Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging

Jasper A. Vrugt[1] and Bruce A. Robinson[1]

Received 23 December 2005; revised 25 August 2006; accepted 8 September 2006; published 17 January 2007.

[1] Predictive uncertainty analysis in hydrologic modeling has become an active area of research, the goal being to generate meaningful error bounds on model predictions. State-space filtering methods, such as the ensemble Kalman filter (EnKF), have shown the most flexibility to integrate all sources of uncertainty. However, predictive uncertainty analyses are typically carried out using a single conceptual mathematical model of the hydrologic system, rejecting a priori valid alternative plausible models and possibly underestimating uncertainty in the model itself. Methods based on Bayesian model averaging (BMA) have also been proposed in the statistical and meteorological literature as a means to account explicitly for conceptual model uncertainty. The present study compares the performance and applicability of the EnKF and BMA for probabilistic ensemble streamflow forecasting, an application for which a robust comparison of the predictive skills of these approaches can be conducted. The results suggest that for the watershed under consideration, BMA cannot achieve a performance matching that of the EnKF method.

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

1. Introduction and Scope

[2] In the last few decades hydrologists have made tremendous progress in the development and application of watershed models for the analysis of hydrologic systems and to provide accurate flood forecasting techniques. Predictions with these models are often deterministic, focusing on the most probable forecast, without an explicit estimate of the associated uncertainty. However, uncertainty arises from incomplete process representation, uncertainty in initial conditions, input, output, and parameter error. Quantifying these uncertainties is necessary to assess model quality and predictive capability.

[3] In the past few years, ensemble-based forecasting methods based on sequential data assimilation approaches have become increasingly popular. State-space filtering methods continuously update the states in the model when new measurements become available. This approach improves model forecast accuracy, and provides a means for explicitly handling the various sources of uncertainty in hydrologic modeling. Most recently, techniques based on the ensemble Kalman filter (EnKF) [*Evensen*, 1994] have been shown to have the power and flexibility required for data assimilation using conceptual watershed models [*Vrugt et al.*, 2005, 2006a; *Moradkhani et al.*, 2005a, 2005b]. In particular, *Vrugt et al.* [2005] presented the simultaneous optimization and data assimilation method (SODA), which uses the EnKF to recursively update the model states while estimating time-invariant values for the model parameters using the shuffled complex evolution Metropolis stochastic ensemble optimization approach (SCEM-UA) [*Vrugt et al.*, 2003]. In addition, *Moradkhani et al.* [2005a, 2005b] presented two different dual-state parameter estimation methods based on the EnKF and sequential Monte Carlo techniques [*Gordon et al.*, 1993].

[4] Despite this progress, a potential limitation of hydrologic uncertainty analyses is that they are typically carried out using a single conceptual mathematical model of the system under study, thereby reducing the size of the plausible model space. Analyses of predictive uncertainty based on a single hydrologic concept are prone to statistical bias and underestimation of uncertainty [*Hoeting et al.*, 1999; *Neuman*, 2003; *Raftery et al.*, 2003, 2005]. This often leads to overconfidence in the predictive capabilities of the model, which the available hydrologic data may not justify.

[5] Notable exceptions in which multimodel ensembles averages were used to quantify conceptual uncertainties include studies by *Shamseldin et al.* [1997], *Shamseldin and O'Connor* [1999], *Georgekakos et al.* [2004], and *Ajami et al.* [2005]. In particular, A. E. Raftery and coworkers at the Seattle School of Statistics recently proposed a formal framework, called Bayesian model averaging (BMA), to optimally combine the predictive capabilities of several competing statistical models and to assess their joint predictive uncertainty. The BMA method does not select a single "best" model, but rather conditions simulations on the entire ensemble of models included. In BMA the overall forecast PDF is a weighted combination of forecast PDFs centered on the individual model forecasts

[1]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

[*Raftery et al.*, 2003, 2005]. The weights are the estimated posterior model probabilities, representing each model's relative forecast skill in the training period. Studies applying the method to a range of different forecasting problems have demonstrated that BMA produces more accurate and reliable predictions than other available multimodel techniques [*George and McCulloch*, 1993; *Raftery et al.*, 1997; *Clyde*, 1999; *Raftery and Zheng*, 2003; *Ye et al.*, 2004; *Ajami et al.*, 2005, 2007; Q. Duan et al., Multi-model ensemble hydrologic prediction using Bayesian model averaging, submitted to *Advances in Water Resources*, 2005]. Another strength is that there are no restrictions on diversity of conceptual and numerical models that can be included, since the model averaging simply requires a numerical representation of each model, and the procedural steps are conducted outside of the scope of the calculations of the individual models.

[6] The present study compares the power and applicability of the EnKF and BMA for probabilistic ensemble streamflow forecasting. Both methods are implemented and compared using a set of conceptual watershed models of varying complexity and degree of parameterization. Models are applied to historical streamflow data from the Leaf River basin in Mississippi, a site for which a long record of hydrologic measurements is available to test the methods.

[7] The remainder of this paper is organized as follows. Section 2 presents a condensed description and overview of the EnKF and BMA methods, and section 3 briefly describes the conceptual watershed models and hydrologic data used in our analysis. In section 4, results are presented for the BMA and EnKF methods, and the usefulness and applicability of these methods for predictive uncertainty analysis in hydrologic rainfall-runoff modeling are compared. In this section the forecast skills and uncertainty bounds of the methods are the primary focus. In section 5 we present and test three possible improvements of the BMA method for hydrologic modeling and forecasting. Finally, a summary with conclusions is presented in section 6.

2. Predictive Uncertainty Analysis Methods

[8] The traditional approach to hydrologic prediction and uncertainty analysis is to postulate a model structure and treat its parameters as being imperfectly known. This approach assumes that model-data mismatches are solely attributable to uncertainty in the parameter values. However, other errors, such as conceptual model errors arising from inadequate representation of physical processes, are typically not handled in an explicit manner. The two techniques compared in the present study are alternative methods for treating errors, including conceptual model errors, more comprehensively.

2.1. Ensemble Kalman Filter

[9] Sequential data assimilation (SDA) methods provide a general framework to explicitly treat input, output, and model structural uncertainty in hydrologic modeling. SDA methods continuously update the states in the model when new measurements become available. In principle, this approach improves the model forecasts and enables evaluation of the forecast accuracy. The prototype of the SDA methods, the Kalman filter (KF) [*Kalman*, 1960] was developed in the 1960s for optimal control of systems governed by linear equations. The KF is a sequential filter method that integrates the model forward in time, and reinitializes or updates the states in the model whenever measurements become available.

