Hydrol. Earth Syst. Sci., 15, 3367–3382, 2011
www.hydrol-earth-syst-sci.net/15/3367/2011/
doi:10.5194/hess-15-3367-2011

Hydrology and Earth System Sciences

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 – Published in Hydrol. Earth Syst. Sci. Discuss.: 30 March 2011
Revised: 20 October 2011 – Accepted: 29 October 2011 – Published: 15 November 2011

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 two different parameter uncertainty estimation methods: the Generalized Uncertainty Likelihood Estimator (GLUE), and the Shuffle Complex Evolution Metropolis (SCEM). Model ensembles are generated 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 presented probabilistic metrics are easily adapted to model simulation ensembles and provide a robust analysis of model performance associated with parameter uncertainty. 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.

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

1 Introduction

In the classic definition, forecast verification is the process of assessing the skill of a forecast or set of forecasts (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; Laio and Tamea, 2007; Bartholmes et al., 2009; Renner et al., 2009; Brown et al., 2010; Demargne et al., 2010; Randrianasolo et al., 2010). One of the earliest attempts at verification was published by Finley (1884) who undertook an evaluation of the success of tornadoes forecasts. His early (and controversial) work sparked interest and a range of alternative methods in probabilistic verification, many of which are in use today (Murphy, 1997). Notable early verification papers in atmospheric and meteorological sciences have since included Cooke (1906) who undertook one of the first extensive verification studies, Ramsey (1926) and de Finetti (1937) who undertook early work in subjective probability theory, Murphy (1966) who overviewed probabilistic predictions and decision making, and Murphy and Epstein (1967) where the authors provided an overview of early development in probabilistic predictions and summarized terminology and definitions in the field.

More recent work on probabilistic verification measures includes Wilks (1997, 1998), numerous papers by Murphy (1991, 1995, 1996, 1997) as well as papers by Murphy and colleagues (e.g. Murphy and Winkler, 1987; Murphy and Wilks, 1998). All methods of verification, from early work by Finley (1884) to recent work by Bradley and Schwartz (2011), involve the comparison of a forecast (or set of forecasts) to the corresponding observation (Wilks, 2006). Murphy and Epstein (1967) lay out simple goals for forecast verification, including: evaluating the *value* of predictions, evaluating the *skill* of predictions, performing *quality control* on the forecast, and finally, investigating the cause(s) of prediction *errors*.

Model evaluation 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. However, the goals of forecast verification and model evaluation (i.e. verification) are analogous. Hydrologists are interested in the *value* and *skill* of their simulations, as well as the potential sources of error in their modeling system (Muleta and Nicklow, 2005; Beven, 2006; Gupta et al., 2006; Clark and Kavetski, 2010; Kavetski and Clark, 2010; Schoups et al., 2010). Despite the solid existence of probabilistic verification measures in the atmospheric and meteorological sciences, few metrics are routinely applied by the hydrologic community. Historically, evaluation of hydrologic models ensembles has been undertaken with standard deterministic measures, such as error, correlation, or bias, typically applied to the ensemble mean or median and occasionally application of a containing ratio metric (Xiong and O'Connor, 2008). 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). The recent growth of probabilistic streamflow estimates in hydrologic modeling, including ensemble data assimilation methods (Kitanidis and Bras, 1980a,b; Evensen, 1994; Margulis et al., 2002; Seo et al., 2003, 2009), multimodeling platforms (Ajami et al., 2007; Duan et al., 2007; Vrugt and Robinson, 2007; Franz et al., 2010), Ensemble Streamflow Prediction (ESP) and other probabilistic forecasting systems (Day, 1985; Krzysztofowicz, 2001; Faber and Stedinger, 2001; Franz et al., 2003, 2008; Bradley et al., 2004; Thirel et al., 2008) and post-processing techniques (Krzysztofowicz and Kelly, 2000; Montanari and Brath, 2004; Coccia and Todini, 2011; Weerts et al., 2011) warrants greater integration of probabilistic model evaluation into the hydrologic community.

There have been few publications on the probabilistic assessment of model performance. 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 ESP 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 and distinguishing model ensemble performance. We utilize two commonly applied parameter estimation methods Generalized Uncertainty Likelihood Estimator (GLUE; Beven and Binley, 1992) and the Shuffled Complex Evolution Metropolis (SCEM; Vrugt et al., 2003) and an operational rainfall-runoff model Sacramento Soil Moisture Accounting Model (SAC-SMA; Burnash et al., 1973) for demonstration purposes. We evaluate the uncertainty associated with model ensembles propagated through parameter estimates, although the metrics presented here are readily transferable to evaluate model performance from other probabilistic systems. We are not undertaking explicit evaluation of the "best" parameter estimation method being used, but rather highlighting how the applied metrics can better inform users on model performance and behavior when different results (ensemble hydrographs) are apparent. We also highlight unique challenges in applying probabilistic verification to hydrologic model ensembles and provide initial guidance on those measures which may be most suitable to the hydrologic community. The study sites, model, parameter estimation methods and verification metrics are presented in Sect. 2. Results from the application of the verification metrics are discussed in Sect. 3. Concluding statements are provided in Sect. 4.

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 hydroclimatic 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 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

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	Area	precipitation	discharge	Calibration	Verification
Gage ID	(km^2)	[mm yr^{-1}]	[mm yr^{-1}]	period	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

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 SAC-SMA 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 that provide relatively equal performance (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 by generating 10,000 parameter sets using Latin hypercube sampling (from a uniform distribution). The SAC-SMA model is run for each of the 10 000 sets. We define behavioral parameters sets as any set that produce simulations with a pre-defined likelihood threshold, using a Nash-Sutcliffe Efficiency (NSE) > 0.30 (Eq. 1, Sect. 2.4). This is a relatively non-restrictive threshold and the approach can result in a large number of behavioral sets.

The second 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

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}	0.1–0.7
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

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).

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)
Correlation	correlation coefficient (r)	joint distribution plots
Accuracy (error)	Nash-Sutcliffe efficiency (NSE), percent bias (PBias), root mean square error (RMSE)	containing ratio (CR)
Conditional Statistics		Reliability diagram , discrimination diagram, resolution
Categorical Statistics	probability of detection, probability of non-detection	Brier score (BS)

2.4 Verification methods

There are an extensive set of forecast verification measures that could be adopted for model evaluation. We selected those that are relevant to the modeling framework in the current study, are commonly applied, 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 five of the seven categories from these two sources (skill scores and confidence are not evaluated) 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. The metrics were applied to the verification period only.

The discharge ensemble (from GLUE or SCEM) are treated as a set of discrete variables by using the individual ensemble values and applying an empirical distribution. We first define $\{x_{(1)}, x_{(2)}, x_{(3)},...x_{(z)}\}$ as the set of simulated discharge values (the discharge ensemble) sorted in ascending order for one timestep (t) from an ensemble of size z.

