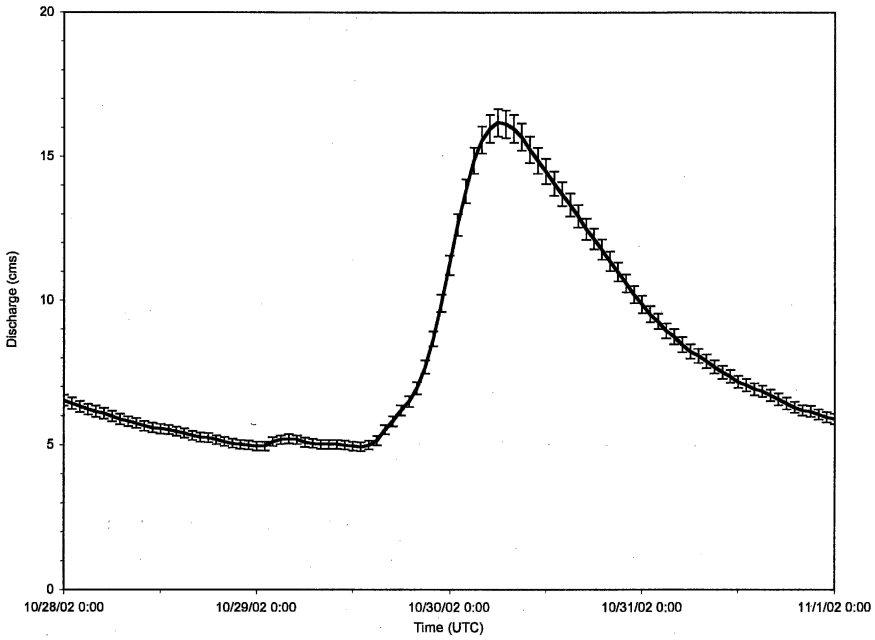

FIG. 8. As in Fig. 4, but for observed streamflow for the 28 Oct 2002 case.

median for the timing of the peak discharge using different rainfall algorithms as model inputs. The spread is a result of parametric uncertainty. Rainfall inputs that rely heavily on rain gauge data (e.g., GAG and MS-LG) produce simulations that encompass the observed time to peak (Fig. 10a). Peak and volume simulations that use either MS or MS-G inputs result in better agreement with observations. Adjustment procedures that rely heavily on *individual* gauge amounts do not result in more accurate peak or volume simulations from the hydrologic model. Figures 7a and 7d indicate that the techniques that are used to adjust QPEs based on point data result in smooth, interpolated accumulations, especially in the lower reaches of the Blue River basin. Note that there are no rain gauge measurements collected by the Mesonet within the Blue River basin. Products that rely more heavily on remote sensing systems (i.e., RAD and MS) resolve no rainfall at all near the basin outlet. Moreover, notice how rainfall maxima in the MS field (narrow part of basin; Fig. 7b) occur in regions where minima occur in the GAG and MS-LG accumulation products (Figs. 7a and 7d). Similar to the 23 October 2002 case, recognition of the most accurate QPE inputs depends on the hydrologic variable being simulated. Peak and volume are forecast best by using input from the MS-G and MS algorithms, respectively. This case indicates that gauge-adjustment strategies that smooth the rainfall field cause a loss in spatial resolution, and poorer hydrologic simulations result.

c. 3 December 2002 case

The final case that is used for the hydrologic evaluation component of this study produced a slightly larger

TABLE 4. As in Table 2, but for the 28 Oct 2002 case.

Time	RO coef	Bias	MAE (min)	Rmse (min)	RPS
GAG	0.14	0.88	0.69	0.69	3.05
MS	0.74	1.25	1.26	1.31	3.65
MS-G	0.42	1.12	**0.58**	**0.61**	2.13
MS-LG	0.13	**0.91**	0.64	0.67	**2.11**
Peak	RO coef	Bias	MAE (cms)	Rmse (cms)	RPS
GAG	0.14	2.99	32.64	40.20	4.28
MS	0.74	**0.63**	**8.10**	**8.74**	1.25
MS-G	0.42	1.43	9.10	13.73	**0.69**
MS-LG	0.13	3.30	37.69	44.62	4.88
Volume	RO coef	Bias	MAE (mm)	Rmse (mm)	RPS
GAG	0.14	3.27	5.23	6.25	5.14
MS	0.74	**1.31**	**0.71**	**0.95**	**0.70**
MS-G	0.42	1.79	1.82	2.14	2.19
MS-LG	0.13	3.75	6.32	7.22	5.97

TABLE 5. As in Table 3, but for the 28 Oct 2002 case.

Time	GAG	MS	MS-G	MS-LG
GAG	0.05	0.99	0.99	0.99
MS		0.04	0.99	0.99
MS-G			0.08	0.42
MS-LG				0.05
Peak	GAG	MS	MS-G	MS-LG
GAG	0.05	0.99	0.99	0.99
MS		0.07	0.99	0.99
MS-G			0.07	0.99
MS-LG				0.05
Volume	GAG	MS	MS-G	MS-LG
GAG	0.08	0.99	0.99	0.99
MS		0.10	0.99	0.99
MS-G			0.06	0.99
MS-LG				0.23

