Hydrol. Earth Syst. Sci. Discuss., 8, 3085–3131, 2011
www.hydrol-earth-syst-sci-discuss.net/8/3085/2011/
doi:10.5194/hessd-8-3085-2011

Hydrology and
Earth System
Sciences
Discussions

Evaluating uncertainty estimates in hydrologic models: borrowing measures from the forecast verification community

K. J. Franz[1] **and T. S. Hogue**[2]

[1]Department of Geological and Atmospheric Sciences, Iowa State University,
Ames, IA 50011, USA
[2]Department of Civil and Environmental Engineering, University of California,
Los Angeles, CA 90095, USA

Received: 17 March 2011 – Accepted: 20 March 2011 – Published: 30 March 2011

Correspondence to: K. J. Franz (kfranz@iastate.edu)

Published by Copernicus Publications on behalf of the European Geosciences Union.

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

Abstract

The hydrologic community is generally moving towards the use of probabilistic estimates of streamflow, primarily through the implementation of Ensemble Streamflow Prediction (ESP) systems, ensemble data assimilation methods, or multi-modeling platforms. However, evaluation of probabilistic outputs has not necessarily kept pace with ensemble generation. Much of the modeling community is still performing model evaluation using standard deterministic measures, such as error, correlation, or bias, typically applied to the ensemble mean or median. Probabilistic forecast verification methods have been well developed, particularly in the atmospheric sciences yet, few have been adopted for evaluating uncertainty estimates in hydrologic model simulations. In the current paper, we overview existing probabilistic forecast verification methods and apply the methods to evaluate and compare model ensembles produced from different parameter uncertainty estimation methods. The Generalized Uncertainty Likelihood Estimator (GLUE), a modified version of GLUE, and the Shuffle Complex Evolution Metropolis (SCEM) are used to generate model ensembles for the National Weather Service SACramento Soil Moisture Accounting (SAC-SMA) model for 12 forecast basins located in the Southeastern United States. We evaluate the model ensembles using relevant metrics in the following categories: distribution, correlation, accuracy, conditional statistics, and categorical statistics. We show that the probabilistic metrics are easily adapted to model simulation ensembles and provide a robust analysis of parameter uncertainty, one that is commensurate with the dimension of the ensembles themselves. Application of these methods requires no information in addition to what is already available as part of traditional model validation methodology and considers the entire ensemble or uncertainty range in the approach.

1 Introduction

Streamflow predictions are vitally important to water supply and natural hazard management. However, hydrologic predictions are inherently based on uncertain modeling systems, stemming from ill-defined parameters, model structure, input (forcing) data and validation data (Muleta and Nicklow, 2005; Beven, 2006; Gupta et al., 2006; Clark and Kavetski, 2010; Kavetski and Clark, 2010; Schoups et al., 2010). Recent trends in hydrologic modeling are to produce probabilistic estimates of streamflow, including through the use of Extended Streamflow Prediction (ESP) systems (Day, 1985; Faber and Stedinger, 2001; Franz et al., 2003; Bradley et al., 2004; Franz et al., 2008; Thirel et al., 2008), ensemble data assimilation methods (Kitanidis and Bras, 1980a,b; Evensen, 1994; Margulis et al., 2002; Seo et al., 2003, 2009) and multi-modeling platforms (Ajami et al., 2007; Duan et al., 2007; Vrugt and Robinson, 2007; Franz et al., 2010). Although probabilistic information is produced using these methods, much of the evaluation on the model ensembles (uncertainty) has traditionally been performed using standard deterministic measures, such as error, correlation, or bias, typically applied to the ensemble mean or median. While creating a deterministic variable simplifies the corresponding model evaluation, deterministic evaluation measures are deficient for fully analyzing probabilistic forecast or model performance (Franz et al., 2003; Bradley et al., 2004; Demargne et al., 2010).

In the classic definition, forecast verification is the process of assessing the skill of a forecast or set of forecasts, and work in this area is found as early as 1884 (Murphy and Winkler, 1987; Jolliffe and Stephenson, 2003; Wilks, 2006). Verification methods have been well developed in the atmospheric sciences (Jolliffe and Stephenson, 2003; Wilks, 2006), and their application to hydrologic forecasts has been progressing in recent years, particularly for probabilistic verification (Franz et al., 2003; Bradley et al., 2004; Verbunt et al., 2006; Bartholmes et al., 2009; Renner et al., 2009; Brown et al., 2010; Demargne et al., 2010; Randrianasolo et al., 2010). All methods of verification involve the comparison of a forecast (or set of forecasts) to the corresponding observation (Wilks, 2006), which is defined by Murphy (1993) as forecast quality.

Model validation is not dissimilar from forecast verification, except that the approach is generally aimed at evaluating the reproduction of historical events rather than the prediction of future events. Despite the existence of probabilistic verification measures, few have been adopted for validating historical hydrologic model ensembles. Advances in uncertainty estimation in hydrologic models has not necessarily been followed by advances in assessment of that uncertainty, and the practice of using deterministic evaluation measures to evaluate the median or mean of an ensemble has persisted. However, existing probabilistic forecast verification methods can easily be implemented to assess and compare ensembles produced from different uncertainty estimation methods. There are several examples of probabilistic assessment of ensembles to evaluate model performance in the literature. Duan et al. (2007) used the ranked probability score to evaluate the outcome of a multi-modeling system. De Lannoy et al. (2006) evaluated model uncertainty for soil moisture using the rank histogram (or Talagrand diagram) and several moments from the probability density functions (such as ensemble spread). Franz et al. (2008) applied probabilistic verification methods to hindcasts produced using two different snow models to assess the impact of the model structure on streamflow predictions. Finally, Shrestha et al. (2009) used the range of the probability interval and number of observations that fell within the interval to assess estimates of model parameter uncertainty in a lumped conceptual model.

The focus of the current study is to provide a succinct overview of a range of available probabilistic verification measures and to demonstrate their application in evaluating the quality of parameter uncertainty estimates in hydrologic model simulations. Subsequently, we discuss the ability of the various measures to provide insight on the performance of model output ensembles. The use of probabilistic verification metrics provide a more comprehensive assessment of simulation uncertainty and robustness compared to the traditional approach of evaluating the ensemble mean or median. To demonstrate applicability of the verification methods, we evaluate uncertainty primarily associated with model parameters, although the approach outlined here is readily transferable to evaluation of uncertainty from all potential sources of error. We use the Generalized Uncertainty Likelihood Estimator (GLUE; Beven and Binley, 1992) and the Shuffled Complex Evolution Metropolis (SCEM; Vrugt et al., 2003) to generate model ensembles for an operational forecast model over a select set of basin in the southeastern United States. We introduce a simple (multi-criteria) modification of the GLUE method resulting in three distinct ensemble sets to which the proposed verification measures can be applied.

2 Methods

2.1 Study sites

We undertake our verification assessment for 12 National Weather Service (NWS) forecast basins located in the Southeastern United States (Table 1). All basins fall within the Southeastern Plains ecoregion delineated by the Environmental Protection Agency (EPA) based on similar hydro-climatic characteristics, geomorphology, vegetation, and soil properties. The watersheds within this region have an array of vegetation types including cropland, pasture, woodland and forest. The streambeds in the southeastern plains have a low-gradient and sandy bottoms. The basins also generally have no precipitation as snow. Data for each basin were collected from the Model Parameter Estimation eXperiment (MOPEX) database and spanned a period of 1 January 1948 to 30 September 2002. This region experiences a moderate climate with average temperature of 17.3 °C and average precipitation of 1360 mm yr^{-1}. The study watersheds range in size from less than 1000 km^2 to almost 10 000 km^2 (Table 1).

2.2 Modeling framework

The SACramento Soil Moisture Accounting (SAC-SMA) model (Burnash et al., 1973) is widely used by the NWS River Forecast Centers (RFCs) for forecasting streamflow in the United States. The SAC-SMA is a conceptual model with a two-layer soil system to continuously account for water storage and flow through the subsurface. The upper layer represents surface soil regimes and interception storage, while the lower layer represents deeper soil layers and groundwater storage (Brazil and Hudlow, 1981). Each layer consists of fast components (free water), driven mostly by gravitational forces, and slow components (tension water), driven by evapotranspiration and diffusion. The SAC-SMA is a saturation excess model; when precipitation amounts exceed percolation and interflow capacities, upper zone storage will overflow and overland flow will occur. Direct runoff also occurs from any impervious areas. There are 16 parameters in the SACSMA, of which 13 were calibrated (Table 2). Inputs to the model are basin-average precipitation and potential evapotranspiration. The model output is channel inflow, which is routed to the basin outlet using a series of five linear reservoirs. The linear reservoir recession coefficient, K, was also optimized along with the 13 SACSMA parameters (Table 2). The SAC-SMA model was run at the daily time-step for each of the study basins. Calibration was conducted using the ten year period 1 October 1979 to 30 September 1989. Model verification was conducted for the period of 1 October 1989–30 September 2002 (a shorter time period was used for the Choctawhatchee and Bogue Chitto Rivers based on the available record; Table 1).

2.3 Parameter identification methods

The Generalized Likelihood Uncertainty Estimator (GLUE) methodology is based on the concept that there is no one optimal parameter set but many parameters sets which give reasonable results, termed equifinality (Zak and Beven, 1999; Beven and Freer, 2001). In the GLUE methodology, feasible parameter ranges must be specified from which many parameter sets will be sampled. The model is run with each parameter set and the output is evaluated against the observed variable of interest using a likelihood function to distinguish behavioral sets (accepted) and non-behavioral sets (rejected). The acceptability of the parameter set is based on a selected likelihood function meeting some threshold criteria which is subjectively pre-defined. The cumulative distribution of the likelihood function values is computed for the acceptable parameter sets. To remove outliers, those sets with a likelihood function that falls within the middle 90% of the distribution are chosen.

In the current study, we apply GLUE using the traditional approach and then utilize a slightly modified version, resulting in two GLUE-based parameter uncertainty ensemble methods. In each case, 10 000 parameter sets are generated using Latin hypercube sampling (from a uniform distribution). The SAC-SMA model is run for each of the 10 000 sets, and the model output is evaluated against observed discharge using one or more objective functions. Four possible objective functions are used in the evaluation step, which are the Nash Sutcliffe efficiency score (NSE), root mean squared error (RMSE), percent bias (Pbias), and the correlation coefficient (R):

$$\mathrm{NSE} = 1 - \left(\sum_{t=1}^{N} (x_t - o_t)^2 \Big/ \sum_{t=1}^{N} (x_t - o_t)^2 \sum_{t=1}^{N} (o_t - \bar{o}_t)^2 \right), \tag{1}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{t=1}^{N} (xt - ot)^2}, \tag{2}$$

$$\mathrm{Pbias} = \left[\sum_{t=1}^{N} (x_t - o_t) \Big/ \sum_{t=1}^{N} (x_t - o_t) \sum_{t=1}^{N} o_t \right] \cdot 100, \tag{3}$$

and

$$R = \frac{n \sum_{t=1}^{N} x_t\, o_t - \left(\sum_{t=1}^{N} x_t \right) \cdot \left(\sum_{t=1}^{N} o_t \right)}{\sqrt{n \sum_{t=1}^{N} x_t^2 - \left(\sum_{t=1}^{N} x_t \right)^2} \cdot \sqrt{n \sum_{t=1}^{N} o_t^2 - \left(\sum_{t=1}^{N} o_t \right)^2}}, \tag{4}$$

where x_t is the simulated discharge, and o_t is the observed discharge at time t, and N is the number of time steps.