[10] The KF is best described by first writing the discrete time evolution of the state vector ψ in the hydrologic model Φ as a stochastic equation [see also *Vrugt et al.*, 2006a]:

$$\psi_{t+1} = \Phi\left(\psi_t, \widetilde{X}_t, \theta\right) + q_t \tag{1}$$

where $\widetilde{X}$ represents the observed forcing, θ is a parameter set, t denotes time, and q_t is a dynamical noise term representing errors in the conceptual model formulation. This stochastic forcing term flattens the probability density function of the states during the integration. The model output predictions are directly related to the model state:

$$y_t = H(\psi_t) \tag{2}$$

where the measurement operator $H(\cdot)$ maps the state-space into the measurement or model output space. In this application, $H(\cdot)$ mathematically defines the relationship between the simulated states and streamflow emanating from the catchment outlet.

[11] The observation equation (2) is assumed to have a random additive error ε_t representing the measurement error:

$$\widetilde{y}_t = H(\psi_t^*) + \varepsilon_t \qquad \varepsilon_t \sim N\left(0, \sigma_t^0\right) \tag{3}$$

where σ^0 denotes the error deviation of the observations, and ψ^* denotes the true model states. At each measurement time, an output observation, $\widetilde{y}_t$, becomes available, and the output forecast error z_t is computed:

$$z_t = \tilde{y}_t - H\left(\psi_t^f\right) \tag{4}$$

and the forecasted states, ψ_t^f, are updated using the standard KF analysis equation:

$$\psi_t^u = \psi_t^f + K_t\left(\tilde{y}_t - H\left(\psi_t^f\right)\right) \tag{5}$$

where ψ_t^u is the updated or analyzed state, and K_t denotes the Kalman gain, the size of which depends on the size of the measurement and model error. The implementation of equation (5) is described further in section 3.

[12] Finally, the analyzed state recursively feeds the next state propagation step in the model:

$$\psi_{t+1}^f = \Phi\left(\psi_t^u, \tilde{X}_t, \theta\right) \tag{6}$$

[13] The virtue of the KF method is that it offers a very general framework for segregating and quantifying the effects of input, output, and model structural error in hydrologic modeling. Specifically, uncertainty in the model formulation and forcing data are specified through the stochastic forcing terms q and ε, whereas errors in the input data are quantified by stochastically perturbing the elements

of X (see *Clark and Slater* [2006] for a viable method). Despite this strength, the recursive KF approach has not received widespread implementation in hydrologic modeling for a variety of practical reasons. Most importantly, the implementation requires that the highly nonlinear original model be rendered into a continuously differentiable state-space form, which requires various modifications and or approximations [*Kitanidis and Bras*, 1980a, 1980b; *Georgakakos et al.*, 1988; *Georgakakos and Sperflage*, 1995; *Refsgaard*, 1998; *Seo et al.*, 2003].

[14] To resolve these issues, *Evensen* [1994] proposed the ensemble Kalman filter (EnKF), which uses a Monte Carlo method to generate an ensemble of model trajectories from which the time evolution of the probability density of the model states and related error covariances are estimated. When an output measurement is available, each forecasted ensemble state vector ψ_t^f is updated to ψ_t^u by means of the linear updating rule given in equation (5). The strength of the gain K_t depends on the strength of the cross covariance between the state variables of interest and the model outputs for which measurements are available. The cross covariance is approximated using the information contained in the ensembles.

[15] The ensemble Kalman filter (EnKF) has gained popularity in many fields of study because of its simple conceptual formulation and relative ease of implementation. The EnKF avoids many of the problems associated with the traditional KF and extended Kalman filter (EKF) method.

2.2. Bayesian Model Averaging

[16] While the EnKF explicitly treats input, output, and model structural uncertainty in hydrologic modeling, the method operates using a single mathematical model. Adopting only a single model may lead to statistical bias, and underestimation of uncertainty, particularly for situations in which there is a fundamental lack of knowledge of the physical conditions or physics that apply to the modeled system. Ensemble Bayesian model averaging (BMA), recently proposed by *Raftery et al.* [2005], is especially developed to better treat conceptual model uncertainty by not only conditioning on a single "best" model, but on an entire ensemble of plausible models.

[17] Let f $\{f_1, \ldots f_K\}$ denote an ensemble of predictions obtained from K different conceptual mathematical models, $M = \{M_1, \ldots, M_k\}$, and Δ be the quantity of interest. In BMA, each ensemble member forecast, f_k, is associated with a conditional PDF, $g_k(\Delta|f_k)$, which can be interpreted as the PDF of Δ given f_k, conditional on f_k being the best forecast in the ensemble. The BMA predictive model can then be expressed as

$$p(\Delta|f_1, \ldots, f_k) = \sum_{k=1}^{K} w_k g_k(\Delta|f_k) \tag{7}$$

where w_k denotes the posterior probability of forecast k being the best one. The w_k values are nonnegative and add up to one, and they can be viewed as weights reflecting an individual model's relative contribution to predictive skill over the training period.

[18] The original ensemble BMA method described by *Raftery et al.* [2005] assumes that the conditional PDFs $g_k(\Delta|f_k)$ of the different ensemble members can be approximated by a normal distribution centered at a linear function of the original forecast, $a_k + b_k f_k$ and standard deviation σ:

$$\Delta|f_k \sim N\left(a_k + b_k f_k, \sigma^2\right) \tag{8}$$

The values for a_k and b_k are bias correction terms that are derived by simple linear regression of Δ on f for each of the individual ensemble members.

[19] Normal predictive distributions seem to be inappropriate for streamflow, and for any other quantity primarily driven by precipitation [*Sloughter et al.*, 2006]. However, we argue that this is a reasonable starting point to illustrate the usefulness of the BMA method. In the final section of this paper we test this hypothesis by examining the performance of the BMA method using the gamma distribution, which is arguably a more appropriate predictive distribution for streamflow forecasting.

[20] The BMA predictive mean can be computed as

$$E[\Delta|f_1, \ldots, f_k] = E[p(\Delta|f_1, \ldots, f_k)] = E\left[\sum_{k=1}^{K} w_k g_k(\Delta|f_k)\right] = \sum_{k=1}^{K} w_k f_k \tag{9}$$

If we denote space and time with subscripts s and t respectively, so that f_{kst} is the kth forecast in the ensemble for place s and time t, the associated variance of the BMA predictive mean in equation (9) can be computed as [*Raftery et al.*, 2005]

$$Var[\Delta_{st}|f_{1st}, \ldots, f_{Kst}] = \sum_{k=1}^{K} w_k \times \left((a_k + b_k f_{kst}) - \sum_{l=1}^{K} w_l (a_l + b_l f_l)\right)^2 + \sigma^2 \tag{10}$$

The variance of the BMA prediction defined in equation (10) consists of two terms, the first representing the ensemble spread, and the second representing the within-ensemble forecast variance.