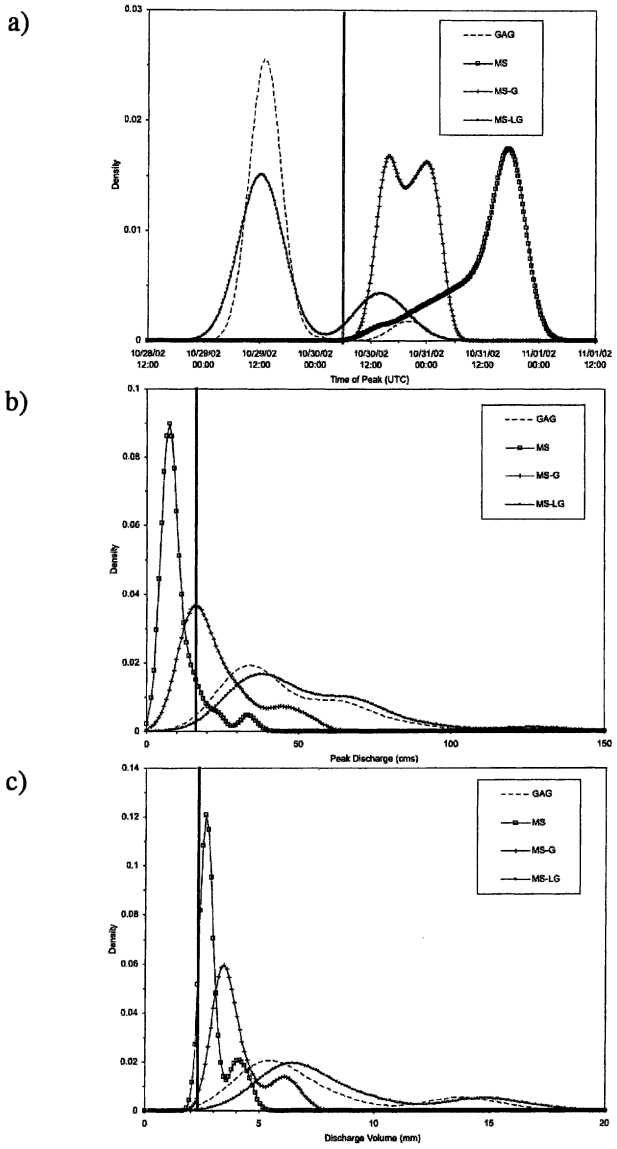

FIG. 9. As in Fig. 5, but for probability density functions for the 28 Oct 2002 case.

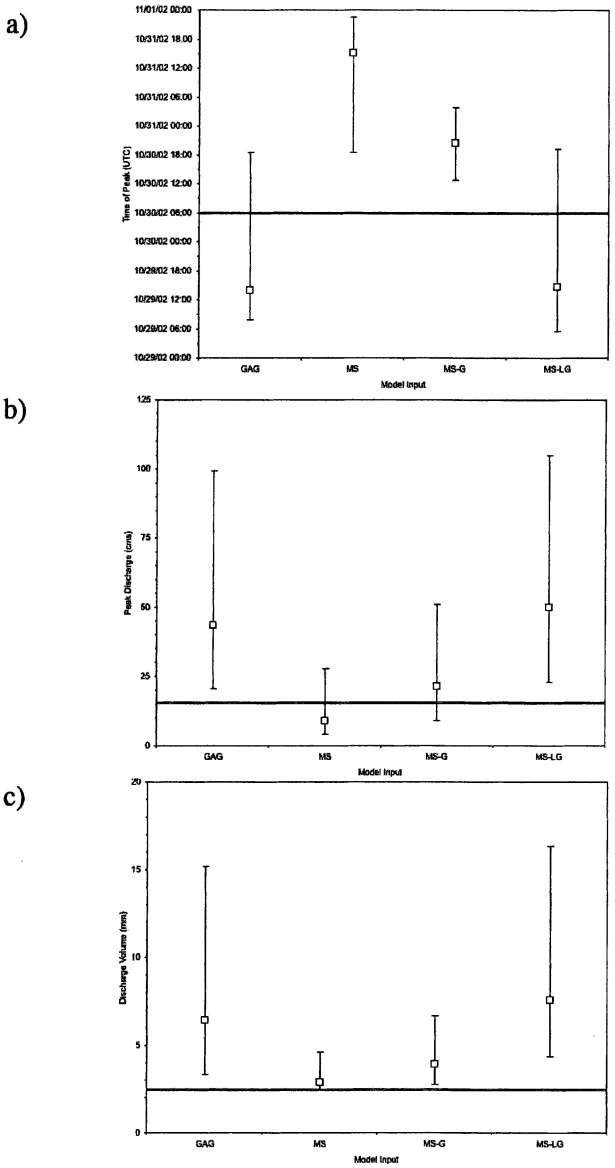

FIG. 10. As in Fig. 6, but for simulation bounds for the 28 Oct 2002 case.

volume but smaller peak discharge than the 28 October 2002 case (Table 1). The storm total accumulations are shown in Figs. 11a–d. Multisensor products are not available for this case. The observed hydrograph is shown in Fig. 12. There is a general agreement in the amounts and locations of precipitation over the Blue River basin between the independent GAG (Fig. 11a) and RAD (Fig. 11b) products. The radar product with mean field bias adjustment (RAD-G; Fig. 11c), however, indicates that overestimation is prevalent over a larger, statewide domain, requiring reductions in the QPEs. Both RAD and RAD-G products show a linear discontinuity near the bottom of the images. Gourley et al. (2003) shows that the KTLX radar reports high reflectivity values relative to its neighboring radars. This apparent miscalibration of the radar hardware may be the root cause requiring the mean field bias adjustment to lower QPEs in the RAD-G product significantly. The quantitative analysis provided herein will illuminate this potential error source in radar-based QPE.