2.3.1 Traditional GLUE approach

The first parameter identification method applies the standard GLUE methodology in which parameter sets are classified as behavioral or non-behavioral based on a predefined threshold using a single objective function. In this first test, any parameter sets that produce a simulation with NSE > 0.30 is classified as behavioral. From the behavioral sets, those that fall within the 90% prediction bounds are applied. This is a relatively non-restrictive threshold and the approach can result in a large number of behavioral sets.

2.3.2 Modified GLUE approach (W-GLUE)

In the second parameter identification method (noted as W-GLUE), more restrictive criteria are applied to try to reduce the potential range of uncertainty while still maintaining a reasonable degree of accuracy in the ensembles of simulated discharge. A secondary goal is to reduce the number of parameters sets that are used in the uncertainty assessment. In the W-GLUE approach, a multi-criteria approach is utilized, where behavioral parameter sets are defined as those that produce a simulation with DRMS < 2 mm, NSE > 0.6, Pbias $< 10\%$ and $R > 0.8$. Each behavioral parameter set is then assigned a weight given by:

$$\mathrm{WTD} = (1 - \mathrm{NSE}) \cdot 0.5 + (1 - R) \cdot 0.25 + \mathrm{DRMS} \cdot 0.15 + |\%\mathrm{BIAS}| \cdot 0.10 \tag{5}$$

where a perfect WTD function would have a value of 0. Although the assigned weights for each criterion are subjective, the combination of objective functions and weights used here is based on experience with the SAC-SMA model and NWS forecast evaluations (Hogue et al., 2000, 2006; Franz et al., 2003). Similar to the standard GLUE method, the cumulative distribution of the WTD values is created, and the parameters sets with WTD values that fall within the 90% predictions bounds are chosen. Again, the goal of the study is to evaluate forecast measures on a range of ensembles from different parameter uncertainty methods, and not to develop the “best” parameter estimation method or combination of objective measures. Modification of the traditional GLUE approach provides a distinct set of ensembles from the traditional approach for use in the evaluation framework.

2.3.3 SCEM

The third parameter identification method uses the SCEM (Vrugt et al., 2003, 2006), which evolved from a combination of previously developed algorithms, including the Shuffled Complex Evolution (SCE-UA; Duan et al., 1992, 1993) and the Metropolis-Hastings (Metropolis et al., 1953; Hastings, 1970). The SCEM-UA uses an initial (random) population of parameters, for which the posterior density is computed using a Bayesian inference scheme (Box and Tiao, 1973). The population is then portioned into complexes, and a parallel sequence from each complex is initiated from the point (parameter set) that contains the highest posterior density. New candidate points are generated for each sequence and a Metropolis-annealing criterion is used to evaluate whether the new point should be added to the current sequence (Vrugt et al., 2006). If successful, new points will randomly replace existing members of the complex. After a prescribed number of iterations, new complexes are formed through shuffling. Evolution and shuffling are repeated until a targeted stationarity is reached in the Gelman-Rubin convergence diagnostic (Gelman and Rubin, 1992).

2.4 Verification methods

There are an extensive set of forecast verification measures that could be adopted for model evaluation. We chose those that had the most relevance for the modeling framework in the current study and have been identified by the hydrologic forecast community as useful measures. The Cooperative Program for Operational Meteorology, Education and Training (COMET®) Meteorology Education and Training (MetEd) web-based module “Introduction to Verification of Hydrologic Forecasts” (for more information see http://www.meted.ucar.edu) and the NWS Hydrologic Verification System

Requirements Team report (NWS, 2006) describe seven forecast verification categories and list several deterministic and probabilistic metrics for each category. Our ensemble evaluation methodology is developed using six of the seven categories from these two sources (skill scores are not used) and a sample of metrics from each category (Table 3). Metrics in the first category are used to assess the distribution properties of the ensembles. Metrics in categories two through five are used to evaluate the joint distribution of the simulations and observations. Finally, category six contains metrics for confidence.

The deterministic metrics are applied to the ensemble median, which is a typical approach for hydrologic ensemble evaluation. The probabilistic metrics are applied to the simulations produced from the behavioral parameters that fell within the 90% uncertainty bounds, which we refer to as the simulation ensemble. In all cases, the ensembles are treated as a set of discrete variables by using the individual ensemble values and applying an empirical distribution. The distribution metrics are applied to both the parameter ensembles and the simulation ensembles. The remaining metrics are applied to only the simulation ensembles.

2.4.1 Distribution properties

Distribution metrics do not measure performance of an ensemble, but provide measures for understanding the characteristics of a data set. Analyzing the distribution of a data set is a first step in understanding the underlying process that generated the numbers (Wilks, 2006). There are many measures of distribution, including the ensemble mean and median, but we are most interested in those that quantify the ensemble spread. Spread indicates the degree of dispersion around a central value. Spread is useful for understanding the influence of the parameter ensemble uncertainty on the simulation ensemble uncertainty, and the relationship between these ensemble uncertainties and the accuracy measures. Three metrics are used to evaluate the ensemble spread: the interquartile range (IQR), median absolute deviation (MAD), and range:

$$\mathrm{IQR} = \frac{1}{N}\sum_{t=1}^{N} q_{0.75}(t) - q_{0.25}(t) \tag{6}$$

$$\mathrm{MAD} = \frac{1}{N}\sum_{t=1}^{N} \mathrm{median} \sum_{i=1}^{n} x_i(t) - q_{0.50}(t) \tag{7}$$

and

$$\mathrm{Range} = \frac{1}{N}\sum_{t=1}^{N} q_0(t) - q_1(t) \tag{8}$$

where $\{x_1, x_2, x_3, \ldots x_n\}$ is the set of simulated discharge values for one timestep (t) from an ensemble of size n and evaluated for all N timesteps, and q_p is the sample quantile that exceeds the portion of the data given by subscript p, where $0 \le p \le 1$ (Wilks, 2006). When Eqs. (6)–(8) are applied to the parameter ensembles, the measures are computed for individual parameters where $\{x_1, x_2, x_3, \ldots x_n\}$ is the set of n values for a single model parameter that are normalized by the possible parameter range (Table 2). Results from all N (14) model parameters are then averaged to obtain a summary of the distribution characteristics.

Standard deviation is a common measure of sample dispersion, but it can be strongly influenced by values far away from the mean. Therefore, we tested two alternative measures: IQR and MAD. IQR is a measure of spread for the central part of the ensemble only and is therefore resistant to outliers; however, the IQR only considers a small range of the ensemble (Wilks, 2006). MAD is an alternative measure that incorporates all the values of the ensemble. MAD is comparable to the standard deviation but less influenced by outliers because it utilizes the median rather than the mean and does not square the difference. Range is the difference between the highest and lowest values in the ensemble giving a measure of the total uncertainty in the ensemble and will be particularly useful for assessing the tradeoff between precision and accuracy in the simulations. A simulation ensemble can be considered accurate if it contains all the observations within the uncertainty bounds; however if the uncertainty bounds are so large that there is little precision in the ensemble, the ensemble is useless for any meaningful decision-making application. Because we computed IQR, MAD, and Range at each simulation timestep, then averaged the values over time, the distribution metrics are not applicable to the ensemble medians in the context of the present study. However, a distribution analysis of the ensemble medians may be relevant under other situations.

2.4.2 Correlation

The joint distribution of the observations and simulations is commonly evaluated through correlation measures or graphically. In the deterministic approach, scatter plots and Eq. (4) are used to assess the correlation between the ensemble median and the observation. In the probabilistic approach, the correlation between ensemble quantiles and the observations are evaluated using quantile plots. Quantile plots are similar to the scatter plots, except select quantiles (q_i) are plotted against the observed, where i can be chosen as 0.10 (10th quantile), 0.25 (25th quantile), etc.

2.4.3 Accuracy

The term accuracy refers to a measure of error in the simulation ensemble when compared to the observation. Equations (1), (2), and (3) are common error measures in hydrology and are used here to assess the accuracy of the ensemble medians (Table 3). A simple measure of ensemble accuracy is the Containing Ratio (CR) (Xiong et al., 2008). CR is the ratio of the number of the observations that fall within the predictions bounds at any timestep, t, to the total number of observations or time steps, N:

$$\mathrm{CR} = \frac{\sum_{t=1}^{N} X(o(t))}{N} \tag{9}$$

$$X(o(t)) = \left\{ \frac{1,\ q_0(t) < o(t) < q_1(t)}{0,\ \text{otherwise}} \right. \tag{10}$$

where $X(o(t))$ equals 1 when the observation falls within the lower ($q_{0.0}$) and upper ($q_{1.0}$) bounds of the simulation ensemble and $X(o(t))$ equals 0 when the observation falls outsize the ensemble bounds.

The CR provides a useful summary measure of the accuracy of the uncertainty bounds, but it does not consider the distribution of the ensemble members. At a minimum, containing the observation within the uncertainty bounds is desired; but an ensemble in which most members fall near the observation (with only a few outliers) is more useful than an ensemble in which the members are equally distributed across many possible flow regimes. In the case of the former, the ensemble probability would give a more accurate indication of the magnitude of the observation that is most likely. In the case of the latter, the ensemble probability would give similar likelihood to many magnitudes of observations and it would not be clear which flow is most likely. The conditional statistics in the next section are used to evaluate the distribution of the simulation ensemble probability relative to possible observations.

2.4.4 Conditional statistics

Murphy and Winkler (1987) set up a general framework for forecast verification based on factorization of the joint distribution of forecasts and observations into the calibration-refinement factorization:

$$p\left(f_i, y_j\right) = p\left(y_j \,|f_i\right) p\left(f_i\right); \quad i = 1, \ldots, I; \quad j = 1, \ldots, J. \tag{11}$$

and the likelihood-base rate factorization:

$$p\left(f_i, y_j\right) = p\left(f_i \,\middle|y_j\right) p\left(y_j\right); \quad i = 1, \ldots, I; \quad j = 1, \ldots, J. \tag{12}$$

Here, f_i denotes the likelihood of a simulated streamflow event at a given time step. The streamflow event can take on any of the I values, i_1, i_2, ... i_I. Likewise, the corresponding observation, y_j, can take on any of the J values j_1, j_2, ... j_J, where $y_j = 1$ if the observation occurred and $y_j = 0$ if the observation did not occur. These two factorizations provide a basis for evaluating the probability distribution of the simulation ensembles.

The conditional distribution $p(y_j|f_i)$ in Eq. (11) is referred to as reliability and is the more familiar measure of the two. Reliability indicates how often the various J outcomes of y_j occur given the simulated likelihood f_i. Ideally:

$$p\left(y = 1|f_i\right) = f_i \tag{13}$$

(Murphy and Winkler, 1987, 1992; Wilks, 2006). That is, the ensembles are considered perfectly reliable if the conditional probability of the observation equals the probability given by the ensemble for that observation.

Discrimination, or $p(f_i|y_j)$ (Murphy and Winkler, 1987; Wilks, 2006), is a less intuitive measure, but very useful for evaluating how well the ensemble represents the likelihood of the observation relative to other possible observations. If $(f_1|y_1) = (f_2|y_1)$, the ensemble was not very discriminatory for event y_1. On the other hand, if $(f_1|y_1) = 1$ and $(f_2|y_1) = 0$, the ensemble was perfectly discriminatory for event y_1.