2.4.1 Distribution properties

There are many measures of distribution, including the ensemble mean and median, but we are most interested in those that quantify the ensemble spread. Three metrics are used to evaluate the distribution of the ensemble members: the interquartile range (IQR), median absolute deviation (MAD), and range:

$$\overline{\text{IQR}} = \frac{1}{n}\sum_{t=1}^{n}\left(q_{0.75}(t) - q_{0.25}(t)\right) \quad (1)$$

$$\overline{\text{MAD}} = \frac{1}{n}\sum_{t=1}^{n} \text{median}_i \left|x_i(t) - x_{\text{med}}(t)\right| \quad (2)$$

and

$$\overline{\text{Range}} = \frac{1}{n}\sum_{t=1}^{n}\left(x_{(1)}(t) - x_{(z)}(t)\right) \quad (3)$$

where $q_{0.75}(t)$ and $q_{0.25}(t)$ are the 75th and 25th percentiles of the ensemble, respectively; $x_i(t)$ represents each ensemble member for timestep t; $x_{\text{med}}(t)$ is the ensemble median; $x_{(1)}(t)$ and $x_{(z)}(t)$ are the lowest and highest valued ensemble members, respectively; and n is the number of timesteps (Wilks, 2006).

Equations (1)–(3) are also used to evaluate the parameter ensembles, in which case $\{x_{(1)}, x_{(2)}, x_{(3)},...\ x_{(z)}\}$ is the set of z values for a single model parameter normalized by the parameter range (Table 2) and n becomes the number of model parameters (14 in this study).

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 the correlation coefficient are used to assess the correlation between the ensemble median and the observation. The correlation coefficient (r) is:

$$r = \frac{n\sum\limits_{t=1}^{N}\left(x_{\text{med}}(t)Q_{\text{obs}}(t)\right) - \left(\sum\limits_{t=1}^{N} x_{\text{med}}(t)\right)\cdot\left(\sum\limits_{t=1}^{N} Q_{\text{obs}}(t)\right)}{\sqrt{n\sum\limits_{t=1}^{N} x_{\text{med}}(t)^2 - \left(\sum\limits_{t=1}^{N} x_{\text{med}}(t)\right)^2}\cdot\sqrt{n\sum\limits_{t=1}^{N} Q_{\text{obs}}(t)^2 - \left(\sum\limits_{t=1}^{N} Q_{\text{obs}}(t)\right)^2}}. \quad (4)$$

where $Q_{\text{obs}}(t)$ is the observation at time t. In the probabilistic approach, the correlation between ensemble quantiles and the observations can be evaluated visually by plotting their joint distributions. Select ensemble quantiles (q_k) are chosen and plotted against the corresponding observation, where $0 \leq k \leq 1$ (e.g. $k = 0.10$ (10th percentile), $k = 0.25$ (25th percentile), etc.). This approach is similar to using a scatter plot.

2.4.3 Accuracy

The term accuracy refers to a measure of error in the simulation ensemble when compared to the observation. Three common deterministic measures of model accuracy are used to assess the ensemble median: the Nash Sutcliffe efficiency score (NSE), root mean squared error (RMSE), and percent bias (Pbias):

$$\text{NSE} = 1 - \left(\sum_{t=1}^{n}\left(x_{\text{med}}(t) - Q_{\text{obs}}(t)\right)^2 \Big/ \sum_{t=1}^{n}\left(Q_{\text{obs}}(t) - \overline{Q_{\text{obs}}(t)}\right)^2\right), \quad (5)$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(x_{\text{med}}(t) - Q_{\text{obs}}(t)\right)^2}, \quad (6)$$

$$\text{Pbias} = \left(\sum_{t=1}^{n}\left(x_{\text{med}}(t) - Q_{\text{obs}}(t)\right) \Big/ \sum_{t=1}^{n} Q_{\text{obs}}(t)\right)\cdot 100\,\%. \quad (7)$$

A simple measure of ensemble accuracy is the Containing Ratio (CR) (Xiong and O'Connor, 2008):

$$\text{CR} = \frac{1}{n}\sum_{t=1}^{n} I\left[Q_{\text{obs}}(t)\right] \quad (8)$$

where $I[\cdot]$ is an indicator function as follows:

$$I\left[Q_{\text{obs}}(t)\right] = \left\{\frac{1,\ x_{(1)}(t) < Q_{\text{obs}}(t) < x_{(z)}(t)}{0,\ \text{otherwise}}\right. . \quad (9)$$

$I[Q_{\text{obs}}(t)]$ equals 1 when the observation falls between the lowest and highest valued ensemble members and $I[Q_{\text{obs}}(t)]$ equals 0 when the observation falls outsize the ensemble bounds.

2.4.4 Conditional statistics

In the previous sections, we presented metrics that compare the simulated discharge values (i.e. median, minimum and maximum of the ensemble) to observed discharge values. In the following section, methods that evaluate probability values from the ensemble for specific discharge events are presented.

We first define $m_i(t)$ as the probability of a simulated streamflow event at a given timestep from the model ensemble, which can take on any of I values $m_1(t)$, m_2(t)... $m_I(t)$ (Wilks, 2006). The corresponding observation ($y_j(t)$) can take on any of J values $y_1(t)$, $y_2(t)$ ... $y_J(t)$. In this study, three possible observations (i.e. $J = 3$) are defined: low flow or a discharge value that is less than the 30th percentile of climatology; middle flow or a discharge value that is between the 30th and 70th percentiles of climatology; and high flow or a discharge value that is greater than the 70th percentile of climatology. Climatology is based on the available discharge data at each site (Table 1).

The probability of a simulated streamflow event is derived by computing the percentage of the ensemble members that fall within each flow category at a given timestep. The probability is rounded up to the nearest tenth probability, therefore the probability will fall within one of ten possible probability bins (0–10 %, >10 %–20 %, etc.). At a given timestep, the observation will have a value of 1 ($y_j(t) = 1$) for the flow category in which it was observed, and a value of 0 ($y_j(t) = 0$) for the flow categories in which it did not occur.

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:

$$p(m_i, y_j) = p(y_j|m_i)\ p(m_i); \quad i = 1, ..., I; \quad j = 1, ..., J \quad (10)$$

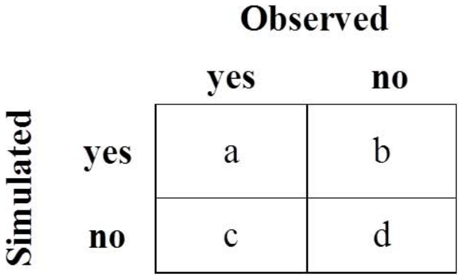
and the likelihood-base rate factorization:

$$p(m_i, y_j) = p(m_i|y_j)\ p(y_j); \quad i = 1, ..., I; \quad j = 1, ..., J. \quad (11)$$

The conditional distribution $p(y_j|m_i)$ in Eq. (10) is the more familiar measure of the two and can be plotted on a reliability diagram as a function of the ensemble probability. The ensemble probability is well calibrated if, for a given flow category, the relative frequency of the conditional event equals the ensemble probability (e.g. $p(y = \text{low flow}|m = 0.1) = 0.1$) and when plotted on the reliability diagram, the conditional event will plot along a 1:1 line (Murphy and Winkler, 1987, 1992; Wilks, 2006). To avoid confusion with the model parameter calibration discussion, hereafter we refer to the calibration of the ensemble probability as reliability.