All statistics indicate that time is simulated best when using RAD rainfall as input for model parameter ensembles, while the ensemble using RAD-G input results in the worst time predictions (Table 6). Different, if not opposite, conclusions are drawn for peak and volume predictions though. The ensemble of RAD-G peak simulations is much lower compared to others as indicated by the ensemble mean bias statistic. The RPS value is also best with the RAD-G product at the 99%

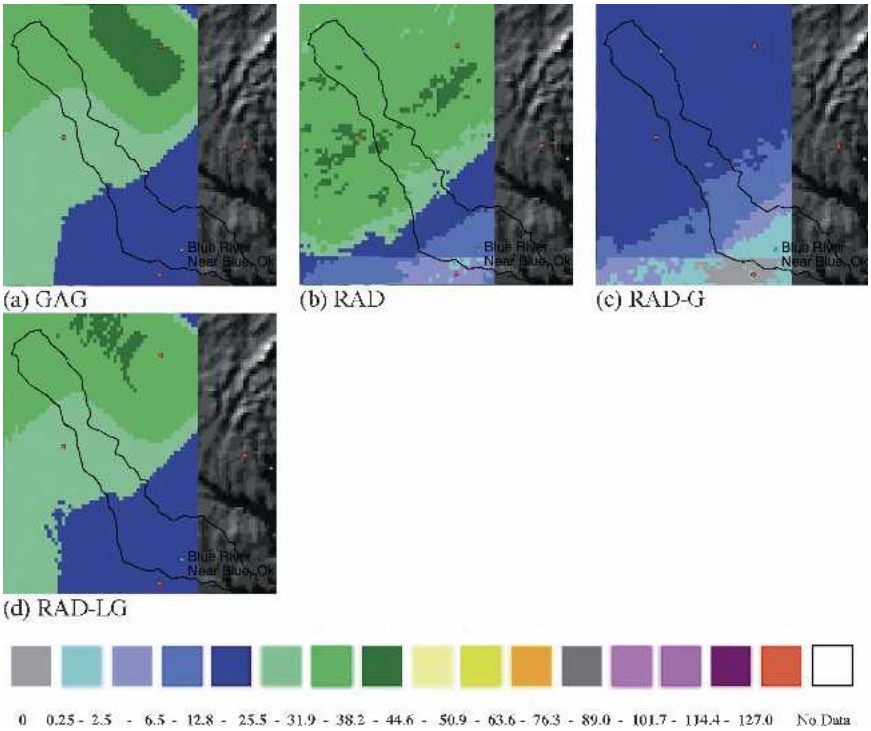

FIG. 11. As in Fig. 3, but for storm total precipitation for the 3 Dec 2002 case.

significance level (Table 7). Similar results are obtained with the volume simulations. Evidently, lower amounts that are estimated by the RAD-G rainfall algorithm are more in tune with an ensemble of simulated hydrographs from the Vflo model.

As noted in previous evaluation cases, a bimodal shape to the pdfs is evident with every ensemble (Figs. 13a–c). Parameter settings under both modes of behavior are examined for simulations of time, peak, and volume. Scalars that are applied to the initial soil satu-

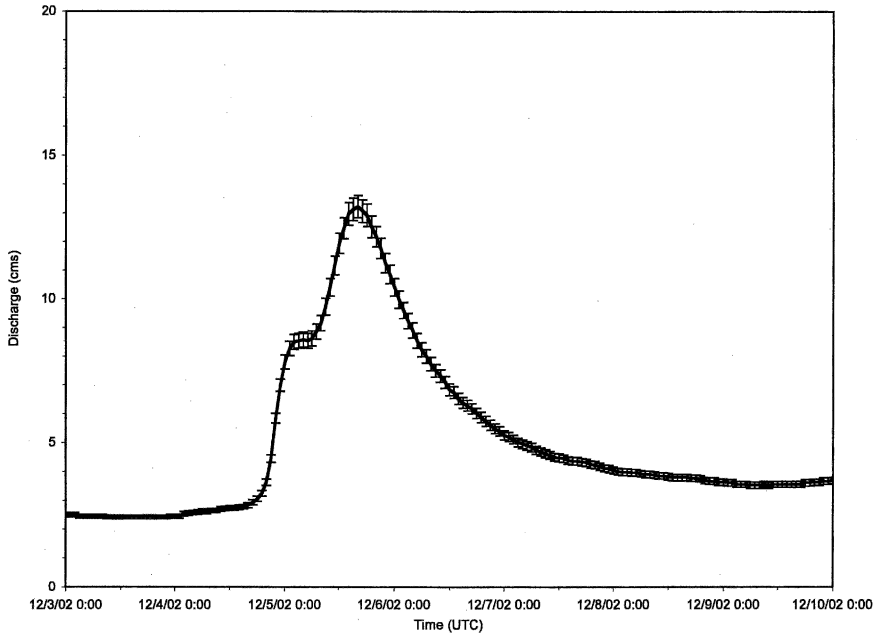

FIG. 12. As in Fig. 4, but for observed streamflow for the 3 Dec 2002 case.

TABLE 6. As in Table 2, but for the 3 Dec 2002 case.

Time	RO coef	Bias	MAE (min)	Rmse (min)	RPS
GAG	0.09	0.94	0.53	0.66	0.94
RAD	0.09	**0.95**	**0.42**	**0.58**	**0.65**
RAD-G	0.19	0.93	0.70	0.75	1.68
RAD-LG	0.09	0.95	0.52	0.66	0.88
Peak	RO coef	Bias	MAE (cms)	Rmse (cms)	RPS
GAG	0.09	4.90	52.22	63.31	5.64
RAD	0.09	4.76	50.44	63.15	5.20
RAD-G	0.19	**1.93**	**14.17**	**23.90**	**0.89**
RAD-LG	0.09	4.87	51.84	62.92	5.54
Volume	RO coef	Bias	MAE (mm)	Rmse (mm)	RPS
GAG	0.09	4.83	9.57	11.36	5.92
RAD	0.09	4.84	9.59	11.27	5.94
RAD-G	0.19	**2.33**	**3.31**	**4.49**	**2.68**
RAD-LG	0.09	4.78	9.44	11.25	5.85

ration parameter (θ) are quite different for the two different modes of time simulations. Smaller relative minima in the simulated time pdfs all correspond to members that use θ values of 100%. Precipitation inputs with different temporal and spatial characteristics interact with complex parameter settings to cause members with 100% soil saturation to deviate nonlinearly from the other members that use unsaturated initial soil conditions. This deviation is not consistent, however, and may result in either later or earlier time simulations. Parameter settings associated with members belonging to the lower density, higher peak, and volume-mode were found to have initial soil saturation settings of 100%. Once again, the bimodal behavior of the pdfs is explained by a nonlinear response in peak and volume simulations to 100% initial soil saturation. As opposed to time simulations, the result of this parameter setting is consistent and, thus, can be expected. All peak and volume simulations are larger, with saturated initial soil conditions. Future work will examine the cause of the sensitivity of hydrologic simulations to this parameter setting.