Reliability and discrimination are displayed graphically and interpretation of the diagrams is explained in the results section. In this application of the metrics, three possible event categories ($I = J = 3$) were used: low flows or <30% of climatology; middle flows or 30%–70% of climatology; and high flows or >70% of climatology, where climatology is based on the available discharge data at each site (Table 1). The simulation ensemble likelihood values were expressed as the probability of non-exceedance and at intervals of 10% (deciles) based on the empirical distribution of the ensemble members.

Terms $p(f_i)$ and $p(y_j)$ are the marginal distributions (or frequencies) of the ensemble probability and observations, respectively. The marginal distribution of the ensembles can be displayed on a frequency diagram to indicate the sharpness, or resolution, of the ensembles. That is, are the ensembles dispersed across many flow ranges and therefore little probability is given to any particular flow event, or are the forecasts highly refined where likelihoods close to 0% and 100% are frequently observed (Wilks, 2006). As forecasts become sharper, the forecast probability becomes more narrowly distributed and is more frequently assigned to the extreme likelihood categories (i.e., 0–10% and >90–100%). Thus, the sample sizes within the middle probability categories are small for sharp forecasts.

2.4.5 Categorical statistics

The categorical statistics listed in Table 3 are used to evaluate dichotomous events. We use these metrics to evaluate the ability of the ensembles to simulate floods. The magnitude of the flood discharge at the outlet gage for each watershed was obtained from the Lower Mississippi River Forecast Center website (http://www.srh.noaa.gov/lmrfc/).

The contingency table is a common method for verifying the joint distribution of non-probabilistic forecasts and observations. This concept is applied here to assess the ability of the ensemble median to identify flood (o_1) and no-flood (o_2) events. Likewise, the ensemble median is classified as flood (x_1) and no-flood (x_2). We set up a 2 × 2 contingency table (Fig. 1), and count all possible observation/simulation pairs. Two measures are used to summarize the 2 × 2 contingency table (Wilks, 2006): the conditional probability $p(x_1|o_1)$, also known as the probability of detection (POD):

$$p\left(x_1|o_1\right) = \mathrm{POD} = \frac{a}{a+c} \tag{14}$$

and $p(x_1|o_2)$, also known as the probability of false detection (POFD):

$$p\left(x_1|o_2\right) = \mathrm{POFD} = \frac{b}{b+d}. \tag{15}$$

The Brier Score (BS) (Brier, 1950), is used to evaluate the accuracy of the simulation ensemble for floods and no-floods events in a probabilistic manner. The BS is the mean squared error of the likelihood given to a particular event (f_t) and the corresponding observation (y_t):

$$\mathrm{BS} = \frac{1}{N}\sum_{t=1}^{N}(f_t - y_t)^2 \tag{16}$$

where N is the number of forecast/observation pairs and is equal to the total number of timesteps, t. The value of f_t is the likelihood given to the observation of interest (e.g., flood). The value of the observation (y_t) is equal to 1 if the observation occurred or 0 if the observation did not occur (Wilks, 2006). A perfect BS is 0, and the score ranges from $0 \le \mathrm{BS} \le 1$.

2.4.6 Sample size

Finally, confidence for evaluation purposes refers to uncertainty in the value of the metric and is applied as appropriate (Table 3). The marginal distribution of the observations, $p(y_i)$, in Eq. (12) is used to assess sample size in the various flow categories and reflects the degree of confidence in the metric (Table 3). With larger samples sizes, the results are more likely to be representative of the behavior of the modeling system. Confidence intervals are demonstrated where graphically possible, but given the large number of samples, the confidence intervals are exceedingly small in most cases.

In the current study, all statistics are computed for daily simulations, for a total of 4745 timesteps (with the exception of 5475 and 4015 for Choctawhatchee and Bogue Chitto Rivers, respectively), therefore the overall sample size is quite large. However, for evaluating extremes (i.e. floods), the sample size and confidence become an important consideration.

3 Results

3.1 Parameter uncertainty estimation

The SCEM produced the largest parameter ensembles (Table 4), with an average of 14 310 sets per site. The GLUE produced an average of 5771 parameter sets per site, and the W-GLUE produced considerably smaller ensembles with an average of 876 parameter sets per site. The W-GLUE method produced only 18 parameter sets for the Rappahannock River basin; this was the smallest set of parameters produced for all methods (Table 4).

For each site and method, the parameter ensemble means were computed and normalized by the feasible parameter range given in Table 2. The distribution of the normalized means indicates that the GLUE parameter ensemble means have the least variation among sites (Fig. 2). Most of the normalized GLUE parameter values are near 0.5, indicating that the parameter ensemble means are located near the middle of the feasible parameter range. In comparison, the parameter ensemble means from the W-GLUE and the SCEM have more variation between sites and are not located near the middle of the parameter range as frequently. Values of parameter *k* (the routing parameter) have the most between-site variability for all methods.

3.2 Distribution properties

The parameter ensemble range reveals that both the GLUE and the W-GLUE parameter ensembles have values that span the entire feasible parameter space, resulting in parameter ranges near 1 (Fig. 3a). Comparison of the parameter ensemble size to the parameter ensemble range indicates that the two metrics are not strongly correlated. Although the W-GLUE parameter ensemble sizes are much smaller, the W-GLUE method produced only a slightly smaller parameter range than the GLUE (Fig. 3a). The SCEM parameter ensembles are the largest, but the range of parameter values are much narrower than the GLUE and W-GLUE parameter sets, spanning less than

30% of the feasible parameter space at all sites. The W-GLUE trial for the Rappahannock River, in which 18 parameter sets were identified, is the only instance where W-GLUE produced a parameter ensemble range similar to SCEM. There is a correlation between the range of the parameter ensemble and the range of the discharge ensemble (Fig. 3b); the methods that have the largest parameter ranges also have the largest discharge ensemble ranges. Larger parameter ensemble ranges also correlate with larger parameter ensemble MADs for all methods (Fig. 3c). As parameter ensemble range increases, MAD increases, indicating that the parameter values deviate more from the median.

The IQR, which is another measure of the parameter distribution, shows similar results to the range analyses (Fig. 3d). The IQR is highest for the GLUE parameter sets and the IQR from the W-GLUE method are only slightly smaller than the GLUE. In addition, similar to the range, larger IQR values for the parameter sets corresponds to larger IQR values for the discharge ensembles. The IQR does reveal characteristics about the SCEM ensembles that are not apparent when evaluating the range only. The SCEM parameter and discharge ensembles vary little among sites, whereas the range values had larger variation. This suggests that the central 50% of the parameter sets are very similar, and the variation seen in the range (Fig. 3a) comes from the upper and lower 25% of the distribution.

3.3 Correlation

Evaluating the correlation of the ensemble medians to the observations reveals values greater than 0.6 for all methods, with the W-GLUE showing the highest values overall. Correlations are lowest for the SCEM (Fig. 4d). Correlation is a summary measure of the relationship between the observations and the ensemble median, whereas the quantile plots provide correlation information about the medians for various quantiles of flow. The quantile plots for select sites for the GLUE and SCEM are shown in Fig. 5. These plots reveal that the ensemble medians are poorly calibrated to lower and higher flows at most sites. Several of the sites (Chunky, Pearl, Bogue Chitto, Ochlocknee

Rivers) have results similar to those of the Leaf River (Fig. 5c and f) in that the median underestimates flows at both the lower and upper end of the range, but overestimates the middle range; this is true for all methods. This results in the overall good correlation seen in Fig. 4d, but it is clear that the calibration between the ensemble median and the discharge changes across flow ranges.

The best fit between the ensemble median and the observations occur for the Chickasawahay River (Fig. 5a and d); however, the lowest flows are underestimated by the median for all methods at this site. The Noxumbee River is one of the few sites in which the low flows are over-estimated by the medians (Fig. 5b and e). Except for the Chickasawahay River, the medians from all methods underestimate the highest flows at all sites. However, flows above 10 mm day^{-1} represent the upper 2% of the distribution at the sites studied, and therefore this underestimation occurs for a very small number of samples.

The $q_{0.10}$ and $q_{0.90}$ of the GLUE ensembles capture the low- and middle-range flows fairly well at most sites (Fig. 5b and c). The high-flow ranges are captured by the $q_{0.10}$ and $q_{0.90}$ of the GLUE only for Choctawhatchee, Chickasawahay (Fig. 5a), Noxumbee, Tar, and Ochlocknee Rivers. For those basins in which the observation did not fall within $q_{0.10}$ and $q_{0.90}$, the observations are most always contained within the upper 10% (between $q_{0.90}$ and $q_{1.0}$) as will be indicated by the containing ratio (CR) discussed in the next section.

3.4 Accuracy

Accuracy measures are used to evaluate the relative performance of the ensemble medians from the three parameter estimation methods. There are many ways to display these measures; however, we chose to combine all values from all basins into a single plot using traditional boxplots. The metrics indicate that on average, the SCEM ensemble medians had lower accuracy scores than the GLUE and W-GLUE methods. SCEM ensemble medians have the highest nRMSE and percent bias values (Fig. 4b and c), and the lowest NSE (Fig. 4d). The ensemble medians of the GLUE and the W-GLUE perform similarly with similar nRMSE and NSE values, although the W-GLUE medians have slightly better scores. There is a general tendency for the ensemble median from all methods to overestimate discharge at the study sites (Fig. 4c), with the exception of one site for the W-GLUE method and four sites for the GLUE method. By plotting all sites together on one diagram, it is shown that the percent bias values from the W-GLUE methods are most consistent across sites.

The CR provides a measure of the accuracy of the range of the ensemble. The GLUE ensembles have CR values very close to 1 (Fig. 6a and b), indicating that the observation is captured almost 100% of the time by the ensemble. The W-GLUE ensembles have very similar results as the GLUE, with very high capture rates. High CR values generally correlate with large ensemble ranges, which provide a higher likelihood that the observation will be contained within the bounds of the ensemble. Likewise, the narrow range of the SCEM ensembles (Fig. 5d–f) is correlated with low CR values (less than 0.3 at all sites). Although the ensemble values are close to the observation values on average (Fig. 5d–f), the SCEM ensembles do poorly for the CR metric. The CR was also computed by category by separating ensembles into one of three flow categories based on where the observation occurred (low, middle, high). Results indicate that the capture rates for each method are fairly consistent across all flow ranges but are slightly better for low flows (Fig. 6c).

3.5 Conditional statistics

In the analyses presented thus far, simulated discharge values (i.e. median, minimum and maximum of the ensemble) were compared to observed discharge values and evaluated using various metrics. However, to characterize the uncertainty information provided by the ensembles, it is necessary to evaluate the accuracy of the likelihood information which is derived from the distribution of the ensemble members. A critical skill of a probabilistic simulation is the ability for the ensemble to indicate which event(s) is most likely, rather than just merely capture the event using large uncertainty bounds.

Reliability diagrams (Fig. 7) illustrate the correlation between the ensemble likelihood for a given flow (low, middle, high) to the frequency of the observations in that flow category. Perfectly reliable ensembles will fall along the 1:1 line. If the conditional distribution falls to the right (left) of the 1:1 line, the ensembles are over-estimating (under-estimating) the likelihood of the observations, or are over-confident (under-confident) (Wilks, 2006). The relative frequency of the probability is shown as an inset on the reliability diagrams to assess sample size in each likelihood bin (a measure of ensemble refinement and confidence). The 90% uncertainty bounds on the reliability values are shown in Fig. 7, but are exceedingly small.