The relative frequencies of the ensemble probabilities ($p(m_i)$) are plotted as an inset on the reliability diagram to indicate the sharpness, or resolution, of the ensembles (Wilks, 2006). Sharp ensembles will have narrowly distributed probability values where probability occurs most frequently in the extreme probability categories (i.e. 0–10 % and >90–100 %).

The likelihood distribution ($p(m_i|y_j)$) is a less intuitive measure, but very useful for evaluating how much probability the ensemble gives to the correct flow category compared to other possible categories. For all instances of an observation occurring in a given flow category, the conditional probability for all possible flows is computed: for example, the ensemble probability of a low flow given a low flow observation (p ($m = \text{low flow}|y = \text{low flow}$)), the ensemble probability of a middle flow given a low flow observation ($p(m = \text{middle flow}|y = \text{low flow})$), and the ensemble probability of a high flow given a low flow observation ($p(m = \text{high flow}|y = \text{low flow})$). These likelihood distributions can then be plotted on the discrimination diagram as a function of the ensemble probability. Ensembles are highly accurate if the majority of the ensemble members frequently fall within the flow category observed (in the previous example, this would be the low flow category), resulting in high probabilities for the observed flow category and low probabilities for the remaining flow categories. For such ensembles, the likelihood distributions for the different possible flows will not overlap to a great degree when plotted on the discrimination diagram and they are considered to have good discrimination for that flow category (Murphy and Winkler, 1987; Murphy et al., 1989; Wilks, 2006).



		Observed	
		yes	no
Simulated	yes	a	b
	no	c	d

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

2.4.5 Categorical statistics

The categorical statistics listed in Table 3 are applicable to dichotomous events where $I = J = 2$. These metrics are used here to evaluate the ability of the GLUE and SCEM 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 to assess the ability of the ensemble median to identify flood ($y(t) = 1$) and no-flood ($y(t) = 0$) events. The model ensemble median is classified as flood ($x_{\text{med}}(t) = 1$) if its value is larger than flood stage and no-flood ($x_{\text{med}}(t) = 0$) if its value is smaller than flood stage. A 2 × 2 contingency table is set up (Fig. 1) and all possible observation/simulation pairs are counted. Two measures are used to summarize the contingency table (Wilks, 2006): the probability of detection (POD):

$$p(x_{\text{med}}(t) = 1|y(t) = 1) = \text{POD} = \frac{a}{a + c} \quad (12)$$

and the probability of false detection (POFD):

$$p(x_{\text{med}}(t) = 1|y(t) = 0) = \text{POFD} = \frac{b}{b + d}. \quad (13)$$

POD values equal to one and POFD values close to zero are optimum.

The Brier Score (BS) (Brier, 1950) is used to evaluate the accuracy of the ensemble for the simulation of flood and no-flood events in a probabilistic manner. The BS is the mean squared error of the ensemble probability ($m_{\text{flood}}(t)$)

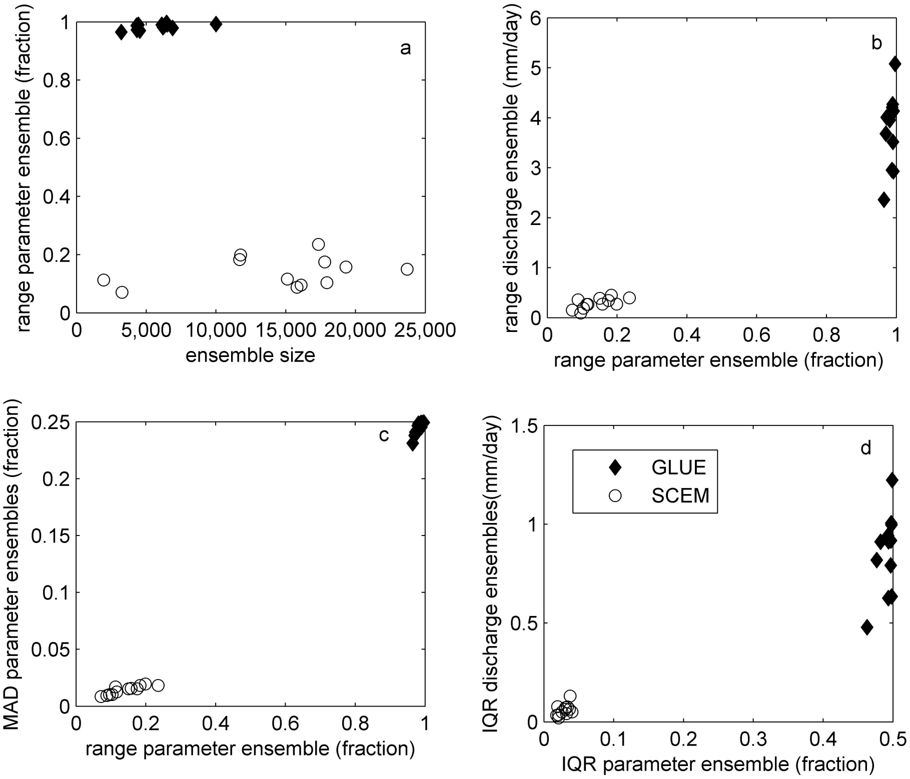


Fig. 2. Comparison of **(a)** parameter ensemble range to parameter ensemble size, **(b)** discharge ensemble range to parameter ensemble range, **(c)** mean absolute deviation (MAD) of the parameter ensembles to parameter ensemble range, and **(d)** interquartile range (IQR) of the discharge ensembles 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.

for flows higher than the flood level and the corresponding observations ($y(t)$) for all timesteps:

$$\mathrm{BS} = \frac{1}{n}\sum_{t=1}^{n}\left(m_{\mathrm{flood}}(t) - y(t)\right)^2. \tag{14}$$

A perfect BS is 0, and the score ranges from $0 \leq \mathrm{BS} \leq 1$ (Wilks, 2006).

3 Results

3.1 Distribution properties

Measures of distribution are first used to understand the nature of the parameter ensembles and their relation to the resulting discharge ensembles. In general we find weak positive correlations between the measures compared (Fig. 2). Correlation between the parameter ensemble size and range (Fig. 2a) is $r = 0.59$ for the GLUE and $r = 0.32$ for the SCEM. Although the SCEM parameter ensembles are larger than the GLUE parameter ensembles, the range of the SCEM parameter ensemble is much narrower, spanning less than 30 % of the feasible parameter space at all sites. The GLUE parameter values span almost the entire parameter space.