The uncertainty associated with the parameter ensembles is shown in Figs. 14a–c. All inputs are capable of producing realistic, behavioral simulations regarding

TABLE 7. As in Table 3, but for the 3 Dec 2002 case.

Time	GAG	RAD	RAD-G	RAD-LG
GAG	0.06	0.90	0.99	0.31
RAD		0.05	0.99	0.83
RAD-G			0.04	0.99
RAD-LG				0.05
Peak	GAG	RAD	RAD-G	RAD-LG
GAG	0.08	0.99	0.99	0.74
RAD		0.06	0.99	0.95
RAD-G			0.04	0.99
RAD-LG				0.06
Volume	GAG	RAD	RAD-G	RAD-LG
GAG	0.12	0.63	0.99	0.94
RAD		0.13	0.99	0.96
RAD-G			0.06	0.99
RAD-LG				0.11

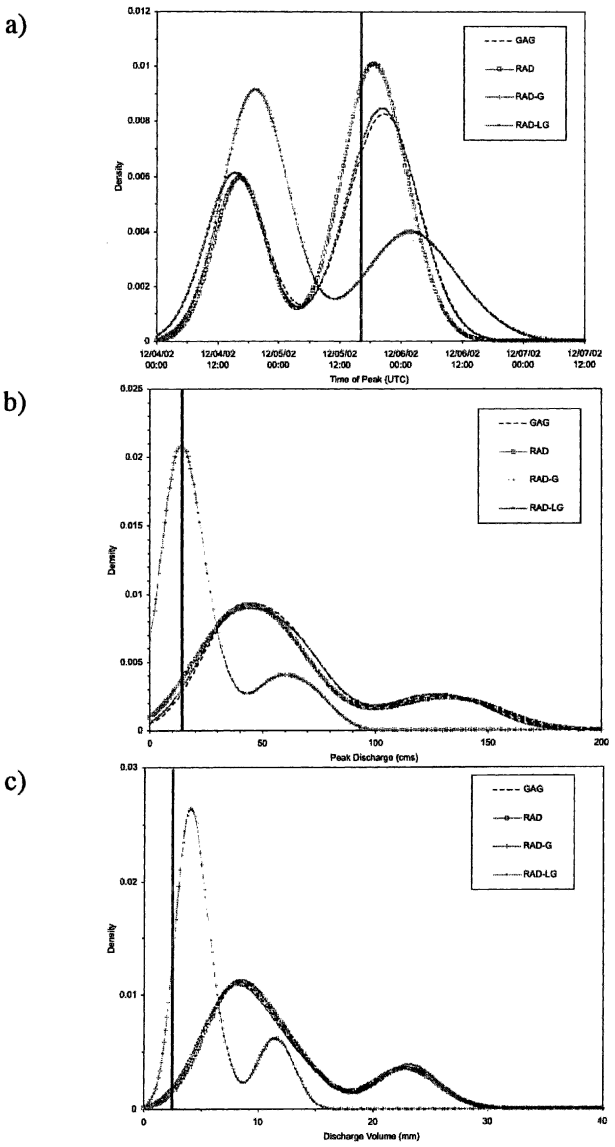

FIG. 13. As in Fig. 5, but for probability density functions for the 3 Dec 2002 case.

the time of maximum discharge. The 90% simulation bounds, using RAD-G inputs, are the only ones that encompass the observed peak discharge value (Fig. 14b) and the observed volume (Fig. 14c). These results point to the RAD-G algorithm as providing the most accurate inputs to the Vflo model when the entire parameter space is considered. Artifacts noted in the RAD algorithm (Fig. 11b) reveal possible calibration errors with the KTLX radar resulting in erroneously high accumulations. In spite of this apparent bias, gauge adjustments that rely only on gauges in close vicinity to the Blue River basin do little to correct the overestimated precipitation amounts. In fact, local gauge adjustments to the radar rainfall field (i.e., RAD-LG) yield products that are nearly indistinguishable from