The reliability diagram indicates that, overall, the ensembles have low reliability and are poorly refined (Fig. 7 and insets, respectively). The GLUE and W-GLUE ensembles have the best reliability on average, and the ensembles are most reliable for high flows (Fig. 7c and f). At a few of the sites, the GLUE and W-GLUE ensembles have quite good reliability for low flows (Fig. 7a and d), but for many sites the ensembles are over-estimating the frequency of low flows. The middle flows on average, tend to be over-estimated by all methods. Also apparent from the reliability diagrams is that the GLUE ensembles never assigned more than a 70% likelihood to middle flows (Fig. 7b) and W-GLUE ensembles never assigned more than a 80% likelihood to middle flows (Fig. 7e).

Reliability diagrams allow evaluation of skill for individual magnitudes of probability; and we note that the skill varies for the studied parameter estimation methods. For example, although the SCEM ensembles over-estimated the likelihood of each flow category on average, they have fairly good reliability for likelihoods less than 40% (Fig. 7g–i). Ensembles from all methods have very good reliability for likelihoods in the range of 0–10% and 10–20% for all flow categories. This is also the category with the highest sample size (Fig. 7, insets). The skill for likelihood bins above 20% varies by site and method. There are sites where the ensembles have good reliability throughout the range of likelihoods for both the GLUE (Fig. 7a and c) and W-GLUE methods (Fig. 7d and f).

Because the ensemble members are highly distributed, the frequency of ensembles that have likelihood in the 0–10% range is very high which indicates poor refinement (Fig. 7 insets). The refinement of the GLUE and W-GLUE, which have the largest parameter ranges, are slightly worse than SCEM ensembles. Because the SCEM has smaller parameter ranges, it produced more ensembles with probabilities in the 90–100% likelihood bin. Although still poorly refined, the SCEM ensembles do have more instances of ensemble likelihood falling in categories above 70%.

Classic calibration and verification approaches evaluate the simulation on the basis of how well the simulation matches the observation at each timestep. The reliability diagrams indicate that this practice does not assure a good calibration of the ensemble likelihoods for the parameter estimation methods evaluated in this study.

The ability of the ensembles to discriminate when observations will fall within one of three possible flow categories (low, middle and high flows) is displayed in the discrimination diagrams in Fig. 8. Results for all sites have been averaged together due to space considerations; for analysis of an individual site, one figure like Fig. 8 would be needed for each site. If an ensemble is highly and correctly discriminatory for a given flow level, the majority of the ensembles (and therefore probability) will fall within the range for that flow level when it is observed. This means very little probability is given to the other flow levels. If the ensembles are discriminatory, then the probability distribution functions of the flow categories will not overlap to a great degree on the discrimination diagram (Murphy et al., 1989). Ideally, the ensembles should assign 100% likelihood to the flow category that was observed and 0% likelihood to any other flow category.

The discrimination diagrams indicate that for all methods there is good discrimination for low flows (Fig. 8a, d, g) and high flows (Fig. 8c, f, i), and poorer discrimination for middle flows (Fig. 8b, e, h). The ensembles have some trouble determining when high flows and low flows were most likely. Looking at just those times when low flows occur, the ensemble probability is more often higher for middle flows than high flows. Likewise, when high flows occur the ensemble probability is more often higher for middle flows than low flows.

The impact of the large range of GLUE and W-GLUE ensembles and small range of the SCEM ensembles can be seen in the results from the discrimination diagrams. Specifically, the ensemble probability of the SCEM is almost always given in the 0–10% or 90–100% probability bins due to the small ensemble spread. GLUE and W-GLUE on the other hand occasionally gives ensemble probability in the middle categories, indicating that the probability is more often distributed between flow categories, rather than focused on one category. While in general the discrimination is highest for the SCEM, the narrow ensembles do lead to occasional failures in simulating the correct flow category when low (Fig. 8g) and high flows (Fig. 8i) occur as indicated by the presence of ensemble probability for middle flows in the 90–100% probability bin. However, the narrow ensemble range of the SCEM does result in better discrimination for middle flows compared to GLUE and W-GLUE. The higher ensemble spread in the GLUE and W-GLUE lead to ensembles that give likelihood to low and high flows when middle flows occur, resulting in relatively poor discrimination for middle flows (Fig. 8b and e).

3.6 Categorical statistics

In the final analysis, three metrics are evaluated for determining how well the ensembles simulated the occurrence of floods. The ensemble median is evaluated using the probability of detection (POD) and the probability of false detection (POFD) using the contingency table in Fig. 1, and the entire ensemble is evaluated using the Brier score (BS). No flood level was available for the Rappahannock River, and therefore this site was not used in the analysis. At least one flood was observed during the evaluation period at all sites.

The POD and POFD are generally displayed using a relative operating characteristic (ROC) curve (Jolliffe and Stephenson, 2003; Wilks, 2006), however, because the POFD was very low (average 1%), a bar graph is used for better illustration (Fig. 9). POD values equal to one and POFD values close to zero are optimum. Cross comparison of POD and POFD values are useful for determining which methods are performing best on average and if there are particular problems at a given site. For example, all methods perform similarly, but the SCEM and W-GLUE have slightly higher POD than the GLUE (the average PODs are 0.56 for GLUE, 0.60 for W-GLUE and SCEM).

Results from previous metrics help lend insight into the results shown in Fig. 9. Recall from Fig. 5 that the medians from all methods tend to underestimate the highest flows at all sites, this negative bias in the upper range of the discharge values results in the low POFD (Fig. 9b). The low bias in the median for the Leaf (Fig. 5c and f) and Chunky Rivers (not shown in Fig. 5) leads to a low POFD (Fig. 9b), but results in no skill for POD (Fig. 9a) at these sites. Chickasawahay River, which has a well calibrated median, also has among the highest POD.

The narrower ensemble bounds of the SCEM result in higher (less optimal) BS at the study sites compared to the other two parameter estimation methods (Fig. 10a). The average BSs are 0.014 for GLUE, 0.013 for W-GLUE, and 0.039 for SCEM. At Chikasawahay, Noxumbee, and Leaf Rivers, the GLUE (Fig. 5a–c) and W-GLUE ensembles capture the high flows well (with the 10–90% uncertainty bounds). As a result, these ensembles assign some level of probability to floods and result in low BS values in this test (Fig. 10a). Although the smaller range and tendency to overestimate high flows (Fig. 5d–f) by the SCEM ensembles have the potential to lead to better flood predictions by assigning higher likelihoods to flood events, the narrow ensemble range leads to poorer overall performance. The SCEM ensembles at Choctawhatchee River strongly underestimate the highest flows. As a result, the BS for the SCEM are poor at this site, while the GLUE and W-GLUE do better because the larger ensemble ranges allowed some of the high flows to be captured. Comparing these results to the frequency at which floods are simulated by each model (Fig. 10b), it is apparent that the SCEM ensembles give 0% change of flows above flood stage on average 95% of the time. The GLUE and W-GLUE give likelihoods of floods more frequently. As a result, the SCEM does more poorly than the other methods for the BS because it gives less probability when flood events occur.

Note that the BS for predictions above flood stage produces the same BS for predictions below flood stage. For this reason, the failure to detect any floods for Leaf and Chunky Rivers (Fig. 9a), produces a very low BS, because they had zero POFD. Thus the BS is being heavily influenced by all the instances of no flood.

4 Discussion

Of the distribution measures evaluated in this study, the range seems most useful for cross-comparison of ensembles and understanding the relationship to discharge ensembles. MAD is secondarily useful. The IQR gives different information about the distribution than the MAD and range, but appears redundant to the range for our analysis. Using these probabilistic metrics, in general, the SCEM outperformed the GLUE and W-GLUE only with respect to discrimination. The low range of the SCEM is a desirable attribute, and allowed for better discrimination of flows in all flow categories. However, the narrow range of the ensemble led to poorer performances in metrics that evaluated the ability of the ensemble to capture the event within the uncertainty bounds, such as CR and BS. The W-GLUE has a slightly better results compared to the GLUE in many cases. The smaller parameter ensemble has the advantage of being more manageable and it was shown that the decrease in ensemble size does not reduce the accuracy of the ensembles or ensemble medians. However, this method still produces very wide parameter ensembles, and therefore, generally does not provide an improvement in ensemble resolution.

The quantile plots and range were most useful for understanding the ensemble performance for other metrics. Quantile-quantile plots also provided information about how the medians performed for various ranges of flow, which could not be gleaned from summary statistics such as Eqs. (1)–(4). The MAD and IQR provided similar information, and the MAD was more useful in assessing the dispersion of the ensembles. The CR is a simple overall measure of ensemble accuracy and a good baseline assessment of ensemble performance. However, as shown, the CR does not provide

Title Page
Abstract | Introduction
Conclusions | References
Tables | Figures
◀| |▶
◀ ▶
Back | Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

information about biases in the ensembles, which is important for understanding why the ensembles, such as the SCEM, failed to capture many of the observations.

Reliability is a common measure of calibration of the ensemble likelihoods. The parameter estimation methods that we tested did very poorly for reliability for our study sites. One likely explanation is that the ensembles spreads were very large, resulting in very low probabilities for any possible observation, therefore the reliability for ensemble probabilities above 20% was very poor due to low sample sizes. This suggests that many of the individual ensemble members are poorly calibrated to the observations. Model calibration traditionally is accomplished by comparing the simulation and observation at a single time-step. Perhaps calibration to the ensemble likelihood using a summary metric that tests reliability as an objective function, would improve the ensemble probability accuracy and the reliability results.

Discrimination also lends insight to the quality of the ensemble calibration. The ensembles did give high likelihoods to low and high flows when they occurred, indicating that the majority of the ensembles are not falsely simulating these flow categories. Simulation of the proper flow category is heavily influenced by the precipitation inputs, which may explain the similar results among the methods. The SCEM does discriminate better than the GLUE or W-GLUE. The simulation of floods is also heavily influenced by the precipitation inputs. However, as POD results for the Leaf River revealed, negatively biased ensembles can fail to produce flows large enough to indicate a flood regardless of the inputs.

Evaluation with respect to flow categories allows for assessment of model performance under different conditions (i.e. low flows versus high flows) that summary measures are unable to portray. The BS had limitations for evaluating flood simulations. Because this summary value includes instances when no flooding occurred, (that is, it is an evaluation of both flood and no-flood events), it is heavily influenced by the large number of samples of no-flood. This leads to very low scores even though the ensembles simulated the very high flows quite poorly on average.

Our choice of flow levels based on climatological thresholds introduces a somewhat arbitrary cut off point for analysis. While the use of reliability does not require the use of flow categories (reliability could be computed for all ensembles collectively), metrics such as discrimination and BS require some degree of categorization of the observations of interest introducing subjectivity into the analysis. Additionally, we chose to use probability intervals of 10%, this interval can be adjusted to varying situations and needs, and to test different confidence levels.

As mentioned above, the POFD and POD are often displayed graphically using a ROC diagram. ROC diagrams can also be used to evaluate continuous and probabilistic forecasts (Mason and Graham, 1999; Jolliffe and Stephenson, 2003). The area under the ROC curve is a common way to summarize the results of the contingency table analysis. The limitation of the POFD and POD metric applied here, and as a result any ROC diagram that could be developed from the values, is that the numerical value of the predictand, in this case the ensemble median, is only used to sort the medians into the 2 × 2 table (Wilks, 2006). The magnitude of the ensemble median as compared to the observation is not evaluated.