[21] Successful implementation of the BMA approach just developed requires specification of the weights and variance of the individual ensemble members generated by the different conceptual models. Following *Raftery et al.* [2005], we estimate the values of w_k, $k = 1,\ldots,K$; and σ^2 by posing the optimization problem in a maximum likelihood context. If we assemble the BMA parameters in the vector θ, the log likelihood function $\ell(\theta)$, corresponding to the predictive model in equation (7), can be expressed as

$$\ell(\theta) = \sum_{s,t} \log\left(\sum_{k=1}^{K} w_k g_k(\Delta_{st}|f_{kst})\right) \tag{11}$$

where the summation is over s and t to include all the observations in the training set (n).

[22] In our experiments, we assume that the maximum likelihood estimate provides a meaningful measure to optimize the BMA weights and variance. There are however,

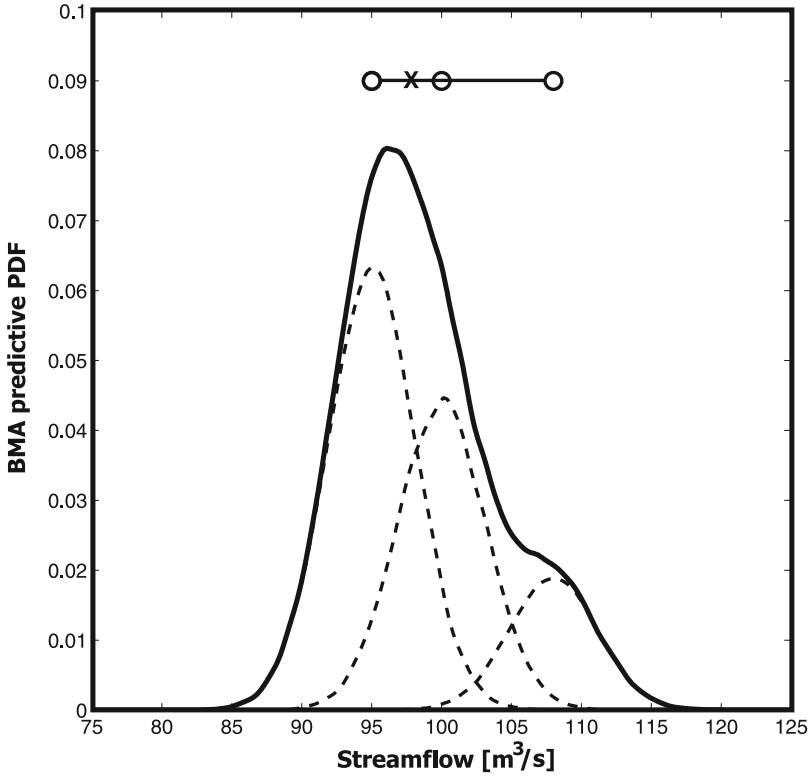

Figure 1. BMA predictive probability distribution function (solid line) and its three ensemble components (dashed lines) for a time series of synthetic streamflow data. The mean and range of individual ensemble forecasts are indicated with the circles and solid horizontal line, respectively, and the verifying observation is indicated with a cross.

various other summary statistics such as the continuous rank probability score (CRPS), root-mean-square error (RMSE), mean average error (MAE), and ignorance score that provide additional and perhaps conflicting information about the optimal parameters of the BMA model, but are deemed important for accurate probabilistic forecasts. In a subsequent paper, we present a multiobjective optimization approach to calibration of forecast ensembles using BMA [*Vrugt et al.*, 2006c].

[23] Unfortunately, no analytical solutions exist that conveniently maximize equation (11), so an alternative approach is required to determine the values of θ. *Raftery et al.* [2003, 2005] recommends the use of the expectation maximization (EM) algorithm to search iteratively for the optimal values of the weights and variances of the individual ensemble members. The EM method is relatively easy to implement, computationally efficient, and the maximization step always satisfies the constraint that the weights are positive and add up to one. Despite these advantages, convergence of the EM algorithm to the global maximum of the likelihood function cannot be guaranteed.

[24] In light of these considerations, we here implement the shuffled complex evolution Metropolis (SCEM-UA) algorithm to estimate the BMA weights and variances. The SCEM-UA algorithm is a general purpose optimization algorithm that uses adaptive Markov chain Monte Carlo (MC^2) sampling to provide an efficient search of the parameter space. The method uses a predefined number of different Markov chains to independently explore the search space. These chains communicate with each other through an external population of points, which are used to continuously update the size and shape of the proposal distribution in each chain. The MC^2 evolution is repeated until the R statistic of *Gelman and Rubin* [1992] indicates convergence to a stationary posterior distribution. To satisfy the constraint that the weights are positive and sum up to one, the feasible parameter space for each of the weights is between 0 and 1, and each SCEM-UA generated combination of weights is first divided by its respective sum before evaluating the BMA model and computing the log likelihood defined in equation (11). A detailed performance comparison of the EM algorithm and SCEM-UA algorithm for estimating the BMA weights and variances is presented by J. A. Vrugt et al. (Calibration of forecast ensembles using adaptive Markov chain Monte Carlo sampling, submitted to *Monthly Weather Review*, 2006) and so will not be repeated here.

[25] To further illustrate the BMA approach, consider Figure 1, which shows the BMA-predicted PDF (thick solid line) for three different conceptual watershed models, using a single synthetically generated streamflow data point. The three individual model forecasts and their optimized conditional distributions (thin dashed lines) are also included. Note that the BMA-predicted PDF is bimodal and broader than any of the individual conditional PDFs, reflecting the fact that there are contrasting model forecasts that disagree with one another. Table 1 summarizes the BMA weights and variances for each of the three individual watershed models for this synthetic example. As was illustrated in Figure 1, there is significant disagreement among the ensemble mem-

Table 1. Individual Model Forecasts, BMA-Derived Weights and Variances, and Verifying Streamflow Observation for the Illustrative Synthetic Example

	Model 1	Model 2	Model 3
Forecast	95.00	100.00	108.00
BMA weight	0.50	0.35	0.15
BMA variance	10.00	10.00	10.00
Observation		98.00	

bers. The highest BMA weight is for model 1, which also exhibits the lowest quadratic forecast errors over the training period (not shown in Table 1).