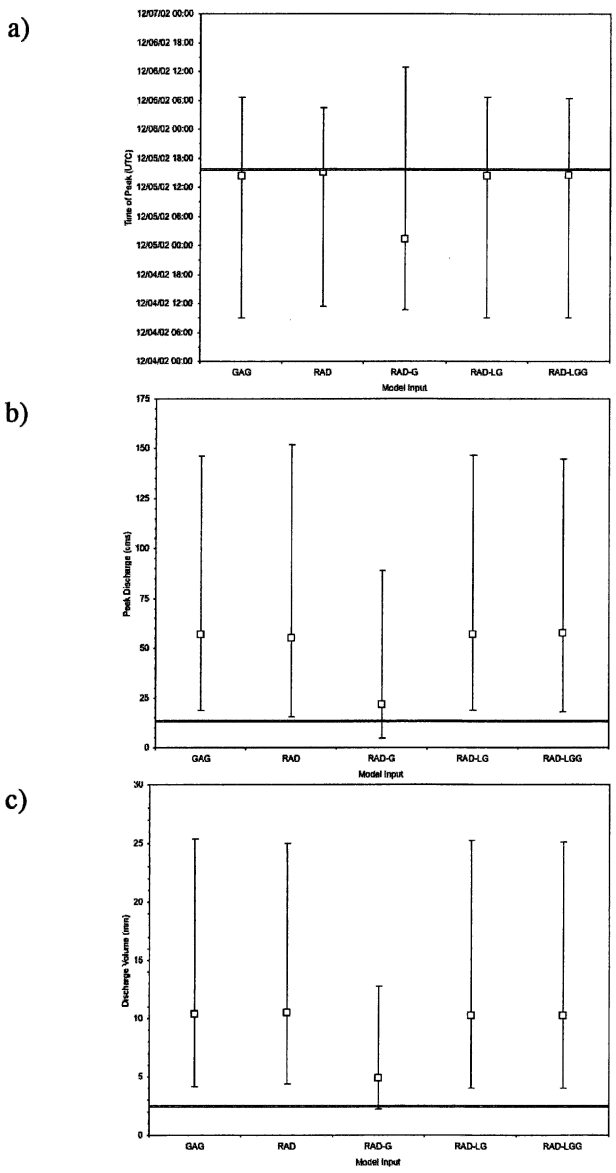

FIG. 14. As in Fig. 6, but for simulation bounds for the 3 Dec 2002 case.

the original one (i.e., RAD). Evidently, there is good radar–gauge agreement at the handful of grid points surrounding the Blue River basin, but these comparisons are not representative of the gridded rainfall over the basin. When over 100 radar–gauge comparisons are made over the entire state, they indicate that the radar-based accumulations are too high. The RAD-G product is, thus, biased low compared to all of the other estimators. In this case, it turns out to be the most accurate input for hydrologic simulations of peak and volume variables. It is noted in all cases how the time simulations are least accurate using the ensemble that produces the most accurate peak and volume simulations.

5. Summary and conclusions

A unique approach is undertaken that places judgment on rainfall estimates not by their agreement with rain gauge measurements, but rather by the skill when the QPEs are input to a given hydrologic model and compared to observed streamflow. Justification for improvements to QPEs is often posed in the context of a need for improved hydrologic forecasting. A methodology is devised to quantify the skill of different QPE inputs at the relevant basin scale by including the range of hydrologic possibilities resulting from parametric uncertainty. Conclusions from this study may be specific to the small sample of events, characteristics of the QPE inputs and the Vflo hydrologic model, or specific hydrologic characteristics of the Blue River basin. Nonetheless, it is the unique methodology of objectively evaluating QPE algorithms from the hydrologic modeling perspective that is the focus of this study. Consistencies in the results from case to case enable the following conclusions to be drawn about the hydrologic evaluation of QPE SUMS rainfall estimates for the three rainfall–runoff events that are examined:

- Rain gauge measurements around the Blue River basin, by themselves, do not provide an accurate depiction of the spatial variability of the rainfall field that is needed for accurate hydrologic simulation. Rain gauges comprising a network over a much larger region are shown to successfully adjust or calibrate remotely sensed QPEs.
- Mean field bias adjustments are found to result in superior hydrologic simulations as compared to "local" adjustment techniques. Latter techniques place more emphasis on individual rain gauge measurements, and spatial details in the original rainfall field are smoothed. These details need to be maintained for skillful hydrologic simulation.
- The inclusion of satellite data in the multisensor algorithm improves QPEs over standard, radar-based products. Satellite data may play an important role in QPE where ground-based radar cannot obtain a representative, low-level sample, which is the case near the outlet of the Blue River basin.

Application of the developed methodologies can be used to evaluate QPE algorithms under development a basin of interest using a hydrologic model of choice. Moreover, the uncertainty analysis techniques can be utilized for other environmental modeling systems. Areas inviting future research involve the objective assessment of the relative contributions (i.e., input versus parametric) to the total uncertainty.

Acknowledgments. Funding for this study was made available by the Department of Education's Graduate Assistance in Areas of National Need Program as well as funding under NOAA-OU Cooperative Agreement NA17RJ1227. Their support is gratefully acknowl-

edged. Oklahoma Mesonet data are provided courtesy of the Oklahoma Mesonet—a cooperative venture between Oklahoma State University and The University of Oklahoma that is supported by the taxpayers of Oklahoma.

APPENDIX A

Verification Measures

a. Statistics based on the ensemble mean

The definitions of *bias*, *mean absolute error* (*MAE*), and *root-mean-square* error (*rmse*) are provided below;

$$\text{bias} = \frac{E[F_i]}{O}, \tag{A1}$$

$$\text{MAE} = E[|F_i - O|], \tag{A2}$$

$$\text{rmse} = \sqrt{E[(F_i - O)^2]}, \tag{A3}$$

where F_i represents a forecast from the ith simulation for time, peak, and volume; O is the observation; and $E[\,]$ is the expected, or mean, value. Bias shows how ensemble means compare to observations in an overall sense. MAE and rmse reveal the degree of scatter or variability between individual forecasts and observations. Biases closest to 1, and MAEs and rmses that are closest to 0, by definition, have the best agreement with observations of streamflow.