The range and IQR of the parameter and discharge ensembles are compared in Fig. 2b and d, respectively. Both give similar information: the parameter ensembles that have values distributed across a larger portion of the parameter space produce discharge ensembles with a larger distribution of values. The IQR does reveal characteristics about the ensembles that are not apparent when evaluating the range only. The IQRs of the SCEM parameter and discharge ensembles vary little among sites; by comparison the range values had more variation. This suggests that the central 50 % of the SCEM parameter sets are very similar, and the variation seen in the range comes from the upper and lower 25 % of the distribution. Figure 2c indicates that as the parameter ensemble range increases, the values deviate more from the median, rather than being concentrated near the median with only a few outliers.

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. The IQR and MAD give different information about the ensemble distributions than the range, but did not add significantly to our analysis.

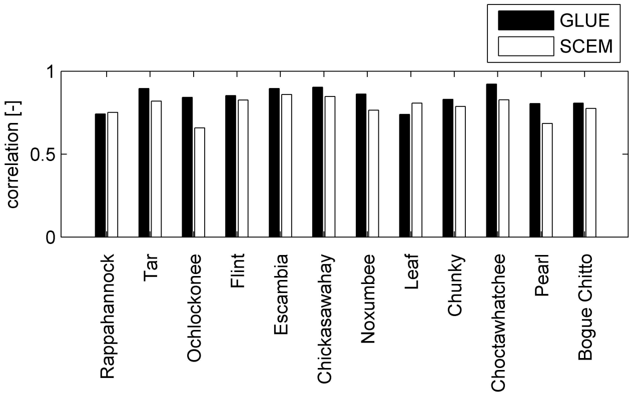


Fig. 3. The correlation coefficient for the ensemble median for all sites.

3.2 Correlation

On average, the ensemble medians from the GLUE and SCEM have good correlation to the discharge observations at the sites studied (Fig. 3). By plotting the joint distribution of the ensemble median and the observations it becomes apparent that the degree of correlation varies across the range of discharge values (Fig. 4, line x_{med}). In addition, the direction of the error (overestimating or underestimating the observation) can vary across the range of discharge values. Results for the Leaf River (Fig. 4c and f) are similar to several sites studied (Chunky, Pearl, Bogue Chitto, Ochlocknee Rivers) in that the ensemble medians underestimate discharge at both the lower and upper end of the range of flows, but overestimate the middle range.

The 10th and 90th percentiles ($q_{0.10}$ and $q_{0.90}$, respectively) of the GLUE ensembles almost always capture the low- and middle-range observations at all sites; whereas, the highest discharge ranges are often only contained within the upper 10 % (between $q_{0.90}$ and $q_{\max}$) (e.g. Fig. 4b and c). The SCEM ensembles (Fig. 4c–f) are narrow relative to the GLUE ensembles (Fig. 4a–c); this was also seen in Fig. 2b. In the Chickasawahay and the Noxumbee, the observation is mostly outside the bounds of the SCEM ensembles. The ability to capture the observation within the ensemble bounds will be quantified in the next section with the CR metric.

As was shown, plots of the joint distribution between the ensemble quantiles and the observation can be used to assess the performance of the median, the range of the ensemble and how well it captures the observation, and biases as a function of discharge magnitude. However, the information conveyed in Fig. 4 is potentially misleading. By depicting the data as a line, it appears as though the frequency of the discharge values is equal across the entire range of flows. This is not the case. At the twelve sites studied, flows above 5 mm day^{-1} represent only 2–6 % of the discharge observations, and therefore the large biases seen in highest discharge values occur for a very small number of samples. An alternative plotting scheme would be to plot the data as points rather than a line. However when 4745 points (the number of model timesteps) were plotted, the points overlapped each other at the lower discharge values (lower left corner of the plot) and the figure was difficult to read, particularly with the inclusion of multiple quantiles. In most modeling studies a long time period is preferred to assure sufficient calibration and verification of the model. However the large number of timesteps poses challenges from both the sampling standpoint (results are dominated by small discharge values which represent the majority of the samples used in computing the statistics) and the visualization standpoint.

3.3 Accuracy

The NSE, RMSE, and Pbias are common measures of model accuracy and are used to evaluate the ensemble median (Fig. 5). While useful for giving a concise assessment of model skill, the skill of the ensemble median for different magnitudes of discharge obviously cannot be understood from these numbers. In Fig. 4 we observed that the GLUE ensemble median tends to overestimate flows less than 5 mm day^{-1} and underestimate flows above that level. The SCEM ensemble medians on the other hand have a tendency to overestimate all flow ranges. As a result, the Pbias (Fig. 5b) and RMSE (Fig. 5c) of the GLUE are smaller than the SCEM, but for high flows neither method tends to be most accurate. Metrics in Fig. 5 measure summarize performance of the ensemble median for all timesteps, therefore, their value will be dominated by the skill of the ensemble median for low flow events. The large number of small discharge observations was also mentioned in the previous section and is a significant limitation of traditional model evaluation methods given that many hydrologic model applications are focused on simulating and predicting very high flows.

The advantage to using the measures in Figs. 3 and 5 is that most people in the hydrologic modeling community are familiar with them. However, comparison of results between different modeling studies is difficult because the accuracy of the model simulations are influenced by data quality and the hydrologic characteristics of the time period studied, which may vary from one study to the next. Additionally, the value of the RMSE scores is a function of the discharge magnitude for the study site. In this study, comparison of the two parameter estimation methods is only possible because we are applying both methods with the same model, time period, and study sites.

One other continuing challenge for the hydrologic community is that there is no standard acceptable values for the NSE, PBias and RMSE. Moriasi et al. (2007) is one of the few examples of an attempt to establish model performance guidelines. The authors recommend that model simulations can be judged as satisfactory if NSE > 5.0 and PBIAS ± 25 % for streamflow. However they specify that these guidelines are for evaluations made on a monthly