3. Case Study

[26] To conduct a case study comparing the EnKF and BMA methods for ensemble streamflow forecasting, a set of rainfall-runoff models was selected that provide different conceptualizations of the watershed under investigation, and generate different calibration responses, to result in a sufficient forecast spread. This will be discussed later on. These models are described in section 3.1. Then an existing historical streamflow data was selected for the purpose of comparing the techniques. This work is described in section 3.2 and discussed in detail thereafter.

3.1. Hydrologic Models Used

[27] The premise of the BMA approach is that the use of multiple models is better than the usual hydrologic practice of adopting a single model. However, working with many models would render the BMA approach impractical [*Neuman*, 2003]. Therefore a suggestion made by *Jefferys and Berger* [1992] is adopted here, whereby only a relatively small set of the most parsimonious models are considered which appear a priori to be hydrologically most plausible and supported by the available data. The conceptual watershed models that we implement in the current study are ABC (3) [*Fiering*, 1967; *Kuczera and Parent*, 1998], GR4J (4) [*Perrin et al.*, 2003], HYMOD (5) [*Boyle et al.*, 2001; *Vrugt et al.*, 2002, 2005], TOPMO (8) [*Oudin et al.*, 2005], AWBM (8) [*Boughton*, 1993; *Marshall et al.*, 2005], NAM (9) [*Nielsen and Hansen*, 1973], HBV (9) [*Bergström*, 1995], and SAC-SMA (13) [*Burnash et al.*, 1973]. These eight models are listed in order of increasing complexity, and the number of user-specified parameters is indicated in parentheses. Inputs to the models include mean areal precipitation (MAP), and potential evapotranspiration (PET). The output is estimated channel streamflow. The parameters of the individual watershed models were optimized using the SCE-UA algorithm, using a classical least squares objective function and 8 years (WY 1953–1960) of historical streamflow data. A detailed description of the different models, including a discussion of their calibration parameters, appears in the cited references.

3.2. Hydrologic Data Used

[28] We apply the EnKF and BMA methods using historical data from the Leaf River watershed (1950 km^2) located north of Collins, Mississippi. For this watershed, 36 years of historical 6-hourly MAP and daily streamflow and PET data are available. Previous studies have indicated that a calibration data set of approximately 8 to 11 years of data, representing a range of hydrologic phenomena (wet, medium and dry years), is desirable to achieve deterministic model calibrations that are consistent and generate good verification and forecasting performance [e.g., *Yapo et al.*, 1996; *Vrugt et al.*, 2006b]. In this study, eight years (WY 1953–1960) of data are used for model calibration/training, and the remaining 28 years on record (WY 1961–1988) were subsequently used to evaluate the forecast performance.

3.3. Implementation of EnKF and BMA

[29] To evaluate the forecast accuracy and associated prediction uncertainty bounds using the EnKF, the least squares estimate derived after SCE-UA calibration, was used to sequentially run the filter for each model over the 28 year evaluation period. The measurement error ε_t in equation (3) was derived using a nonparametric time-differencing approach [*Vrugt et al.*, 2005, 2006a]. In each model, input and model structural errors are treated jointly as a combined stochastic forcing term, hereafter simply called model error. The size of this error is estimated as a function of flow level q_t in equation (1) during the training period each time the EnKF derived ensemble streamflow forecast is evaluated against the corresponding observation. We assume the model error of the individual ensembles $q_{i,t}$ to be normally distributed and centered on the mean ensemble forecast:

$$q_{i,t} \sim N\left(\overline{y_t}, \sqrt{\frac{1}{T}\sum_{i=1}^{T} z_{i,t}^2}\right) - y_{i,t} \tag{12}$$

where $\overline{y_t}$ represents the EnKF-derived mean ensemble forecast, and T denotes the size of the ensemble. The advantage of this approach is that we can reliably estimate the size of this joint stochastic error term as function of flow level. This relationship between flow level and model error, derived during the training period, was subsequently used to evaluate the performance of the EnKF during the evaluation period.

[30] For BMA, an ensemble of streamflow forecasts was first produced by evaluating the SCE-UA estimates of the parameters over the calibration and evaluation period. Because the individual hydrologic models were calibrated first using the training data set and the SCE-UA algorithm, a linear bias correction of the individual ensemble members was deemed unnecessary prior to optimization of the BMA weights and variance. Moreover, our analyses including a linear bias correction term have demonstrated to somewhat reduce the bias but to result in similar values for the other performance measures, such as the RMSE, MAE, ignorance score, and coverage. Actually, in conjunction with a sliding window training approach, as discussed in section 5.3, a linear bias correction was shown to destabilize the results. So, the values for a_k and b_k, $k = 1, \ldots, K$ in equation (8) were set to 0 and 1 respectively.

[31] A (global) linear bias correction, as suggested by [*Raftery et al.*, 2003, 2005], is too simple to be useful in hydrologic modeling. Such an approach is essentially not powerful enough to remove heteroscedastic and non-Gaussian model errors from the individual forecasts in the ensemble.