An additional graphical approach for correlation analysis is called the rank histogram. Although we chose not to show the rank histograms in this manuscript because of space requirements, it is frequently used in the forecast verification literature. The rank histogram is a measure of statistical consistency and is used to determine if the ensemble includes the observations as equiprobable members (Hamill and Collucci, 1997; Wilks, 2006). Rank histograms are a useful tool for evaluating the ensemble spread and forecast confidence, and can reveal deficiencies in the ensemble calibration that can be connected to characteristics of the reliability diagrams. For a discussion of rank histogram application and interpretation see Hamill (2001).

Another common measure not applied here is the ranked probability score (RPS). RPS is similar to the BS, except that more than two possible outcomes (discharge categories) can be defined (Epstein, 1969; Wilks, 2006). The measure is sensitive to the distance between the categories forecasted and the true observation, and increasingly penalizes forecast the more probability is given to categories far from the category in which the observation occurred. For analysis of continuous variables such as discharge, the use of the categories introduces a degree of subjectivity to the verification process as the number and type of categories chosen will affect the final value of the RPS. The continuous ranked probability score (CRPS) is the application of the RPS to an infinite number of categories, and is therefore sensitive to the entire range of possible observed values (Hersbach, 2000). Both the BS (Murphy, 1973; Wilks, 2006) and the CRPS (Hershbach, 2000) can be decomposed into reliability, resolutions and uncertainty to give more insight into the forecast system performance that cannot be interpreted from the single BS or CRPS alone.

5 Concluding remarks

When evaluating ensembles or simulations with an associated uncertainty, deterministic metrics are often applied to the median or expected value. This practice ultimately removes a significant amount of ensemble information from the evaluation process. We have demonstrated a number of metrics that are traditionally applied for verification of probabilistic forecasts, and have shown these to be informative for evaluation and comparison of streamflow simulations. A considerable amount of information about the relative utility of the uncertainty estimation methods can be gleaned when treating the simulations in a probabilistic manner. The probabilistic metrics provide an analysis of model uncertainty, one that is commensurate with the dimension of the ensembles themselves.

Advanced probabilistic verification metrics developed for forecast verification provide a rigorous platform by which modeling methods can be evaluated and cross-compared. The application of these methods require no information in addition to what is already available as part of the traditional model validation methodology, except that it considers the entire ensemble or uncertainty range in the approach. Common problems such as identifying thresholds or appropriate distributions still exist, however, theses measure are much more informative about the true nature of model uncertainty estimates than simple deterministic measures. Through our efforts in this and future papers, we hope to advance discussion about evaluation of simulation uncertainty and more robust model verification measures.

Acknowledgements. This research was partially supported by a grant from the National Oceanic Atmospheric Administration (NOAA) National Weather Service (#NA07NWS4620013). We would also like to thank Alexandre Dulac and Sonya Lopez for their contributions to this work.

References

Ajami, N. K., Duan, Q., and Sorooshian, S.: An integrated hydrologic Bayesian multimodel combination framework: Confronting input, parameter, and model structural uncertainty in hydrologic prediction, Water Resour. Res., 43, W01403, doi:10.1029/2005WR004745, 2007.

Bartholmes, J. C., Thielen, J., Ramos, M. H., and Gentilini, S.: The european flood alert system EFAS - Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts, Hydrol. Earth Syst. Sci., 13, 141–153, doi:10.5194/hess-13-141-2009, 2009.

Beven, K.: A manifesto for the equifinality thesis, J. Hydrol., 320, 18–36, 2006.

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

Beven, K. and Freer, J.: Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology, J. Hydrol., 249, 11–29, doi:10.1016/S0022-1694(01)00421-8, 2001.

Box, G. E. P. and Tiao, G. C.: Bayesian Inference in Statistical Analysis, Addison-Wesley, Reading, Mass., 1973.

Bradley, A. A., Schwartz, S. S., and Hashino, T.: Distributions-Oriented Verification of Ensemble Streamflow Predictions, J. Hydrometeorol., 5(3), 532–545, 2004.

Brazil, L. E. and Hudlow, M. D.: Calibration procedures used with the National Weather Service Forecast System, in: Water and Realted Land Resource Systems, edited by: Haimes, Y. Y. and Kindler, J., Pergamon, Tarrytown, NY, 457–466, 1981.

Brier, G. W.: Verification of forecasts expressed in terms of probabilities, Mon. Weather Rev., 78, 1–3, 1950.

Brown, J., Demargne, J., Seo, D.-J., and Liu, Y.: The Ensemble Verification System (EVS): a software tool for verifying ensemble forecasts of hydrometeorological and hydrologic variables at discrete locations, Environ. Modell. Softw., 25, 854–872, doi:10/1016/j.envsoft.2010.01.009, 2010.

Burnash, R. J., Ferral, R. L., and McGuire, R. A.: A Generalized Streamflow Simulation System Conceptual: Modeling for Digital Computers, Joint Federal-State River Forecast Center, Sacramento, CA, 1973.

Clark, M. P. and Kavetski, D.: Ancient numerical daemons of conceptual hydrological modeling: 1. Fidelity and efficiency of time stepping schemes, Water Resour. Res., 46, W10510, doi:10.1029/2009WR008894, 2010.

Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Res. Pl.-ASCE, 111, 157–170, 1985.

De Lannoy, G. J. M., Houser, P. R., Pauwels, V. R. N., and Verhoest, N. E. C.: Assessment of model uncertainty for soil moisture through ensemble verification, J. Geophys. Res., 111, D10101, doi:10.1029/2005JD006367, 2006.

Demargne, J., Brown, J., Liu, Y., Seo, D. J., Wu, L., Toth, Z., and Zhu, Y.: Diagnostic verification of hydrometerological ensembles, Atmos. Sci. Lett., 11(1), 114–122, 2010.

Duan, Q., Sorooshian, S., and Gupta, V.: Effective and efficient global optimization for conceptual rainfall-runoff models, Water Resour. Res., 28, 1015–1031, 1992.

Duan, Q., Gupta, V. K., and Sorooshian, S.: A Shuffled Complex Evolution Approach for Effective and Efficient Global Optimization, J. Optimiz. Theory App., 76(3), 501–521, 1993.

Duan, Q., Ajami, N. K., Gao, X., and Sorooshian, S.: Multi-model ensemble hydrologic prediction using Bayesian model averaging, Adv. Water Resour., 30, 1371–1386, 2007.

Epstein, E. S: A Scoring System for Probability Forecasts of Ranked Categories, J. Appl. Meteorol., 8, 985–987, 1969.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99(C5), 10143–10162, 1994.

Faber, B. A. and Stedinger, J. R.: Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts, J. Hydrol., 249, 113–133, 2001.

Franz, K. J., Hartmann, H. C., Sorooshian, S., and Bales, R.: Verification of National Weather Service Ensemble Streamflow Predictions for Water Supply Forecasting in the Colorado River Basin, J. Hydrometeorol., 4, 1105–1118, 2003.

Franz, K. J., Hogue, T., and Sorooshian, S.: Operational snow modeling: Addressing the challenges of an energy balance model for National Weather Service forecasts, J. Hydrol., 360, 48–66, 2008.

Franz, K. J., Butcher, P., and Ajami, N. K.: Addressing snow model uncertainty for hydrologic prediction, Adv. Water Resour., 33, 820–832, 2010.

Gelman, A. and Rubin, D. B.: Inference from iterative simulation using multiple sequences, Stat. Sci., 7, 457–472, 1992.

Gupta, H. V., Beven, K. J., and Wagener, T.: Model calibration and uncertainty estimation, in: Encyclopedia of Hydrologic Sciences, edited by: Anderson, M. G. and Mcconnell, J. J., John Wiley, New York, 2015–2031, 2006.

Hamill, T. M.: Interpretation of the rank histogram for verifying ensemble forecasts, Mon. Weather Rev., 129, 550–560, 2001.

Hamill, T. M. and Collucci, S. J.: Verification of Eta-RSM short-range ensemble forecasts, Mon. Weather Rev., 125, 1312–1327, 1997.

Hastings, W. K.: Monte-Carlo sampling methods using Markov Chains and their applications, Biometrika, 57, 97–109, 1970.

Hersbach, H.: Decomposition of the continuous ranked probability score for ensemble prediction systems, Weather Forecast., 15, 559–570, 2000.

Jolliffe, I. T. and Stephenson, D. B.: Forecast Verification: A Practioner's Guide in Atmospheric Science, John Wiley and Sons, Chichester, 240 pp., 2003.

Kavetski, D. and Clark, M. P.: Ancient numerical daemons of conceptual hydrological modeling: 2. Impact of time stepping schemes on model analysis and prediction, Water Resour. Res., 46, W10511, doi:10.1029/2009WR008896, 2010.

Kitanidis, P. K. and Bras, R. L.: Real-Time Forecasting With a Conceptual Hydrologic Model 2. Applications and Results, Water Resour. Res., 16(6), 1034–1044, 1980a.

Kitanidis, P. K. and Bras, R. L.: Real-Time Forecasting with a Conceptual Hydrologic Model I. Analysis of Uncertainty, Water Resour. Res., 16(6), 1025–1033, 1980b.

Margulis, S. A., McLaughlin, D., Entekhabi, D., and Dunne, S.: Land data assimilation and estimation of soil moisture using measurements from the Southern Great Plains 1997 Field Experiment, Water Resour. Res., 38(12), 1299, doi:10.1029/2001WR001114, 2002.

Mason, S. J. and Graham, N. E.: Conditional Probabilities, Relative Operating Characteristics, and Relative Operating Levels, Weather Forecast., 14, 713–725, 1999.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E.: Equations of state calculations by fast computing machines, J. Chem. Phys., 21, 1087–1091, 1953.

Muleta, M. K. and Nicklow, J. W.: Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model, J. Hydrol., 306, 127–145, 2005.

Murphy, A. H.: A new vector partition of the probability score, J. Appl. Meteor., 12, 595-600, 1973.

Murphy, A. H.: What is a good forecast? An essay on the nature of goodness in weather forecasting, Weather Forecast., 7, 692–698, 1993.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Mon. Weather Rev., 115, 1330–1338, 1987.

Murphy, A. H. and Winkler, R. L.: Diagnostic verification of probability forecasts, Hydrol. Process., 7, 435–455, 1992.

NWS: National Weather Service River Forecast Verification Plan. Report of the Hydrologic Verification System Requirements Team, October 2006, US Department of Commerce, NOAA/NWS, Silver Spring, Maryland, http://nws.noaa.gov/oh/rfcdev/docs/Final_Verification_Report.pdf (last access: October 2010), 2006.

Randrianasolo, A., Ramos, M. H., Thirel, G., Andréassian, V., and Martin, E.: Comparing the Scores of hydrologic ensemble forecasts issued by two different hydrological models, Atmos. Sci. Lett., 11, 100–107, 2010.

Renner, M., Werner, M. G. F., Rademacher, S., and Sprokkereef, E.: Verification of ensemble flow forecasts for the River Rhine, J. Hydrol., 376, 463–475, doi:10.1016/j.jhydrol.2009.07.059, 2009.

Schoups, G., Vrugt, J. A., Fenicia, F., and van de Giesen, N. C.: Corruption of accuracy and efficiency of Markov chain Monte Carlo simulation by inaccurate numerical implementation of conceptual hydrologic models, Water Resour. Res., 46, W10530, doi:10.1029/2009WR008648, 2010.