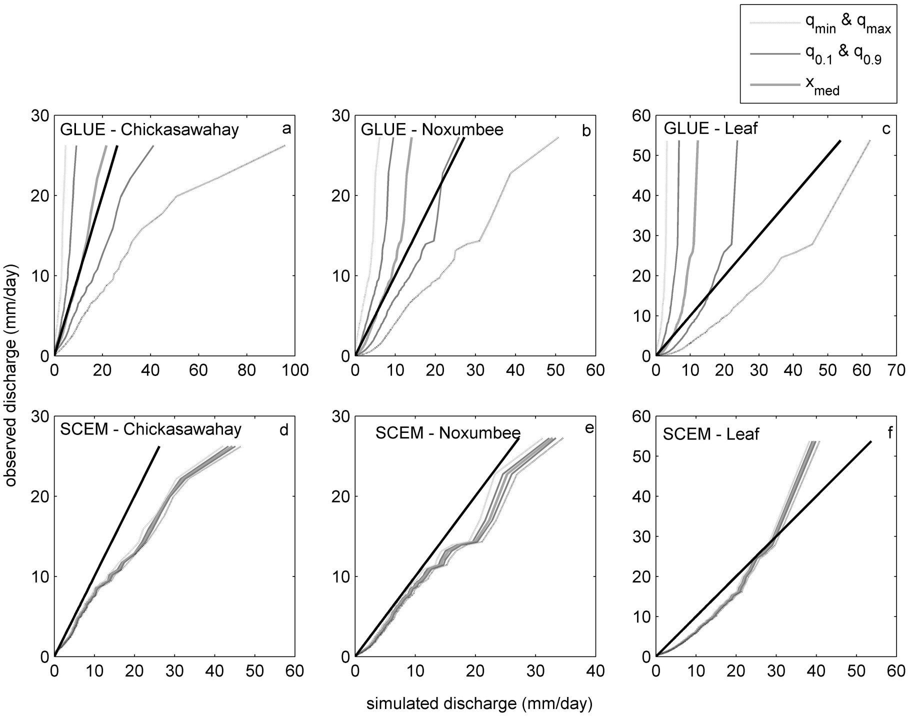


Fig. 4. The joint distribution of the lowest (q_{min}), highest (q_{max}), 10th ($q_{0.1}$), and 90th ($q_{0.9}$) percentiles, and the median (x_{med}, or 50th percentile) of the discharge ensembles and the observations from the **(a)–(c)** GLUE and **(d)–(f)** SCEM parameter estimation methods for select sites. The solid black line in the figures is the 1:1 line and indicates perfect correlation between the simulated and observed discharge.

timestep. Therefore they may not be applicable to this study which uses a daily timestep.

The CR is used to assess the accuracy of the ensemble bounds rather than focusing only on the median (Fig. 6). Given the biases and narrow range of the SCEM ensembles observed in Fig. 4, it is not surprising that the CR values are lower for this parameter estimation method. GLUE, which has much wider bounds captures the observations at a higher rate. Thus, CR is positively correlated with the range of the discharge and parameter ensembles for these parameter estimation methods (Fig. 6a and b). We also compute the CR by category by separating ensembles into one of three flow categories based on which observation occurred (low, middle, or high flow). The CR for each method are fairly consistent across all flow ranges but are slightly better for low flows (Fig. 6c).

In its standard application, the CR provides a useful summary of the accuracy of the uncertainty bounds, but does not consider the distribution of the ensemble members. It also cannot reveal whether the ensemble is over- or under-estimating the observation. More detailed information about the ensemble member distributions and associated performance can be obtained by considering multiple intervals within the ensemble, rather than the ensemble bounds only such as through the application of the rank histogram (Hamill and Collucci, 1997; Hamill, 2001; Wilks, 2006) or spread-bias diagram (Brown et al., 2010).

At a minimum, containing the observation within the uncertainty bounds is desired. However, an ensemble in which most members fall near the observation (indicating high probability for that observation) is more useful than an ensemble in which the members are equally distributed across many possible flow values (indicating similar probability for many possible observations). The conditional statistics in the next section are used to evaluate the probability distribution of the ensemble.

3.4 Conditional statistics

The relative frequencies of the observations given the ensemble probability for low, middle or high flows are shown on the reliability diagrams (Fig. 7). Based on the reliability diagram, we can conclude that the ensembles have low reliability on average. Reliability also tends to be best in probability bins of 30 % and less (i.e. 0–10 %, 10–20 %, 20–30 %) for all three flow categories. For probability of 40% and higher, the ensembles almost always overestimate the frequency of the observations. For a number of sites, the GLUE ensembles display good reliability for low flows and for high flows in the 0–10 % and 90–100 % probability bins (Fig. 7a and c). Reliability diagrams also reveal conditional biases in the ensembles. For example, the frequency of the observations is under-estimated for low probabilities but over-estimated

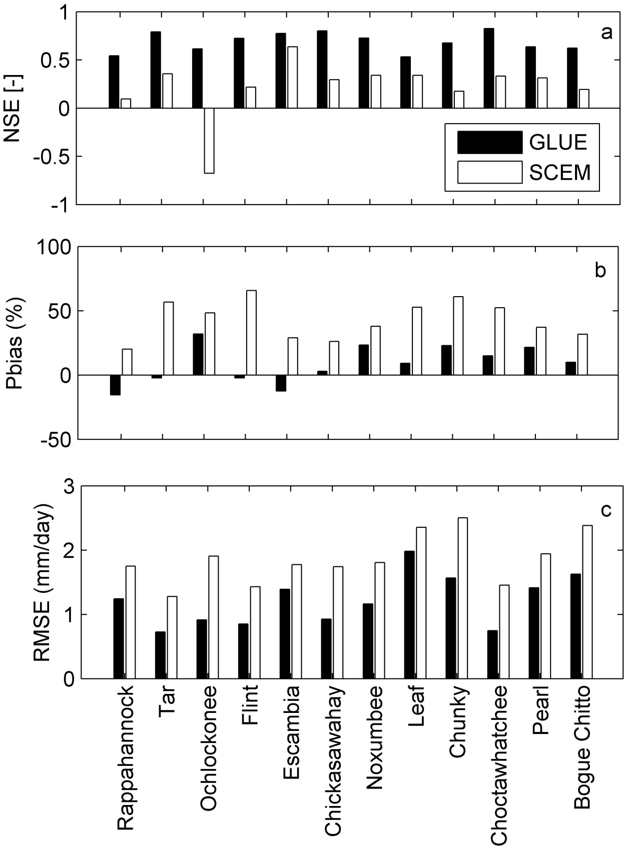


Fig. 5. The **(a)** Nash Sutcliffe Efficiency (NSE), **(b)** percent bias (Pbias), and **(c)** root mean square error (RMSE) for the ensemble median for all sites.

for high probabilities by the SCEM ensembles at many sites (Fig. 7d–f).

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 that the ensemble probabilities will reflect the frequency of the observations. However, because the number of samples in the middle probability bins is so small, particularly for the SCEM (Fig. 7d–f, insets), interpretation of the reliability results for those bins is difficult. The poor reliability results for ensemble probabilities between 20 % and 80 % (Fig. 7d–f) may be because there are too few samples to provide a good assessment of the ensemble.

The high frequency of ensemble probabilities within the 0–10 % and 90–100 % probability bins for the SCEM (Fig. 7d–f, inset) indicate high refinement or resolution. The GLUE ensembles are comparably less refined with more instances of ensemble probability in the 20–80 % range (Fig. 7a–c, inset). The small range of the SCEM ensembles (Fig. 2a and b) led to higher refinement of the ensemble probabilities.

For illustration purposes, results from the discrimination analysis have been averaged together for all sites for each parameter estimation methods (Fig. 8); for analysis of an individual site, one figure like Fig. 8 would be needed for each site. All methods produce ensembles with good discrimination for low flows (Fig. 8a and d) and high flows (Fig. 8c and f), and poorer discrimination for middle flows (Fig. 8b and e). When either a high or low flow occurs, the ensembles have the most difficulty discriminating the probability of the observed flow from the probability of middle flows. But they are skillful in not giving large probability to the extreme opposite category.