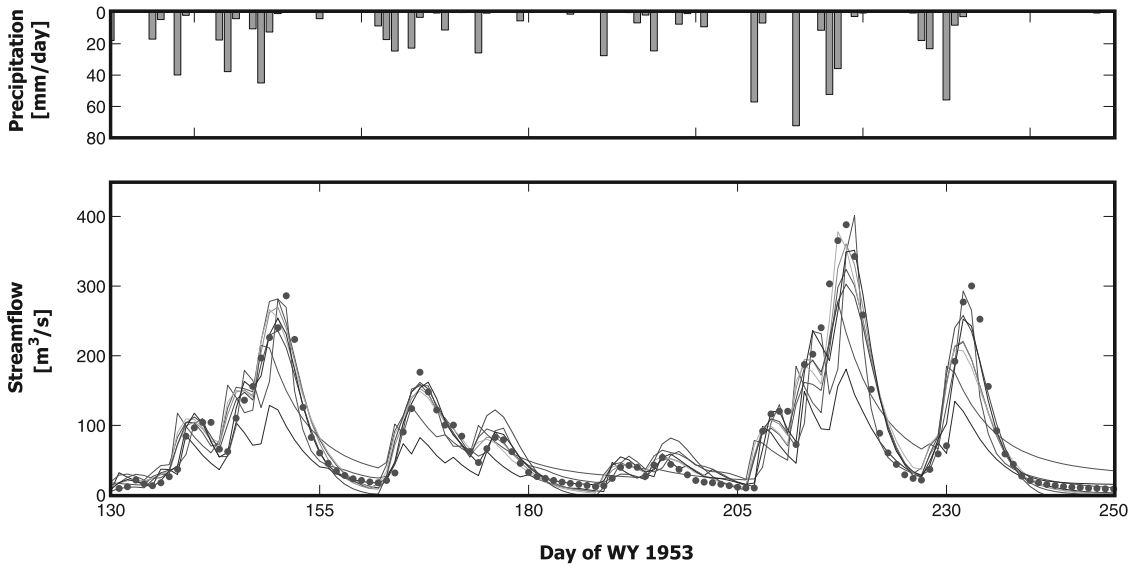

Figure 2. Streamflow predictions of the individual models of the BMA ensemble for a representative portion of the calibration period. The circles represent the verifying observations.

Instead, what is needed to improve the BMA method is a more sophisticated and probably (local) nonlinear bias correction method, perhaps using information from modeling errors in the immediate past history. Construction of better bias correction methods is beyond the scope of the current paper, and needs to be investigated in future work.

[32] To obtain some insights on the probabilistic properties of the BMA ensemble, consider Figure 2, which presents the deterministic forecasts of the individual, calibrated models of the ensemble for a representative period between 6 June and 30 September 1953. Note that the spread of the ensemble is sufficient and generally brackets the observations (the circles). The different calibrated models appear to provide different forecasts. This is a desirable characteristic and prerequisite for accurate ensemble streamflow forecasting with the BMA method.

4. Results and Discussion

[33] The results derived with the EnKF and BMA are summarized in Tables 2–4 and Figures 3–6 and will be discussed here. To streamline the discussion, the BMA-derived forecast statistics over the calibration and evaluation period are presented first, with results benchmarked against the performance of the individual calibrated watershed models that comprise the ensemble. Then, the BMA results are compared to those obtained using the EnKF, considering both the accuracy and precision of the methods. The accuracy of both methods is evaluated using summary statistics of the 1-day-ahead mean ensemble streamflow forecasts during the calibration and evaluation period. For precision, the size of the prediction uncertainty output ranges of the ensemble streamflow forecasts of the two methods are compared.

4.1. Model Accuracy: Individual Watershed Models and BMA

[34] Table 2 presents summary statistics of the 1-day-ahead streamflow forecasts for the 8-year calibration (WY 1953–1960) and the 28-year evaluation (WY 1961–1988) period for the eight individual conceptual watershed models considered in this study. The forecast statistics of the BMA predictive model and associated weights and variances for the entire calibration period are also listed. The weights

Table 2. Summary Statistics (RMSE, CORR, and BIAS) of the 1-day-Ahead Streamflow Forecasts Using the Individual Calibrated Models and BMA Predictive Model for the Calibration and Evaluation Period[a]

	Calibration (WY 1953–1960)			Evaluation (WY 1961–1988)			BMA	
Model	RMSE	CORR	BIAS, %	RMSE	CORR	BIAS, %	Weight	Variance
ABC	31.67	0.70	15.59	50.24	0.75	−4.51	0.02	120.70
GR4J	19.21	0.90	7.51	25.18	0.93	9.42	0.15	120.70
HYMOD	19.03	0.90	−0.38	28.63	0.91	2.22	0.15	120.70
TOPMO	17.68	0.92	−0.59	27.39	0.92	2.18	0.08	120.70
AWBM	26.31	0.80	6.37	42.04	0.80	6.79	0.03	120.70
NAM	20.22	0.89	−4.10	32.77	0.88	2.91	0.04	120.70
HBV	19.44	0.90	−0.04	31.23	0.90	7.42	0.04	120.70
SAC-SMA	16.45	0.93	6.69	21.89	0.95	12.12	0.49	120.70
BMA	16.24	0.93	4.64	22.29	0.95	8.52		

[a]The BMA weights and variance for the individual watershed model are also listed.

Table 3. Summary Statistics (RMSE, CORR, BIAS) of the 1-day-Ahead Streamflow Forecasts Using an Ensemble Kalman Filter Approach of the Watershed Models Retained in the Analyses[a]

	Calibration (WY 1953–1960)			Evaluation (WY 1961–1988)		
Model	RMSE	CORR	BIAS, %	RMSE	CORR	BIAS, %
HYMOD	13.60	0.95	1.37	21.94	0.95	2.08
TOPMO	12.98	0.96	0.31	21.32	0.95	1.14
NAM	15.57	0.94	10.30	26.47	0.92	8.12
HBV	13.62	0.95	−0.29	21.97	0.95	2.28
SAC-SMA	11.82	0.96	−0.47	17.02	0.97	0.26
BMA	16.24	0.93	4.64	22.29	0.95	8.52

[a]The filter performance was evaluated using the maximum likelihood parameter estimates derived from the SCE-UA calibration. The statistical results of the BMA method are also included (from Table 1).

reflect the ensemble member's overall performance over the training period, relative to the other members.

[35] The results in Table 2 suggest that it is not possible to select a single "best" watershed model that minimizes the RMSE and %BIAS of the forecast error, while simultaneously also maximizing the correlation (CORR) between the simulated and measured time series of streamflow. In this analysis, the Sacramento Soil Moisture Accounting Model (SAC-SMA) has the most consistent performance, with the lowest quadratic forecast error during the calibration and evaluation period. However, the SAC-SMA model forecasts exhibit considerable bias.

[36] Second, note that the theoretical benefit of using multimodel averaging is not realized in this instance: the BMA deterministic forecast given by equation (9) has an average quadratic forecast error that is of similar magnitude than the best performing model (SAC-SMA) in the ensemble. This is true for both the calibration and evaluation period, suggesting that the BMA approach does not necessarily improve predictive capabilities. This finding is consistent with results from a previous study [*Georgekakos et al.*, 2004], which showed that the forecast error of the mean prediction of a multimodel ensemble was of similar magnitude as the forecast error of the best model in the ensemble. As indicated previously, a linear bias correction would have resulted in similar values for the RMSE of the BMA model. It is possible that the hoped for gain in performance improvement with a multimodel ensemble approach might have been realized had we implemented a different form for the BMA conditional predictive distribution (rather than normal), and/or allowed for dynamic updating of the BMA weights and variance using a sliding window training approach. We test these alternatives in section 5.1.