b. Gaussian kernel density estimation

Following Silverman (1986), a kernel density estimate is computed as follows:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{A4}$$

where $\hat{f}(x)$ is an estimate of the data density, h is the smoothing parameter or bandwidth, K is the kernel, X_i is the ith prediction, and n is the number of ensemble members (with each supplying a prediction). The kernel is actually a function itself, which satisfies the following condition:

$$\int_{-\infty}^{\infty} K(x)\,dx = 1. \tag{A5}$$

In this case, the kernel is nonnegative and from (A5) satisfies the requirements for a pdf. In (A4), note how the pdf is computed from the sum of individual kernels at the prediction locations. In this sense, the estimated pdf is a compilation of several "bumps," where the shape of the bump is defined by the kernel estimator being used. In this study, a Gaussian kernel is applied. Thus, the kernel density estimate becomes

$$\hat{f}(x) = \frac{1}{\sigma\sqrt{2\pi}}\sum_{i=1}^{n} e^{-\{1/2[(X_i - x)/\sigma]^2\}}, \tag{A6}$$

where σ is the bandwidth. The choice of the bandwidth defines how smooth the estimated pdf will be. Smaller (larger) bandwidths lead to more (less) bumps in the estimated pdf that may or may not be real. In essence, small (large) bandwidths place lower (higher) weight on observations that are further away. If only one pdf were being estimated, then the bandwidth could be chosen subjectively to minimize the insignificant bumps, while retaining the "true" shape of the pdf. In this study, many pdfs are estimated, requiring an objective choice of the smoothing parameter σ. Silverman (1986) recommends using a bandwidth that is adaptive to the range or spread of predictions. The following equation is used to adaptively adjust the bandwidth to accommodate each dataset:

$$\sigma = 0.79\text{Rn}^{-1/5}, \tag{A7}$$

where R is the interquartile range of the predictions.

c. Ranked probability score (RPS)

The formal definition of RPS is provided below:

$$\text{RPS} = \sum_{m=1}^{J}\left[\left(\sum_{i=1}^{m} y_i\right) - \left(\sum_{i=1}^{m} o_i\right)\right]^2, \tag{A8}$$

where y_i is the cumulative probability assigned to the ith category, o_i is the cumulative probability of the observation in the ith category, and J is the number of categories. The RPS values require the selection of categories. Ten categories were chosen based on observed timep, and volume observations. The first rime category was chosen to be the observed time of maximum discharge minus 1 day, with a class interval of 5 h. The first peak and volume categories were chosen to be 25% of the observed variables with class intervals equal to the first category. The choice of categories is a subjective one, but is based entirely on observed data, not on the simulations. This enables the RPS scores to be compared to one another in an objective way.

d. Resampling technique

The resampling technique pools together all of the ensemble members from two different ensembles. Two new ensembles of the same size are created by randomly selecting members from the pooled population. RPS values are computed for the new ensembles and are stored. This procedure is repeated 1000 times. Each RPS *difference* is computed and used to produce a cumulative distribution. The cumulative distribution is then used to determine the probability of obtaining the original *RPS* difference. These confidence levels are computed for each ensemble that is being compared. For the 23 October 2002 event, there are as many as seven different model parameter ensembles representing seven different model inputs (see Table 1). The confidence intervals are, thus, presented in a 7×7 matrix where repeated results are omitted.

APPENDIX B

Vflo Model Formulation

The 1D conservation of mass (B1) and momentum (B2) equations, commonly referred to as the Saint–Venant equations, are used to derive the governing equations in the Vflo model;

$$\frac{\partial h}{\partial t} + \frac{\partial (uh)}{\partial x} = r - i, \tag{B1}$$

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} + g\left(\frac{\partial y}{\partial x} - S_o + S_f\right) = \frac{u(r-i)}{h}, \tag{B2}$$

where u is the 1D component of velocity, h is the flow depth, r is the rainfall rate, i is the soil infiltration rate, g is the acceleration due to gravity, S_o is the bed slope, and S_f is the friction slope. The momentum equation (B2) is simplified by making the assumptions utilized in the kinematic wave analogy. The local acceleration ($\partial u/\partial t$), horizontal momentum advection ($u\partial u/\partial x$), hydrostatic pressure ($\partial y/\partial x$), and forcing terms on the right-hand side of [(B2); $u(r - i)/h$] are all assumed to be at least an order of magnitude smaller than the friction and bed slope terms. Thus, the momentum equation (B2) reduces to

$$S_o = S_f. \tag{B3}$$

Neglecting the aforementioned terms to form the kinematic wave equations assumes that the slope of the energy grade line and water surface elevation are parallel to the land surface slope. It is also assumed that the flow is uniform within the element being considered. In nature, water flow over flat terrain may be subcritical. In this event, changes in water levels may result in waves propagating upstream. These backwater effects can only be accounted for in the local acceleration, advection of horizontal momentum, and hydrostatic pressure terms. All of these terms are neglected in the kinematic wave model; thus, solutions of flow depth in areas where backwater effects are present may not be accurate using this model. Next, an appropriate relationship must be utilized that relates the flow velocity to the flow depth such as Manning's equation in SI units:

$$u = \frac{R^{2/3} S_f^{1/2}}{n}, \tag{B4}$$

where R is the hydraulic radius [defined in (B5) below] and n is the Manning roughness coefficient,

$$R = \frac{hw}{2h + w}, \tag{B5}$$

where h is the flow depth and w is the elemental flow width. The hydraulic radius in a rectangular channel can be approximated by the flow depth when the flow width is assumed to be much greater than the flow depth. Combining this approximation with the result from (B3) reduces Manning's equation to

$$u = \frac{h^{2/3} S_o^{1/2}}{n}. \tag{B6}$$

The simplified form of the momentum equation above can be substituted into the continuity equation (B1) and rearranged to yield the following governing equation, used for overland flow in the Vflo model:

$$\frac{\partial h}{\partial t} + \frac{S_o^{1/2}}{n}\frac{\partial h^{5/3}}{\partial x} = r - i. \tag{B7}$$

The continuity equation may also be expressed in terms of the cross-sectional area A instead of the flow depth h. This leads to the formulation of the conservation of mass for channelized flow:

$$\frac{\partial A}{\partial t} + \frac{\partial Q}{\partial x} = q, \tag{B8}$$

where Q is the channel flow rate and q is the rate of lateral flow entering the channel per unit length. Analogous to the treatment of (B1), it can be shown that substitutions and approximations of the momentum equation effectively relate the flow rate Q to a cross-sectional area A using the Manning equation.