The low range of the SCEM (Fig. 2a and b) led to better discrimination of flows in all flow categories compared to the GLUE because the probabilities of the SCEM ensembles were better resolved (more probabilities in the extreme probability bins). However, the narrow range of the SCEM ensemble led to poorer performances in metrics that evaluated the ability of the ensemble to capture the event within the uncertainty bounds (i.e. the CR, Fig. 6). The higher ensemble spread in the GLUE (Fig. 2a and b) led to ensembles that tended to distribute the probability among the possible flow categories. For example, GLUE ensembles frequently assigned probability to low and high flows when middle flows occur, resulting in relatively poor discrimination for middle flows (Fig. 8b). Therefore, while the CR is high for the GLUE, the discrimination is lower compared to the SCEM.

Unlike some of the other methods presented here, reliability and discrimination are easy to interpret for an individual parameter estimation method or site without the need to compare to another example. The use of flow categories in this section allows for assessment of model performance under different conditions (i.e. low flows versus high flows) providing more information than the summary measures can. However, the 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, metrics such as discrimination require some degree of categorization. Additionally, we chose to use probability intervals of 10 %, another subjective decision that can be adjusted to varying situations and needs.

3.5 Categorical statistics

In the final analysis, three metrics are used to evaluate the simulation of flood events. 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.

POD and POFD are used to evaluate the ensemble median (Fig. 9). The POD and POFD are generally displayed using a relative operating characteristic (ROC) curve (Mason and Graham, 1999; 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). Both

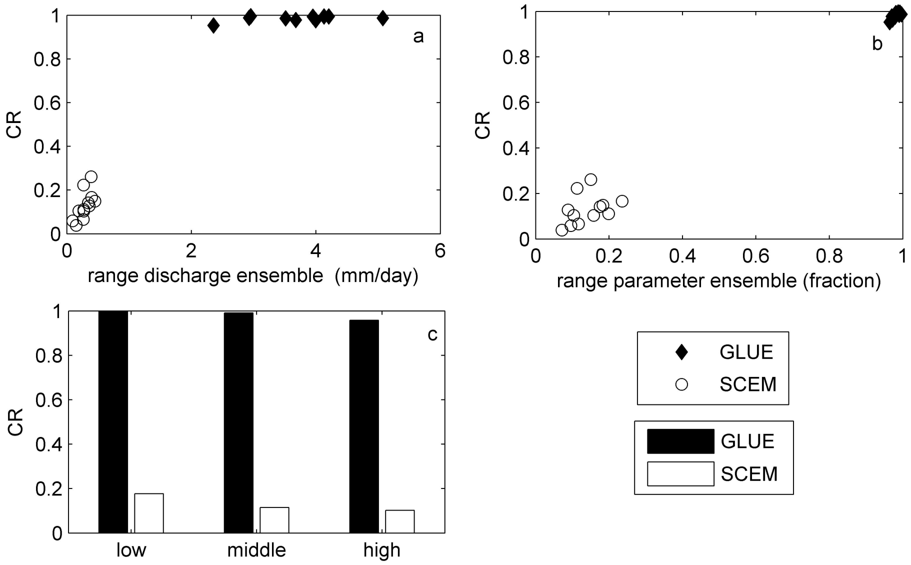


Fig. 6. Comparison of the containing ratios (CR) to the **(a)** discharge ensemble ranges and **(b)** parameter ensemble ranges; and **(c)** the average CR from the study sites by flow category (low, middle, high).

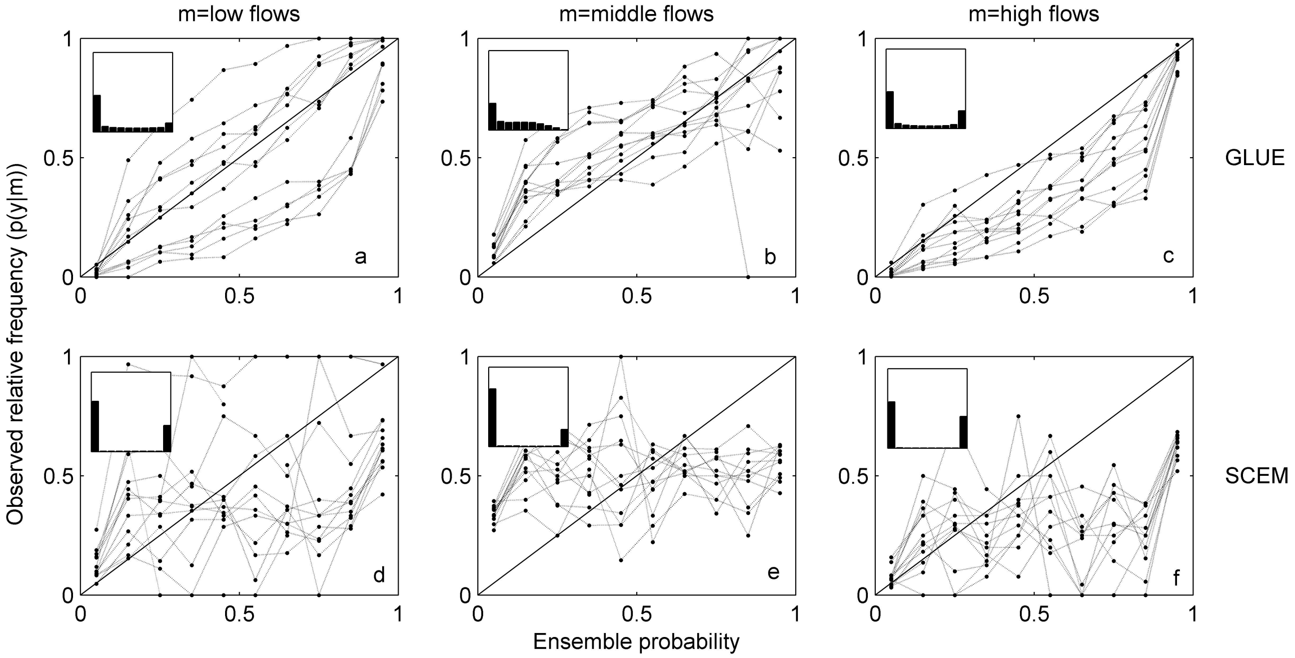


Fig. 7. Reliability diagrams for simulations of **(a,d)** low, **(b,e)** middle and **(c,f)** high flows from the **(a)**–**(c)** GLUE, and **(d)**–**(f)** SCEM methods. Each line represents a separate study site. Perfectly reliable ensembles will fall along the 1:1 line. If the ensembles are over-estimating (under-estimating) the conditional distribution will fall to the right (left) of the 1:1 line (Wilks, 2006).Probability frequency diagrams for the simulations are shown in the inset, where the y-axis is the relative frequency of the ensemble probability and the x-axis is the ensemble probability value.