Seo, D. J., Koren, V., and Cajina, N.: Real-Time Variational Assimilation of Hydrologic and Hydrometeorological Data into Operational Hydrologic Forecasting, J. Hydrometeorol., 4(3), 627–641, 2003.

Seo, D. J., Cajina, L., Corby, R., and Howieson, T: Automatic state updating for operational streamflow forecasting via variational data assimilation, J. Hydrol., 367(3–4), 255–275, 2009.

Shrestha, D. L., Kayastha, N., and Solomatine, D. P.: A novel approach to parameter uncertainty analysis of hydrological models using neural networks, Hydrol. Earth Syst. Sci., 13, 1235–1248, doi:10.5194/hess-13-1235-2009, 2009.

Thirel, G., Rousset-Regimbeau, F., Martin, E., and Habets, F.: On the Impact of Short-Range Meteorological Forecasts for Ensemble Streamflow Predictions, J. Hydrometeorol., 9, 1301–1317, 2008.

Verbunt, M., Zappa, M., Gurtz, J., and Kaufmann, P.: Verification of a coupled hydrometeorological modeling approach for alpine tributaries in the Rhine basin, J. Hydrol., 324, 224–238, doi:10.1016/j.jhydrol.2005.09.036, 2006.

Vrugt, J. A. and Robinson, B. A.: Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, Water Resour. Res., 43, W01411, doi:10.1029/2005WR004838, 2007.

Vrugt, J. A., Gupta, H. V., Bouten, W., and Sorooshian, S.: A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrological model parameters, Water Resour. Res., 39, 1201, 2003.

Vrugt, J. A., Gupta, H. V., Nuallain, B., and Bouten, W.: Real-Time Data Assimilation for Operational Ensemble Streamflow Forecasting, J. Hydrometeorol., 7(3), 548–565, 2006.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd edition, Academic Press, Amsterdam, 627 pp., 2006.

Xiong, L. and O'Connor, K. M.: An empirical method to improve the prediction limits of the GLUE methodology in rainfall-runoff modeling, J. Hydrol., 349, 115–124, 2008.

Zak, S. and Beven, K.: Equifinality, sensitivity and predictive uncertainty in the estimation of critical loads, Sci. Total Environ., 236, 191–214, 1999.

Table 1. Study basins, basin area, and annual average precipitation and discharge for the period of record 1979–2002. Calibration and verification periods started 1 October and ended 30 September of the years indicated.

Site Name	USGS Gage ID	Area [km^2]	precipitation [mm yr^{-1}]	discharge [mm yr^{-1}]	Calibration period	Verification period
Rappahannock River Near Fredericksburg, VA	01668000	4134	1047.0	358.4	1979–1989	1989–2002
Tar River at Tarboro, NC	02083500	5654	1143.5	327.9	1979–1989	1989–2002
Ochlockonee River Nr Havana, FL	02329000	2953	1335.7	326.2	1979–1989	1989–2002
Flint River at Montezuma, GA	02349500	7511	1193.3	366.1	1979–1989	1989–2002
Choctawhatchee River at Caryville, FL	02365500	9062	1405.0	534.7	1979–1989	1989–1994
Escambia River Near Century, FL	02375500	9886	1470.3	544.1	1979–1989	1989–2002
Noxubee River at Macon, MS	02448000	1989	1388.9	464.9	1979–1989	1989–2002
Leaf River Nr Collins, MS	02472000	1924	1479.4	517.9	1979–1989	1989–2002
Chunky River Nr Chunky, MS	02475500	956	1419.0	467.4	1979–1989	1989–2002
Chickasawhay River at Leakesville, MS	02478500	6967	1459.0	495.3	1979–1989	1989–2002
Pearl River at Edinburg, MS	02482000	2341	1390.2	455.4	1979–1989	1989–2002
Bogue Chitto River near Bush, LA	02492000	3142	1597.8	626.1	1979–1989	1989–2000

Table 2. SACSMA model parameters and feasible range.

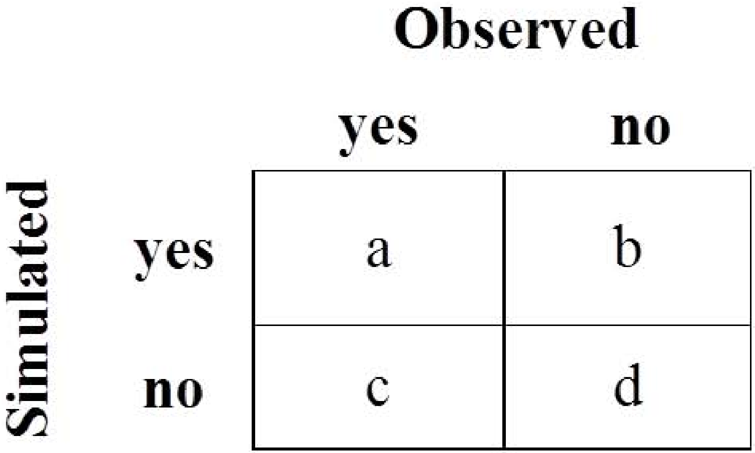
Parameter	Description	Units	Range
UZTWM	Upper-zone tension water maximum storage	mm	1–150
UZFWM	Upper-zone free water maximum storage	mm	1–150
LZTWM	Lower-zone tension water maximum storage	mm	1–500
LZFPM	Lower-zone free water primary maximum storage	mm	1–1000
LZFSM	Lower-zone free water supplementary storage	mm	1–1000
UZK	Upper-zone free water lateral depletion rate	day^{-1}	.1–.7
LZPK	Lower-zone primary free water depletion rate	day^{-1}	0–0.2
LZSK	Lower-zone supplementary free water depletion rate	day^{-1}	0.01–0.5
ADIMP	Additional impervious area	decimal fraction	0–0.4
PCTIM	Impervious fraction of the watershed	decimal fraction	0–0.1
ZPERC	Maximum percolation rate	dimensionless	1–249
REXP	Exponent of the percolation equation	dimensionless	0.5–4.5
PFREE	Fraction of water percolating from upper zone directly to lower-zone free water storage	decimal fraction	0–0.8
K	Five-level linear reservoir constant	dimensionless	0.0–0.9
RIVA	Riparian vegetation	decimal fraction	0
SIDE	Ratio of deep recharge to channel base flow	decimal fraction	0.3
RSERV	Fraction of lower-zone free water not transferable to lower-zone tension water	decimal fraction	0

Table 3. Statistical measures used for evaluation of parameter estimation methods and their respective categories.

Categories	Deterministic metrics	Probabilistic metrics
Distribution Properties		median, mean, range, inter-quartile range (IQR), median absolute deviation (MAD), CDF
Correlation	scatter plots, correlation coefficient	quantile plots, rank histogram
Accuracy (error)	Nash-Sutcliffe efficiency (NSE), percent bias (%Bias), root mean square error (RMSE)	containing ratio (CR)
Conditional Statistics		reliability , discrimination, resolution
Categorical Statistics	probability of detection, probability of non-detection	Brier score (BS)
Confidence	sample size	sample size, confidence interval

Table 4. Number of parameters generated for each watershed by each parameter estimation method.

	Number of parameters		
Site Name	GLUE	W-GLUE	SCEM
Rappahannock River Near Fredericksburg, VA	3213	18	11 758
Tar River at Tarboro, NC	6502	978	17 949
Ochlockonee River Nr Havana, FL	6111	655	1942
Flint River at Montezuma, GA	4307	409	3252
Choctawhatchee River at Caryville, FL	4424	699	16 101
Escambia River Near Century, FL	6823	1481	17 340
Noxubee River at Macon, MS	6881	1811	23 698
Leaf River Nr Collins, MS	6447	1209	15 103
Chunky River Nr Chunky, MS	4535	180	19 302
Chickasawhay River at Leakesville, MS	10 000	925	17 787
Pearl River at Edinburg, MS	6192	1702	15 793
Bogue Chitto River near Bush, LA	4324	447	11 693



		Observed	
		yes	no
Simulated	yes	a	b
	no	c	d

Fig. 1. Contingency table displaying the relationships between counts **(a–d)** of event pairs.

HESSD

8, 3085–3131, 2011

Evaluating uncertainty estimates in hydrologic models

K. J. Franz and T. S. Hogue

Title Page
Abstract Introduction
Conclusions References
Tables Figures
◀▶
◀ ▶
Back Close
Full Screen / Esc
Printer-friendly Version
Interactive Discussion

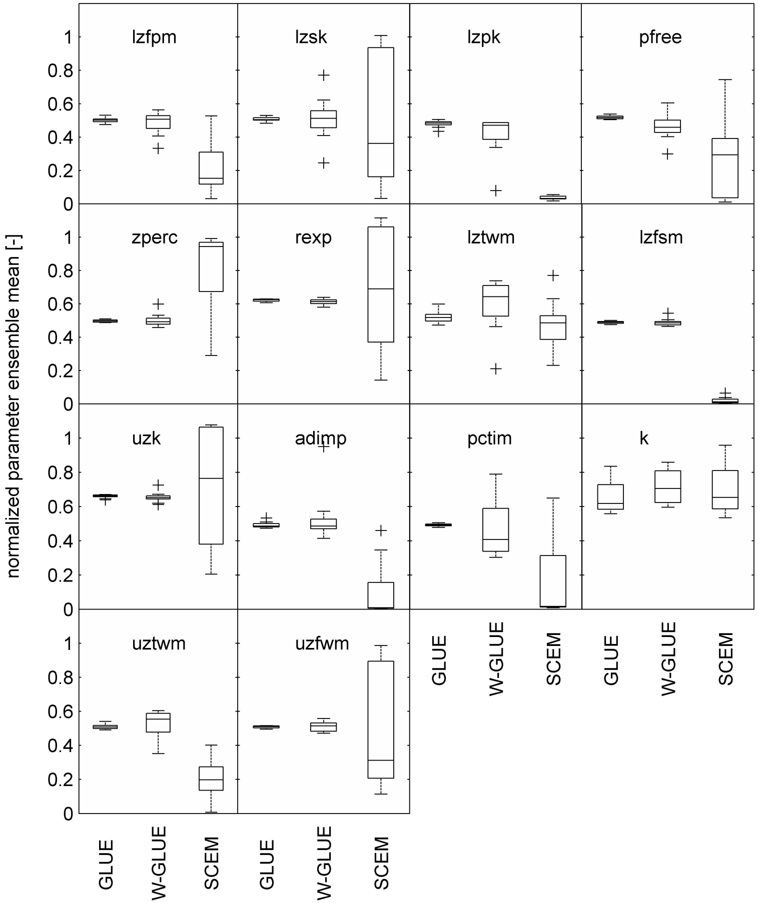


Fig. 2. Comparison of the parameter ensemble means across study basins for each parameter estimation method. The mean parameter values were normalized by the respective feasible parameter range and averaged for each site.