The forcing function on the right-hand side of (B7) is the difference between the soil infiltration rate and rainfall rate, or rainfall excess. The Green and Ampt equation is used in the treatment of soil infiltration in the Vflo model. The infiltration rates are equal to rainfall intensities as long as the rainfall intensities are less than the potential infiltration rates. When the rainfall rate equals and exceeds the potential infiltration rate, the soil is saturated and water becomes ponded at the surface. This ponded water is now available for overland or channelized flow to the adjacent, downstream grid cells. The potential infiltration rate is given as follows:

$$i(t) = K\left(\frac{\psi \Delta\theta}{I(t)} + 1\right), \tag{B9}$$

where K is the saturated hydraulic conductivity, θ is the soil moisture content, ψ is the soil suction at wetting front, and $I(t)$ is the cumulative infiltration. The cumulative infiltration prior to soil saturation is calculated from the rainfall rates accumulated over time. When the rainfall intensity exceeds the infiltration rate in (B9), the time to ponding (t_p) and the associated cumulative infiltration at ponding time (I_p) have been reached and are used below. After ponding has occurred, the following equation describes the cumulative infiltration I:

$$I - I_p - \psi\Delta\theta \ln\left(\frac{\psi\Delta\theta + I}{\psi\Delta\theta + I_p}\right) = K(t - t_p); \tag{B10}$$

I must be solved for implicitly using a method such as a Newton iteration. At this point, the cumulative infiltra-

tion may be inserted in (B9) to yield the infiltration rate after ponding has occurred. This completes the formulation of the governing equations [(B7)–(B8)] for overland and channel flow in the Vflo model.

REFERENCES

Austin, P. M., and A. Bemis, 1950: A quantitative study of the bright band in radar precipitation echoes. *J. Meteor.,* **7,** 165–171.

Barnes, S. L., 1964: A technique for maximizing details in numerical weather map analysis. *J. Appl. Meteor.,* **3,** 396–409.

Beven, K., and A. Binley, 1992: The future of distributed models: Model calibration and uncertainty prediction. *Hydrol. Process.,* **6,** 279–298.

Bradley, A. A., and K. W. Potter, 1992: Flood frequency analysis of simulated flows. *Water Resour. Res.,* **28,** 2375–2385.

Bringi, V. N., and V. Chandrasekar, 2001: *Polarimetric Doppler Weather Radar: Principles and Applications.* Cambridge University Press, 636 pp.

Chow, V. T., D. R. Maidment, and L. W. Mays, 1988: *Applied Hydrology.* McGraw-Hill Series in Water Resources and Environmental Engineering, McGraw-Hill, 572 pp.

Doviak, R. J., and D. S. Zrnic, 1993: *Doppler Radar and Weather Observations*. 2d ed. Academic Press, 562 pp.

Droegemeier, K. K., and Coauthors, 2000: Hydrological aspects of weather prediction and flood warnings: Report on the ninth prospectus development team of the U.S. Weather Research Program. *Bull. Amer. Meteor. Soc.,* **81,** 2665–2680.

Entekahbi, D., and Coauthors, 2002: *Report of a Workshop on Predictability and Limits to Prediction in Hydrologic Systems*. Committee on Hydrologic Science, National Research Council, National Academy Press, 118 pp.

Fabry, F., G. L. Austin, and D. Tees, 1992: The accuracy of rainfall estimates by radar as a function of range. *Quart. J. Roy. Meteor. Soc.,* **118,** 435–453.

Faures, J. M., D. C. Goodrich, D. A. Woolhiser, and S. Sorooshian, 1995: Impact of small-scale spatial rainfall variability on runoff simulation. *J. Hydrol.,* **173,** 309–326.

Frank, E., M. Borga, and E. N. Anagnostou, 1999: Hydrological modeling of mountainous basins using radar rainfall data. Preprints, *29th Int. Conf. on Radar Meteorology,* Montreal, QC, Canada, Amer. Meteor. Soc., 717–720.

Freer, J., K. Beven, and B. Ambroise, 1996: Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach. *Water Resour. Res.,* **32,** 2161–2173.

Goodrich, D. C., J.-M. Faures, D. A. Woolhiser, L. J. Lane, and S. Sorooshian, 1995: Measurement and analysis of small-scale convective storm rainfall variability. *J. Hydrol.,* **173,** 283–308.

Gourley, J. J., and C. M. Calvert, 2003: Automated detection of the bright band using WSR-88D radar data. *Wea. Forecasting,* **18,** 585–599.

——, J. Zhang, R. A. Maddox, C. M. Calvert, and K. W. Howard, 2001: A real-time precipitation monitoring algorithm—Quantitative Precipitation Estimation and Segregation Using Multiple Sensors (QPE SUMS). Preprints, *Symp. on Precipitation Extremes: Prediction, Impacts, and Responses*, Albuquerque, NM, Amer. Meteor. Soc., 57–60

——, R. A. Maddox, K. W. Howard, and D. W. Burgess, 2002: An exploratory multisensor technique for quantitative estimation of stratiform rainfall. *J. Hydrometeor.,* **3,** 166–180.

——, B. Kaney, and R. A. Maddox, 2003: Evaluating the calibrations of radars: A software approach. Preprints, *31st Int. Conf. on Radar Meteorology*, Seattle, WA, Amer. Meteor. Soc., 459–462.

Joss, J., and A. Waldvogel, 1990: Precipitation measurements and hydrology. *Radar in Meteorology*, D. Atlas, Ed., Amer. Meteor. Soc., 577–606.

Kitchen, M., and P. M. Jackson, 1993: Weather radar performance at long range—Simulated and observed. *J. Appl. Meteor.,* **32,** 975–985.

Legates, D. R., and T. L. DeLiberty, 1993: Precipitation measurement biases in the United States. *Water Resour. Bull.,* **29,** 855–861.