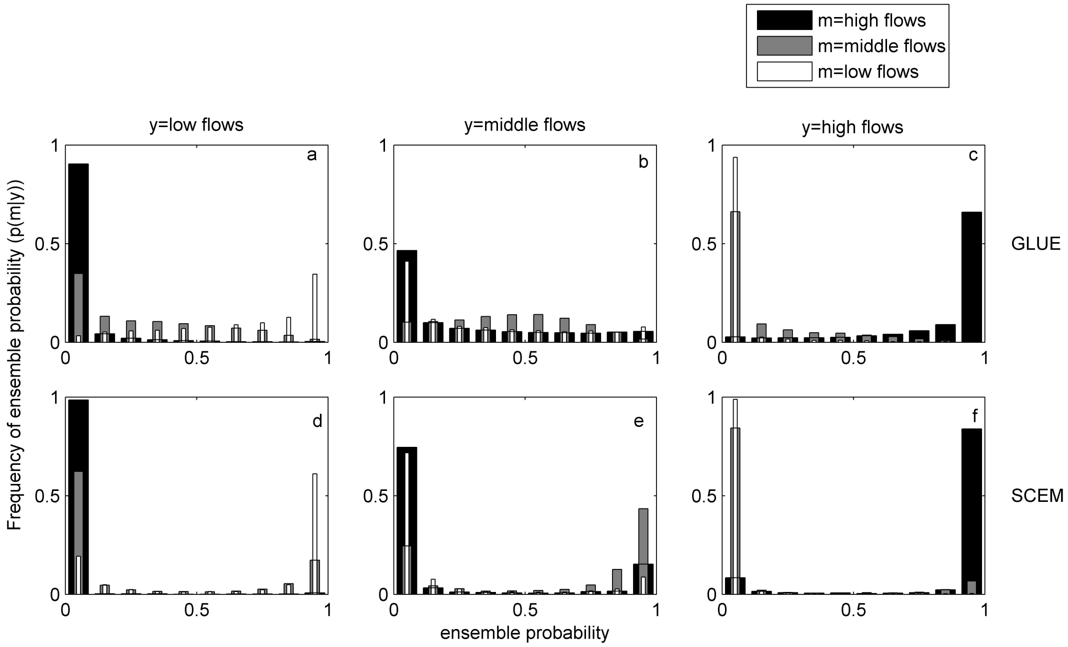


Fig. 8. Discrimination diagrams for simulation ensembles when the observations were in the **(a,d)** low, **(b,e)** middle, and **(c,f)** high flow categories from the **(a)**–**(c)** GLUE and **(d)**–**(f)** SCEM methods. The diagram depicts the average of results from all sites.

methods perform similarly. Recall from Fig. 4 that the medians tend to underestimate the highest flows, this negative bias in the upper range of the discharge values results in low POFD scores (Fig. 9b). The large negative bias in the median for the Leaf (Fig. 4c and f) and Chunky Rivers (not shown in Fig. 4) leads to a low POFD (Fig. 9b), but results in no skill for POD (Fig. 9a) at these sites.

The BS is used to evaluate the ensemble probability for a flood event (Fig. 10a). While the perfect BS is zero, this is another metric for which the number itself holds little meaning without comparison to other ensembles. When evaluating forecasts, the BS from the forecasts are often compared to climatology. In our comparison, the GLUE has slightly better BSs than the SCEM because the range of the ensembles is larger and, therefore, the ensemble tend to assign some probability to flood events when they occur. 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 % chance of flows above flood stage on average 95 % of the time. The GLUE gives probability for floods more frequently. As a result, the SCEM does more poorly than the GLUE for the BS.

Note that the BS for predictions above flood stage produce the same BS for predictions below flood stage. For this reason, the failure for the ensemble to simulate floods for Leaf and Chunky Rivers (Fig. 9a), produces a very low BS (Fig. 10a) because they predict no-floods well. Because this summary value is an evaluation of no-flood as well as flood events, it is heavily influenced by the large number of timesteps with no-flood. This leads to very low BSs event though the ensembles tended to have larger biases for very high flows.

4 Concluding remarks

When evaluating ensembles of simulations, 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 sampling of metrics that are traditionally applied for verification of forecasts, and have shown these to be informative for evaluation and comparison of ensemble streamflow simulations. A considerable amount of information about the uncertainty estimation methods can be obtained when treating the simulations in a probabilistic manner.

A critical skill of a probabilistic simulation is the ability to indicate which flow is most likely, rather than just merely capture the event using large uncertainty bounds. 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. As shown, discrimination and reliability diagrams give information about the accuracy and precision of the uncertainty estimates. The use of flow

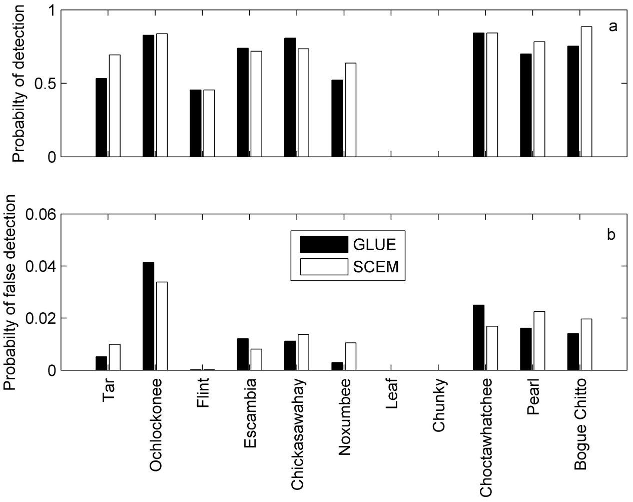


Fig. 9. The **(a)** probability of detection for floods and **(b)** probability of non-detection for floods for the ensemble medians at each site.

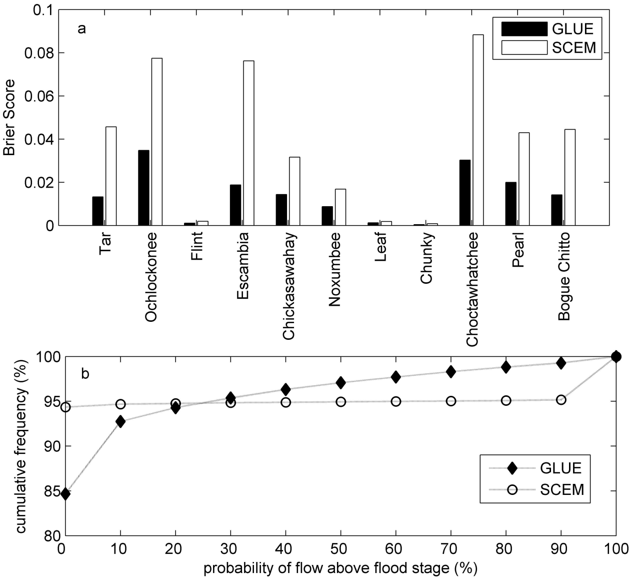


Fig. 10. The **(a)** Brier score for simulations of flood events for each site, and **(b)** cumulative frequency of the ensemble probability for flow above flood stage.

categories and the joint distribution plots allow analysis of the ensembles for discharge levels of interest.