HESSD

8, 3085–3131, 2011

Evaluating uncertainty estimates in hydrologic models

K. J. Franz and T. S. Hogue

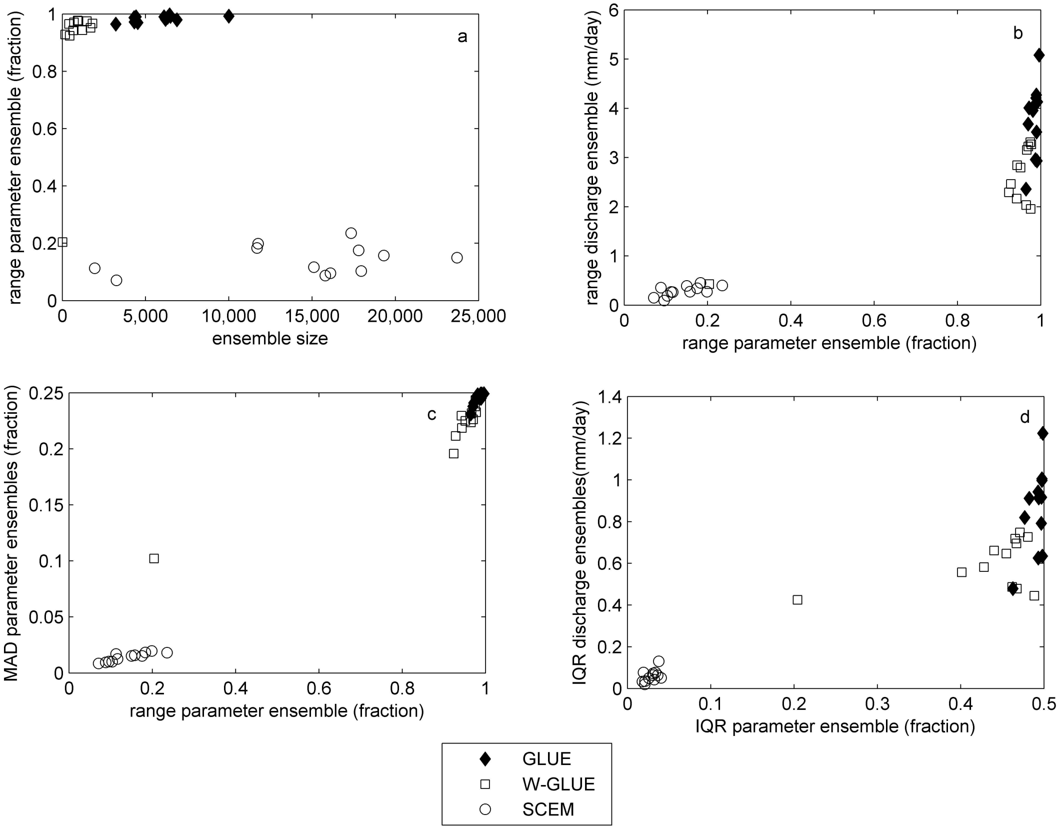


Fig. 3. Comparison of **(a)** parameter ensemble range to parameter ensemble size, **(b)** discharge ensemble range to parameter ensemble range, **(c)** mean absolute deviation (MAD) of the parameter ensembles to parameter ensemble range, and **(d)** interquartile range (IQR) of the discharge ensembles to IQR of the parameter ensembles for each study site. Parameters were normalized by their feasible range before computing the average parameter ranges, MADs and IQRs.

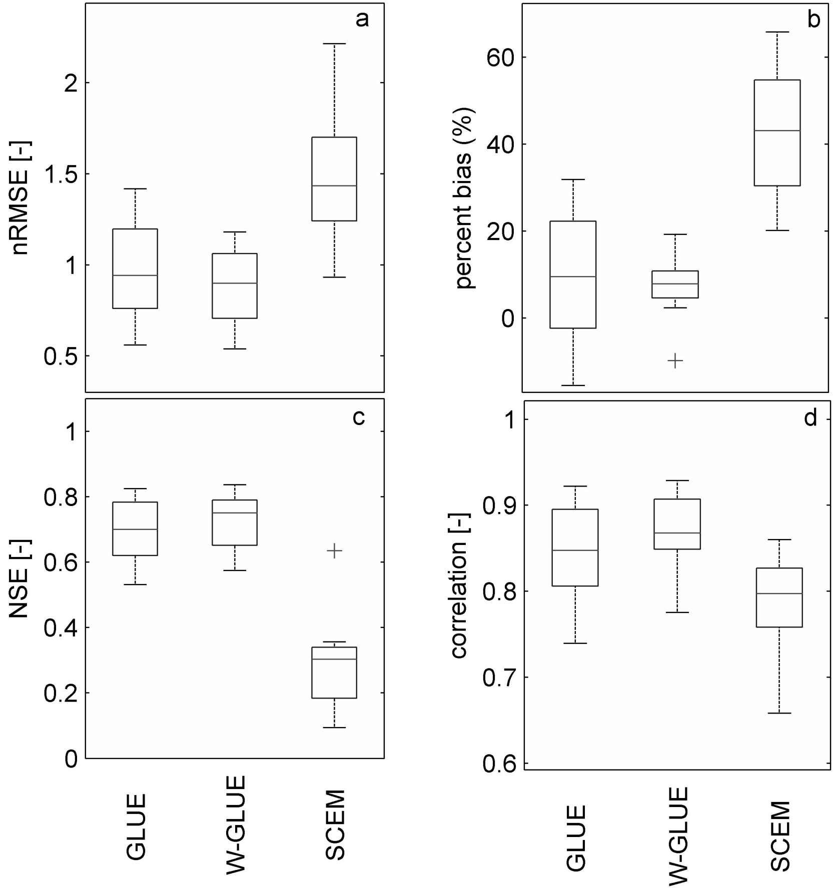


Fig. 4. Box plots of the deterministic measures **(a)** root mean square error normalized by the basin mean discharge (nRMSE), **(b)** percent bias, **(c)** Nash Sutcliffe efficiency (NSE), and **(d)** correlation for the discharge ensemble median for all sites for the verification period.

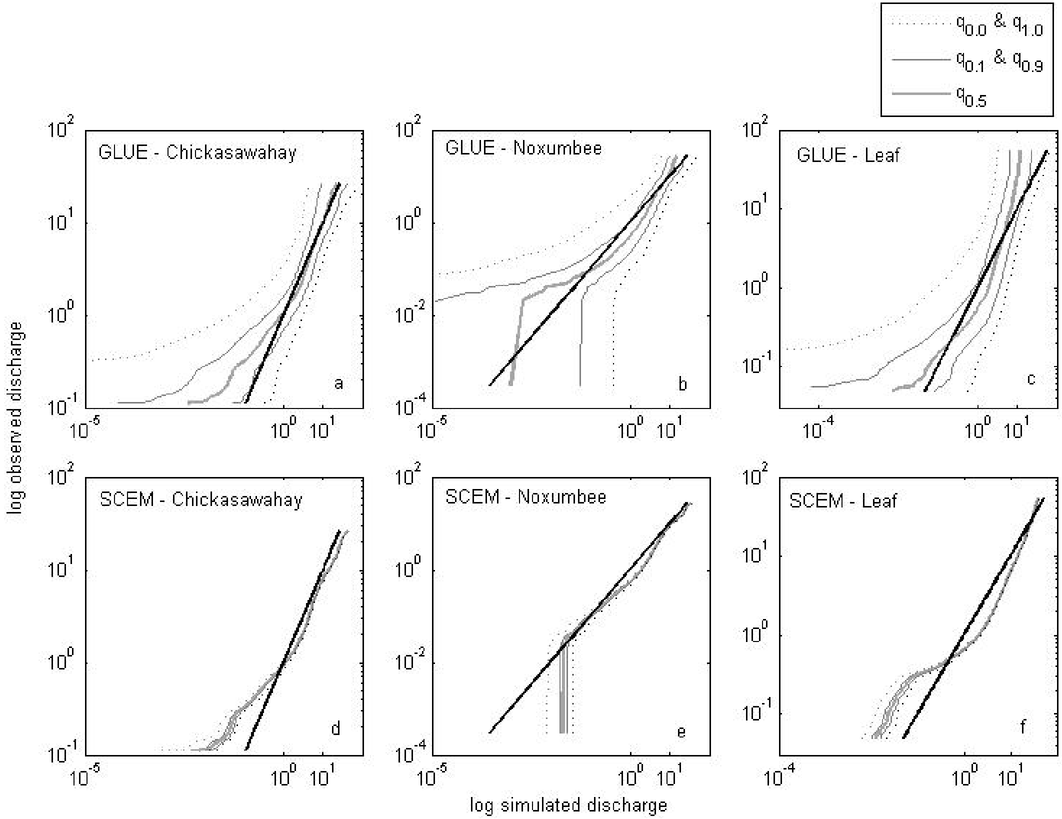


Fig. 5. Quantile plots comparing the 0th, 10th, 50th, 90th, and 100th quantiles of the discharge ensembles from the **(a–c)** GLUE and **(d–f)** SCEM methods to the observations for select sites. The data is displayed on a log scale.

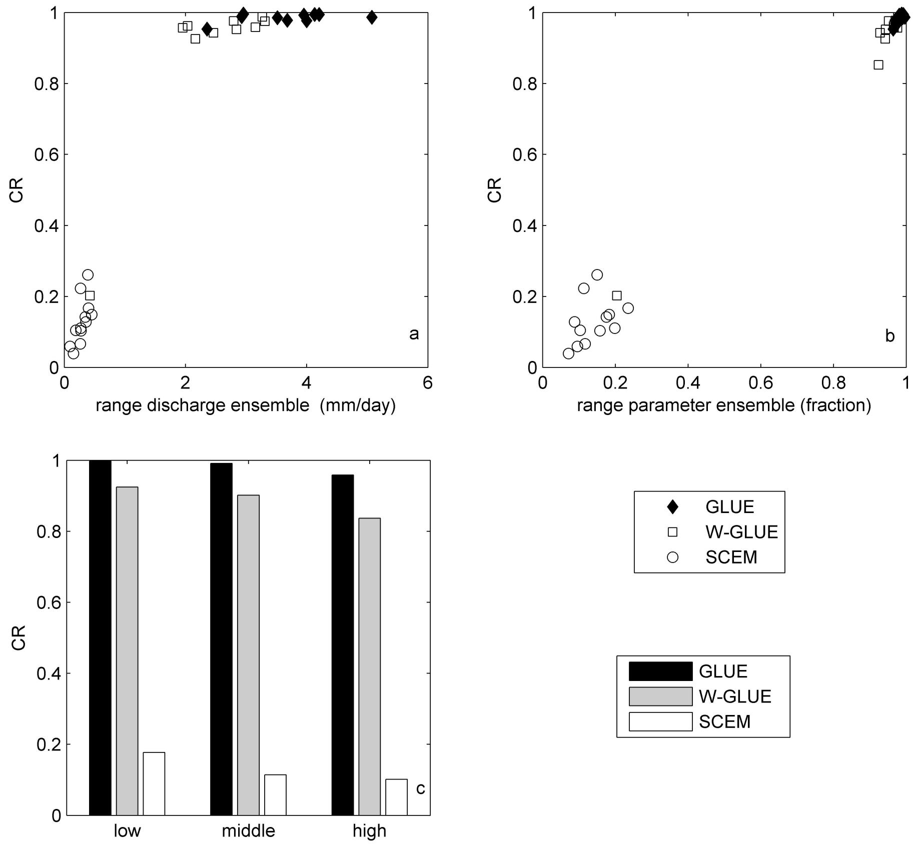


Fig. 6. Comparison of the containing ratios (CR) to the **(a)** discharge ensemble ranges and **(b)** parameter ensemble ranges; and **(c)** the average CR from the study sites by flow category.