[37] Another characteristic of the BMA results is that the rank order of the weights does not completely track the inverse rank order of the forecast RMSE values of the individual models. For instance, note that the TOPMO model is ranked number 2 in RMSE forecast performance during the calibration period, but receives very little weight in the BMA predictive model. In contrast, the GR4J and HYMOD model exhibit poorer predictive capabilities in the training period, but receive more weight in the computation of the BMA predictive PDF. This behavior demonstrates that cross correlations between individual forecasts in the streamflow ensemble influence the derived weights of the individual members. This result is probably due to fundamental similarities of the models, each being a lumped parameter model involving a network of connected storage volumes.

[38] Finally, note that the BMA-derived variance of the individual watershed models listed in Table 2 expresses relatively little uncertainty around the individual model forecasts, much smaller than the average one-step-ahead quadratic forecast error. Seemingly, the fitted BMA variance has little hydrologic meaning, and can only be meaningfully interpreted in conjunction with the ensemble spread. It might be possible to further improve the BMA method by explicitly incorporating information about the size of the model and measurement errors into the fitting of the mixture model. Section 5 addresses this issue more fully.

[39] To simplify the subsequent analysis, we discarded the ABC, AWBM and GR4J models, and carried the HYMOD, TOPMO, NAM, HBV and SAC-SMA conceptual

Table 4. Probabilistic Properties of the Ensemble Streamflow Forecasts Derived With the EnKF and BMA Methods for Different Flow Levels for the 8-year Calibration and 28-year Evaluation Period[a]

	Calibration (WY 1953–1960)					Evaluation (WY 1961–1988)				
	0–10	10–50	50–200	200 →	All	0–10	10–50	50–200	200 →	All
HYMOD	97.31	96.71	96.43	97.50	97.02	97.07	96.80	96.59	97.71	96.92
TOPMO	98.03	96.63	96.44	97.47	97.41	97.80	96.84	96.58	97.52	97.29
NAM	98.70	96.73	96.49	97.44	97.82	98.59	96.85	96.77	98.17	97.72
HBV	97.97	97.09	96.80	97.62	97.56	97.11	97.33	96.66	97.44	97.13
SAC-SMA	96.20	96.08	96.25	97.50	96.19	96.19	96.18	96.30	97.17	96.23
BMA	99.88	96.94	82.67	35.29	96.17	99.98	94.79	83.43	58.71	94.84

[a]The values represent the percentage of streamflow observations contained in the 95% confidence interval of the EnKF- and BMA-forecasted ensembles. For the EnKF results, statistics are provided for each of the watershed models retained in the analysis.

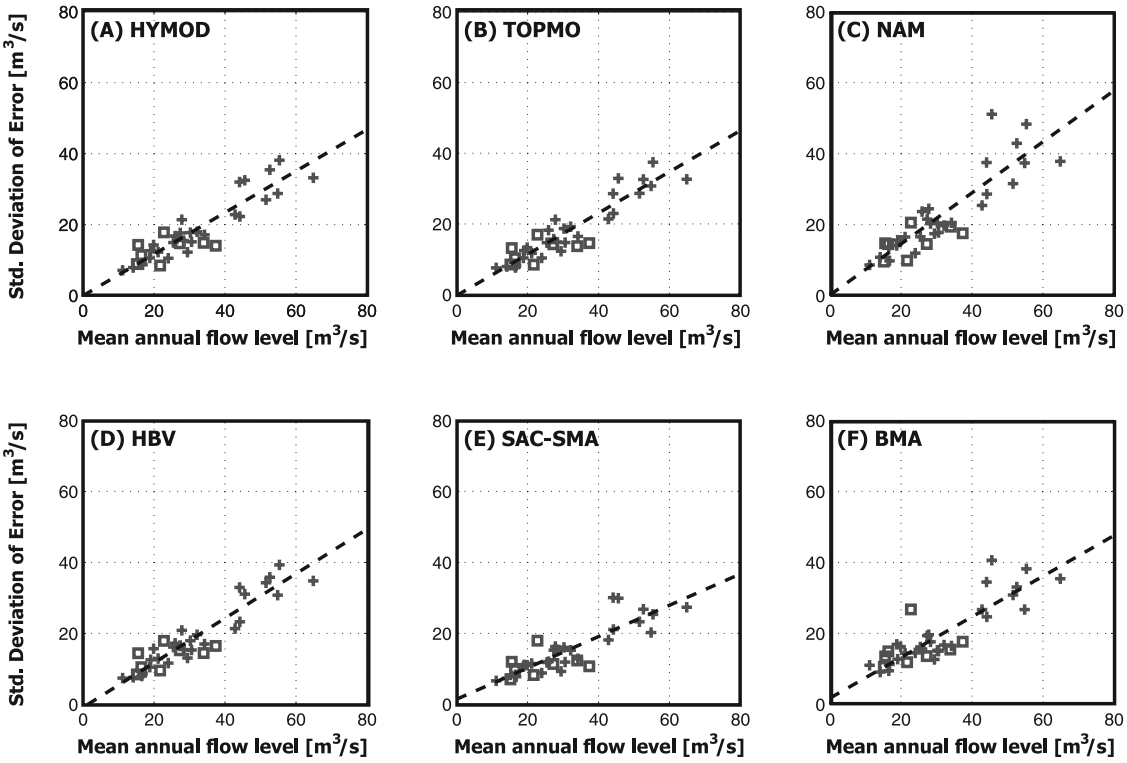

Figure 3. Annual mean standard deviation of the 1-day-ahead streamflow forecast errors as function of mean annual flow level. EnKF results for the (a) HYMOD, (b) TOPMO, (c) NAM, (d), HBV, and (e) SAC-SMA conceptual watershed models and (f) results for the BMA predictive model. Calibration years are plotted with squares, and evaluation years are plotted using pluses.

watershed models forward in the EnKF studies presented below.

4.2. Model Accuracy: Mean Ensemble Forecast of EnKF

[40] Table 3 presents summary statistics of the EnKF-based, 1-day-ahead streamflow forecasts for the calibration and evaluation period for the five remaining conceptual watershed models. These results indicate that the EnKF method provides consistently better values of the RMSE, correlation (CORR), and bias (BIAS) statistics than the multimodel BMA approach, during both the calibration and evaluation period. When state updating is employed, only the NAM model fails to outperform the BMA method.