Marselek, J., 1981: Calibration of the tipping-bucket raingage. *J. Hydrol.,* **53,** 343–354.

Nystuen, J. A., 1999: Relative performance of automatic rain gauges under different rainfall conditions. *J. Atmos. Oceanic Technol.,* **16,** 1025–1043.

O'Bannon, T., 1997: Using a terrain-based hybrid scan to improve WSR-88D precipitation estimates. Preprints, *28th Conf. on Radar Meteorology*, Austin, TX, Amer. Meteor. Soc., 506–507.

Ogden, F. L., H. O. Sharif, S. U. Senarath, J. A. Smith, M. L. Baeck, and J. R. Richardson, 2000: Hydrologic analysis of the Fort Collins, Colorado flash flood of 1997. *J. Hydrol.,* **228,** 82–100.

Peters, J. C., and D. J. Easton, 1996: Runoff simulation using radar rainfall data. *Water Resour. Bull.,* **32,** 753–760.

Romanowicz, R. J., K. Beven, and J. A. Tawn, 1994: Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach. *Statistics for the Environment 2: Water Related Issues*, V. Barnett and F. Turkman, Eds., Wiley, 297–318.

Sanchez-Diezma, R., D. Sempere-Torres, J. D. Creutin, I. Zawadzki, and G. Delrieu, 2001: Factors affecting the precision of radar measurement of rain. An assessment from an hydrological perspective. Preprints, *30th Int. Conf. on Radar Meteorology,* Munich, Germany, Amer. Meteor. Soc., 573–575.

Sauer, V. B., and R. W. Meyer, 1992, Determination of error in individual discharge measurements. U.S. Geological Survey Open-File Rep. 92-144, 21 pp.

Silverman, B. W., 1986: *Density Estimation for Statistics and Data Analysis.* Chapman and Hall, 175 pp.

Seo, D.-J., and J. P. Breidenbach, 2002: Real-time correction of spatially nonuniform bias in radar rainfall data using rain gauge measurements. *J. Hydrometeor.,* **3,** 93–111.

——, ——, R. Fulton, and D. Miller, 2000: Real-time adjustment of range-dependent biases in WSR-88D rainfall estimates due to nonuniform vertical profile of reflectivity. *J. Hydrometeor.,* **1,** 222–240.

Smith, C. J., 1986: The reduction of errors caused by bright bands in quantitative rainfall measurements made using radar. *J. Atmos. Oceanic Technol.,* **3,** 129–141.

Smith, J. A., D.-J. Seo, M. L. Baeck, and M. D. Hudlow, 1996: An intercomparison study of NEXRAD precipitation estimates. *Water Resour. Res.,* **32,** 2035–2045.

Smith, M. B., D.-J. Seo, V. I. Koren, S. M. Reed, Z. Zhang, Q. Duan, F. Moreda, and S. Cong, 2004: The distributed model intercomparison project (DMIP): Motivation and experiment design. *J. Hydrol.,* **298,** 4–26.

Straka, J. M., D. S. Zrnic, and A. V. Ryzhkov, 2000: Bulk hydrometeor classification and quantification using polarimetric radar data: Synthesis of relations. *J. Appl. Meteor.,* **39,** 1341–1372.

Vieux, B. E., and J. E. Vieux, 2002: Vflo™: A real-time distributed hydrologic model. User guide, Vieux and Associates, Inc., 12 pp.

——, and F. G. Moreda, 2003: Ordered physics-based parameter adjustment of a distributed model. *Advances in Calibration of Watershed Models, Geophys. Monogr.,* Vol. 6, Amer. Geophys. Union, 267–281.

——, Z. Cui, and A. Gaur, 2004: Evaluation of a physics-based distributed hydrologic model for flood forecasting. *J. Hydrol.,* **298,** 155–177.

Warner, T. T., E. A. Brandes, J. Sun, D. N. Yates, and C. K. Mueller, 2000a: Prediction of a flash flood in complex terrain.

Part I: A comparison of rainfall estimates from radar, and very short range rainfall simulations from a dynamic model and an automated algorithmic system. *J. Appl. Meteor.,* **39,** 797–814.

——, D. N. Yates, and G. H. Leavesley, 2000b: A community hydrometeorology laboratory for fostering collaborative research by the atmospheric and hydrologic sciences. *Bull. Amer. Meteor. Soc.,* **81,** 1499–1506.

Westrick, K. J., and C. F. Mass, 2001: An evaluation of a high-resolution hydrometeorological modeling system for prediction of a cool-season flood event in a coastal mountainous watershed. *J. Hydrometeor.,* **2,** 161–180.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Wilson, J., and E. Brandes, 1979: Radar measurement of rainfall–A summary. *Bull. Amer. Meteor. Soc.,* **60,** 1048–1058.

Woodley, W. L., A. R. Olson, A. Herndon, and V. Wiggert, 1975: Comparison of gage and radar methods of convective rain measurement. *J. Appl. Meteor.,* **14,** 909–928.

Yates, D. N., T. T. Warner, and G. H. Leavesley, 2000: Prediction of a flash flood in complex terrain. Part II: A comparison of flood discharge simulations using rainfall input from radar, a dynamic model and an automated algorithmic system. *J. Appl. Meteor.,* **39,** 815–825.

Young, B., A. A. Bradley, W. F. Krajewski, and A. Kruger, 2000: An evaluation study of NEXRAD multisensor precipitation estimates for operational hydrologic forecasting. *J. Hydrometeor.,* **1,** 241–254.

Zawadzki, I. I., 1975: On radar-raingauge comparison. *J. Appl. Meteor.,* **14,** 1430–1436.

Zrnic, D. S., and A. V. Ryzhkov, 1999: Polarimetry for weather surveillance radars. *Bull. Amer. Meteor. Soc.,* **80,** 389–486.