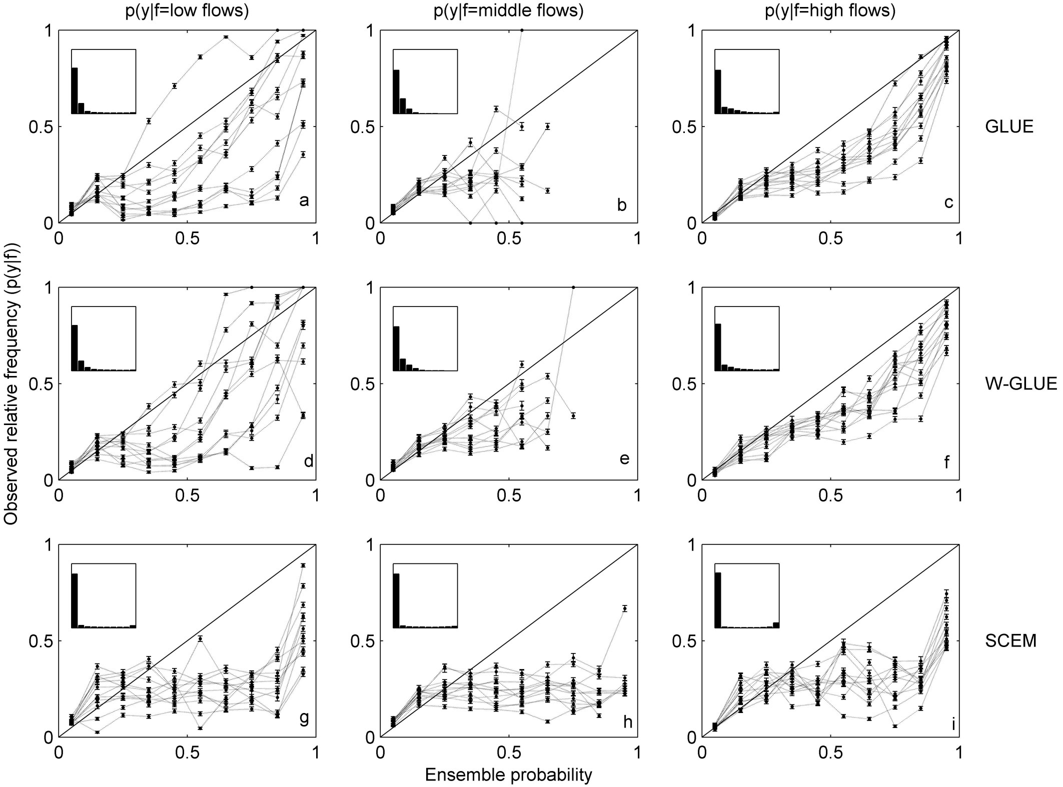


Fig. 7. Reliability diagrams for **(a, d, g)** low, **(b, e, h)** middle and **(c, f, i)** high flow simulations from the **(a–c)** GLUE, **(d–f)** W-GLUE and **(g–i)** SCEM methods. Each line represents a separate study site. Probability frequency diagrams for the simulations are shown in the inset, where the y axis is the ensemble frequency and the x-axis is the ensemble probability.

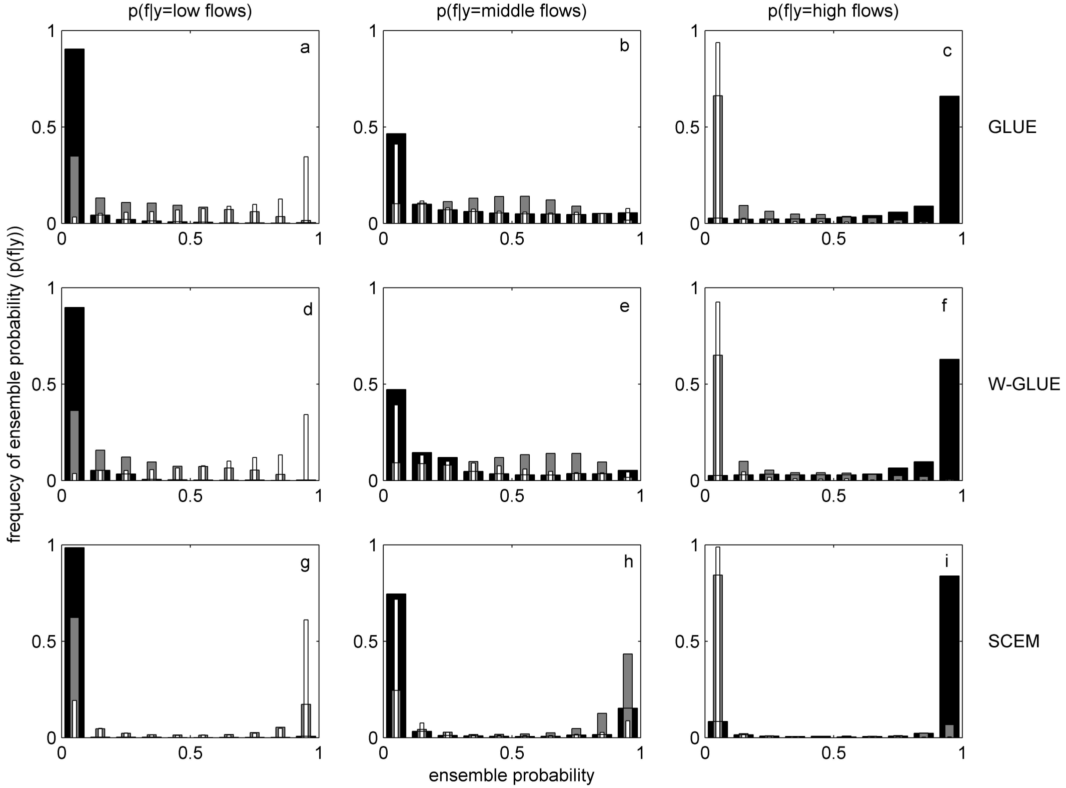


Fig. 8. Discrimination diagrams for simulation ensembles when the observations were in the **(a, d, g)** low flow, **(b, e, h)** middle flow, and **(c, f, i)** high flow categories from the **(a–c)** GLUE, **(d–f)** weighted GLUE, and **(g–i)** SCEM methods. The diagram depicts the average of all sites.

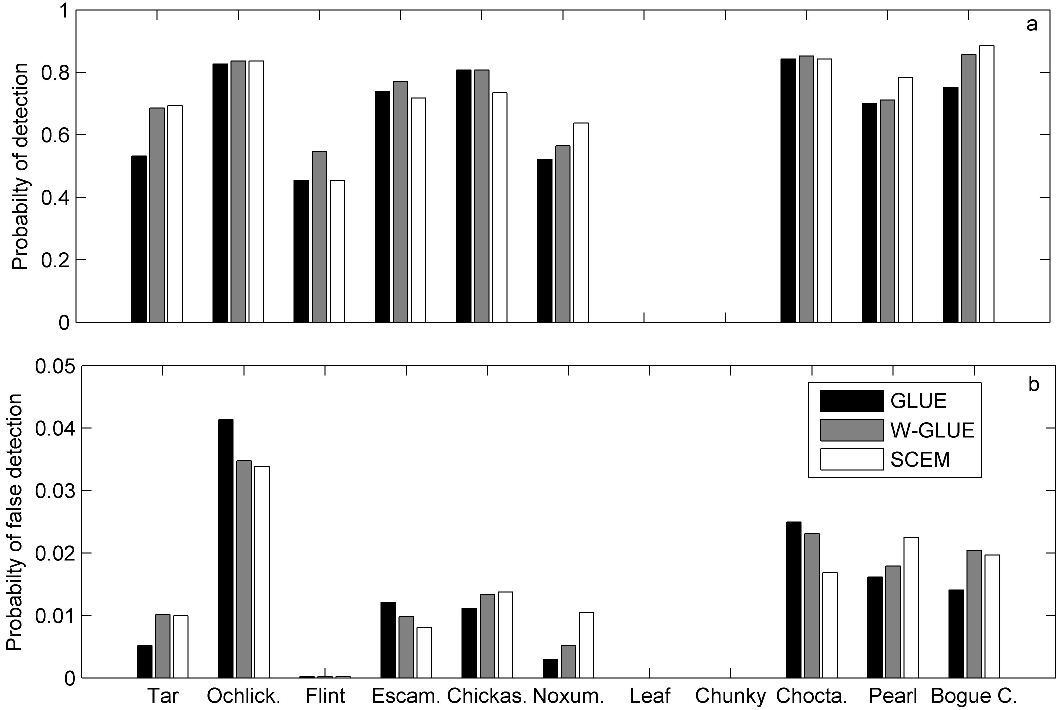


Fig. 9. The **(a)** probability of detection for floods and **(b)** probability of non-detection for floods for the ensemble medians at each site.

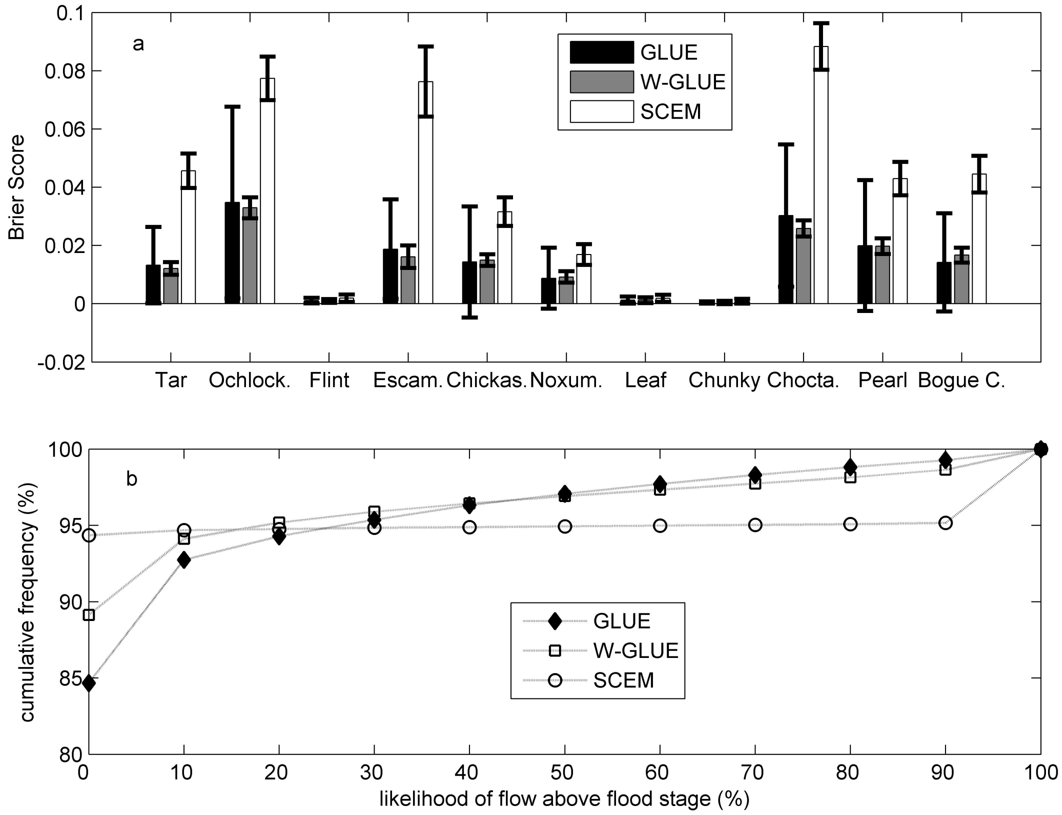


Fig. 10. The **(a)** Brier score for the categories of flood and no-flood evaluated for the ensembles for each site, and **(b)** cumulative frequency of the ensemble probability for flow above flood stage.