[41] Another indication of this result is presented in Figure 3. For each model, the annual standard deviation of the EnKF-derived mean ensemble 1-day-ahead streamflow forecast errors for each of the individual watershed models are plotted against the mean annual flow level. For comparison, Figure 3f presents the same results for the BMA predictive model, using the weights and variances listed in Table 2. In all cases, the symbols appear to fall on a straight line, indicating that all the different watershed models are well optimized and that the calibration and evaluation performance are mutually consistent. However, with the exception of the NAM model, the EnKF-derived points for the individual watershed models are less scattered and closer to the origin, and the slopes of the regression lines are smaller. This indicates that the EnKF forecasts exhibit more reliable behavior than the BMA-derived predictions.

[42] This conclusion is further confirmed in Figure 4, which presents a statistical check of the autocorrelation functions of the forecasted residuals for the EnKF and BMA methods. Autocorrelation is a measure of the serial dependency of the error residuals between model predictions and corresponding data. Ideally, this measure should be centered on zero at all lags, implying a lack of systematic errors. The EnKF-derived forecast errors are essentially white for all of the watershed models, indicating that most of the bias in the 1-day-ahead predictions is removed by the recursive state adjustments. In contrast, the BMA forecasts exhibit significant autocorrelation at the first lag, suggesting that the method has more difficulty accurately forecasting the 1-day-ahead streamflow observations.

[43] Considered in total, this comparison suggests that the EnKF method yields forecasts that are superior to those obtained with the BMA approach. This is probably not surprising as the EnKF has the flexibility to explicitly account for all sources of uncertainty and employs dynamic state variable updating, whereas BMA only accounts for uncertainties in the model structure. In section 5.3, we will implement the BMA approach using a sliding window approach with dynamic updating of the weights and variance of the individual watershed models in the ensemble.

4.3. Model Precision: Ensemble Spread of EnKF and BMA

[44] The results presented so far have focused on the forecast statistics of the EnKF- and BMA-derived mean ensemble prediction, without explicitly considering the spread of the ensemble forecast. In other words, accuracy

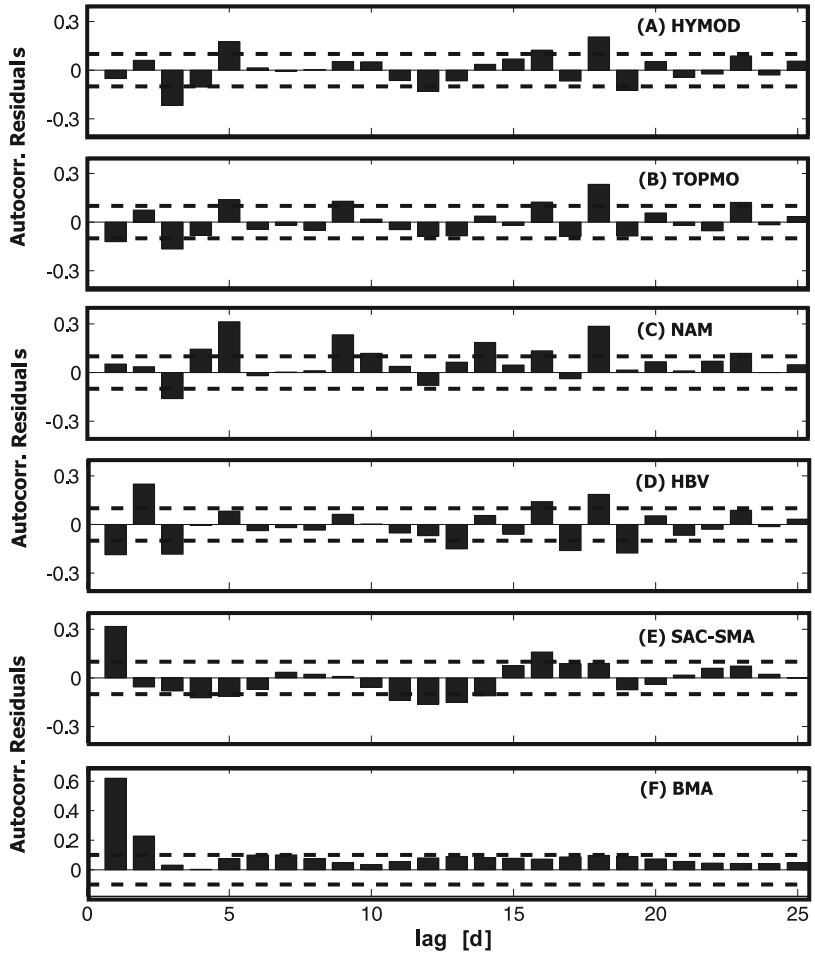

Figure 4. Autocorrelations functions of the time series of 1-day-ahead forecast errors for the combined calibration and evaluation period. EnKF-derived mean ensemble forecast for the (a) HYMOD, (b) TOPMO, (c) NAM, (d) HBV, and (e) SAC-SMA conceptual watershed models and (f) results corresponding to the BMA predictive model. The dashed lines denote the theoretical upper and lower 99% significance intervals of a time series of white residuals.

has been emphasized rather than precision. Precision can be measured from the size of the hydrograph prediction uncertainty ranges. Ideally, the ensemble spread should be as small as possible, but consistent with observations, so that the predictive PDF is as sharp as possible. Stated differently, if the model is required to generate a probabilistic forecast at a given confidence level, say, 95%, then the predictions should encompass 95% of the observations with as small an uncertainty band as possible.

[45] Figure 5 presents a comparison of the 95% hydrograph prediction uncertainty derived with the EnKF and BMA methods for a representative period in WY 1977 for the Leaf River watershed in Mississippi. The EnKF-derived results for the simplest (Figure 5b, HYMOD) and most complex watershed models (Figure 5c, SAC-SMA) are shown, along with the ensemble prediction spread for the BMA-predictive model (Figure 5d). The total uncertainty is indicated by the shaded region, and the prediction corresponding to the mean ensemble forecast is indicated with the black line. The box plots denote the median, lower and upper quartile values of the confidence intervals for the streamflow data.

[46] Figures 5b and 5c show that the EnKF-computed spread of the ensemble streamflow forecasts generally brackets the observations, but is quite large, especially in response to high-flow events. This result provides strong support for the claim by *Sorooshian and Dracup* [1980] that streamflow data exhibit nonhomogeneous (heteroscedastic) errors. Furthermore, the prediction uncertainty ranges for the 13-parameter SAC-SMA model are significantly smaller than the spread of the forecast ensemble obtained with the more parsimonious five-parameter HYMOD model. This is true for all flow levels, indicating that the SAC-SMA model structure is more capable of representing the rainfall-runoff transformation for the humid Leaf River watershed. This conclusion agrees with the earlier interpretation of the SCE-UA derived summary statistics of the 1-day-ahead forecast (Tables 2 and 3) of the individual calibrated watershed models. Therefore the size of the uncertainty bounds is directly controlled by the size of the model error, and will generally be smaller for more complex watershed models.