We have identified some challenges when using forecast verification metrics for model ensemble evaluation. First, most forecast verification metrics were developed for forecasts of a single variable (e.g. rain or no rain, or peak discharge) to occur over some forecast interval, whereas model simulations produce a continuous variable most often evaluated at the model timestep. This means that in the case of evaluating model simulations, the sample size will likely be very large. Furthermore, the number of timesteps with low flows will be very large relative to the higher flows and model skill for low flows will dominate the results. Because low flows are often the range of least interest, approaches to limit the influence of low discharge events on the statistics should be investigated. One possible approach to deal with variations in sample sizes across flow regimes is to evaluate categories of flows as shown. But careful consideration of the influence of the sample size and sampling distribution on the confidence of the verification metric, an issue not addressed in this study, should be taken (Bradley et al., 2003; Wilks, 2006). Because probabilistic statistics rely on a significant number of model-observation pairs to obtain meaningful results (Wilks, 2006), evaluation of the model uncertainty associated with flood events will be limited by small sample sizes in most cases. Common problems such as identifying flow and probability thresholds or appropriate distributions exist and, because they may be treated differently in different studies, will limit the ability to compare results across different studies. Finally, we did not test for time-dependent clustering of the ensemble members or independence of the events analyzed, such as described by Christoffersen (1998), to determine statistical correctness. There is significant memory in a sequence of hydrologic model outputs and hydrologic observations, which violates the assumption of sample independence. Investigation of this issue with respect to hydrologic model and forecast verification is a recommended topic for future studies.

Nonetheless, advanced probabilistic verification metrics developed for forecast verification provide a rigorous platform by which modeling methods can be evaluated and compared. The application of these metrics requires no information in addition to what is already available as part of the traditional model validation methodology, and allows consideration of the entire ensemble or uncertainty range in the approach. These measures are much more informative about the 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. We would like to thank James Brown and two anonymous reviewers for their helpful comments on an earlier draft of this paper. We would also like to thank Alexandre Dulac and Sonya Lopez for their contributions to this work. This research was partially supported by a grant from the National Oceanic Atmospheric Administration (NOAA) National Weather Service (#NA07NWS4620013) and the National Aeronautics and Space Administration (#NNX10AQ77G).

Edited by: A. Montanari

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, 1973.

Bradley, A. A. and Schwartz, S. S.: Summary verification measures and their interpretation for ensemble forecasts, Mon. Weather Rev., 3075–3089, 2011.

Bradley, A. A., Hashino, T., and Schwartz, S. S.: Distributions-oriented verification of probability forecast for small data samples, Weather Forecast., 18, 903–917, 2003.

Bradley, A. A., Schwartz, S. S., and Hashino, T.: Distributions-oriented verification of ensemble streamflow predictions, J. Hydrometeorol., 5, 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, N.Y., 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.

Coccia, G. and Todini, E.: Recent developments in predictive uncertainty assessment based on the model conditional processor approach, Hydrol. Earth Syst. Sci., 15, 3253–3274, doi:10.5194/hess-15-3253-2011, 2011.

Cooke, W. E.: Forecasts and verifications in Western Australia, Mon. Weather Rev., 34, 23–24, 1906.

Christoffersen, P. F.: Evaluating interval forecasts, Int. Econ. Rev., 39, 841–862, 1998.

Day, G. N.: Extended streamflow forecasting using NWSRFS, J. Water Resour. Plann. Manage., 111, 157–170, 1985.

De Finetti, B.: Foresight: its logical laws, its subjective sources, in: Studies in Subjective Probability, edited y: Kyburg Jr., H. E. and Smokler, H. E., Wiley, New York, 1964, 94–158, 1937.

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, 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. Opt. Theory App., 76, 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.

Evensen, G.: Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, J. Geophys. Res., 99, 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.

Finley, J. P.: Tornado predictions, Am. Meteorol. J., 1, 85–88, 1884.

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.: Snow model verification using ensemble prediction and operational benchmarks, J. Hydrometeorol., 9, 1402–1415, 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.

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 I. analysis of uncertainty, Water Resour. Res., 16, 1025–1033, 1980a.

Kitanidis, P. K. and Bras, R. L.: Real-Time Forecasting with a conceptual hydrologic model 2. applications and results, Water Resour. Res., 16, 1034–1044, 1980b.

Krzysztofowicz, R.: The case for probabilistic forecasting in hydrology, J. Hydrol., 249, 2–9, 2001.

Krzysztofowicz, R. and Kelly, K. S.: Hydrologic uncertainty processor for probabilistic river stage forecasting, Water Resour. Res., 36, 3265–3277, 2000.

Laio, F. and Tamea, S.: Verification tools for probabilistic forecasts of continuous hydrological variables, Hydrol. Earth Syst. Sci., 11, 1267–1277, doi:10.5194/hess-11-1267-2007, 2007.

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, 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.

Montanari, A. and Brath, A.: A stochastic approach for assessing the uncertainty of rainfall-runoff simulations, Water Resour. Res., 40, W01106, doi:10.1029/2003WR002540, 2004.

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

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 note on the utility of probabilistic predictions and the probability score in the cost–loss ratio decision situation, J. Appl. Meteorol., 5, 534–537, 1966.

Murphy, A. H.: Forecast verification: Its complexity and dimensionality, Mon. Weather Rev., 119, 1590–1601, 1991.

Murphy, A. H.: A coherent method of stratification within a general framework for forecast verification, Mon. Weather Rev., 123, 1582–1588, 1995.

Murphy, A. H.: The Finley affair: A signal event in the history of forecast verification, Weather Forecast., 11, 3–20, 1996.

Murphy, A. H.: Forecast Verification, in: Economic value of weather and climate forecasts, edited by: Katz, R. W. and Murphy, A. H., Cambridge University Press, Cambridge, MA, 240 pp., 1997.

Murphy, A. H. and Epstein, E. S.: A note on probability forecasts and "Hedging", J. Appl. Meteorol., 6, 1002–1004, 1967.

Murphy, A. H. and Wilks, D. S.: A case study of the use of statistical models in forecast verification: Precipitation probability forecasts, Weather Forecast., 13, 795–810, 1998.

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.

Murphey, A. H., Brown, B. G., and Chen, Y.: Diagnostic verification of temperature forecasts, Weather Forecast., 4, 485–501, 1989.

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

Ramsey, F. P.: Truth and Probability, in: The Foundations of Mathematics and Other Logical Essays, edited by: Braithwaite, R. B.,Humanities Press, New York, 1950, 156–198, 1926.

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, 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, 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, 548–565, 2006.

Weerts, A. H., Winsemius, H. C., and Verkade, J. S.: Estimation of predictive hydrological uncertainty using quantile regression: examples from the National Flood Forecasting System (England and Wales), Hydrol. Earth Syst. Sci., 15, 255–265, doi:10.5194/hess-15-255-2011, 2011.

Wilks, D. S.: Resampling hypothesis tests for autocorrelated fields, J. Climate, 10, 65–82, 1997.

Wilks, D. S.: Multisite generalizations of a daily stochastic precipitation generation model, J. Hydrol., 210, 178–191, 1998.

Wilks, D. S.: Statistical Methods in the Atmospheric Sciences, 2nd Edn., 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.