[47] Regarding the BMA results (Figure 5d), note that the predicted time evolution of the spread of the ensemble is quite different than the EnKF-derived results. It is evident that the BMA method is placing more confidence in the

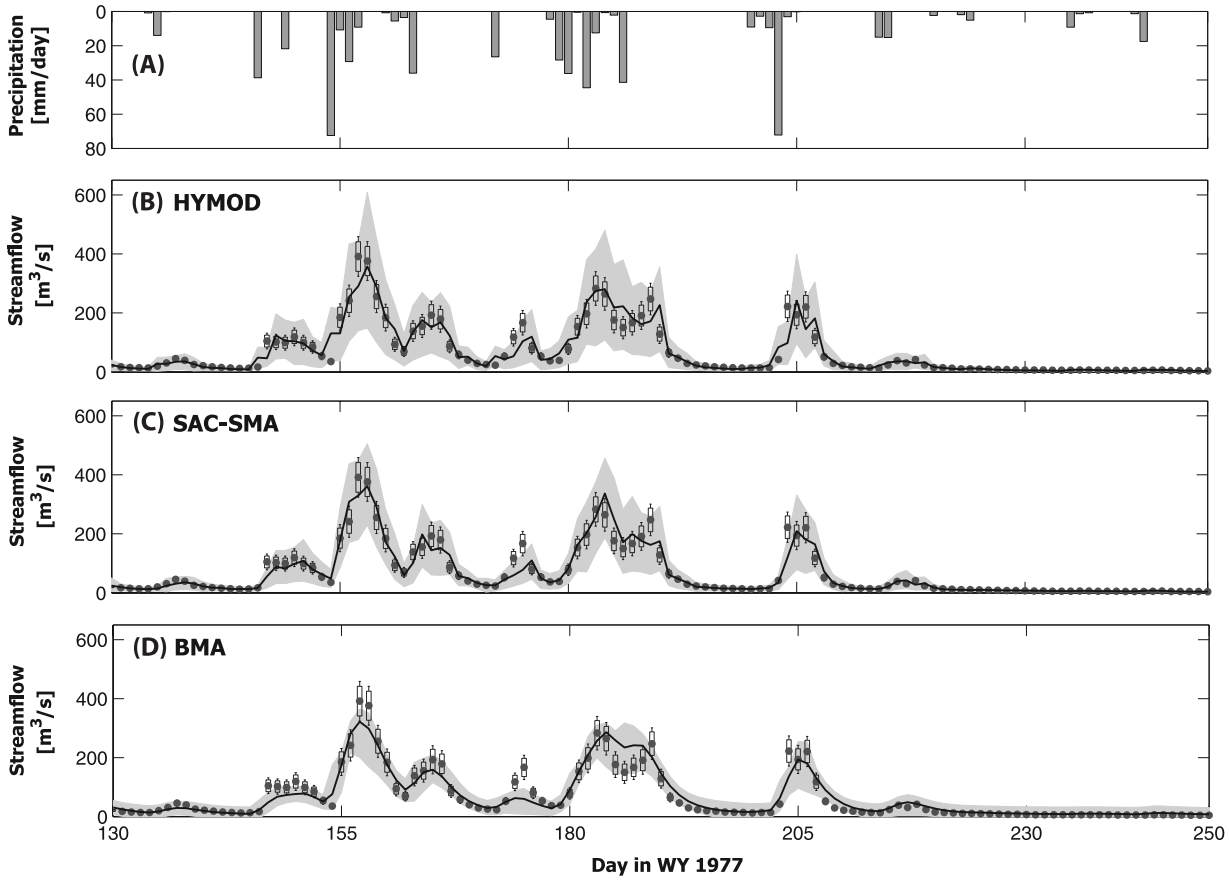

Figure 5. (a) The 95% streamflow hydrograph prediction uncertainty ranges for a representative portion of the evaluation year 1977. EnKF-derived results with the (b) HYMOD and (c) SAC-SMA models and (d) BMA ensemble spread. In each case, the mean ensemble prediction is indicated using a solid line, and the ensemble spread is represented by the shaded region. Streamflow measurement uncertainty is indicated using box plots.

forecasting of peak flows, at the expense of considerable uncertainty associated with low flows. This finding calls into question the realism of the EnKF-estimated streamflow uncertainty ranges, as well as the assumption of a single variance with flow level of the individual models in the BMA ensemble. Hence this approach essentially assumes homoscedasticity of errors with flow level, an assumption that seems particularly inappropriate for daily streamflow data.

[48] To test the validity of the EnKF- and BMA-forecasted ensemble spreads, consider Table 4, which lists the percentage of streamflow observations contained in the uncertainty bounds defined at the 95% confidence intervals. Both the EnKF-derived results of the individual watershed models and the multimodel BMA results are provided for both the calibration and evaluation periods. A significant departure from 95% would indicate that the predictive uncertainty is either underestimated or overestimated, and would call into question the validity of the modeling approach for performing accurate probabilistic streamflow forecasting.

[49] There is significant disparity in the performance of the EnKF and BMA method. The EnKF-derived uncertainty bounds are statistically appropriate for all considered watershed models and flow levels, suggesting that the size of the measurement and model errors, and their dependency on flow level, are consistent with the statistical properties of the streamflow observations. In general, the EnKF-computed spread of the ensemble retains about 97% of the streamflow data at the 95% confidence level. In contrast, the uncertainty ranges derived with the BMA method are clearly inconsistent at the various flow levels, degrading significantly at high flows. The method places too much confidence in the individual watershed models of the ensemble during flood events.

5. Refining the BMA Approach for Streamflow Forecasting

[50] The results so far have demonstrated that the BMA method needs to be refined to be useful for flood forecasting. Here we briefly explore three possible implementation improvements. The results of these analyses are presented in Table 5 and Figures 6 and 7. In Figure 6, the BMA-derived 95% hydrograph prediction uncertainty ranges are indicated by the lightly shaded region, while the prediction of the mean ensemble forecast is indicated with the solid line.

5.1. Normal BMA Predictive PDF With Heteroscedastic Variance

[51] The classical BMA method implements a single variance of the conditional PDFs of the individual forecasts

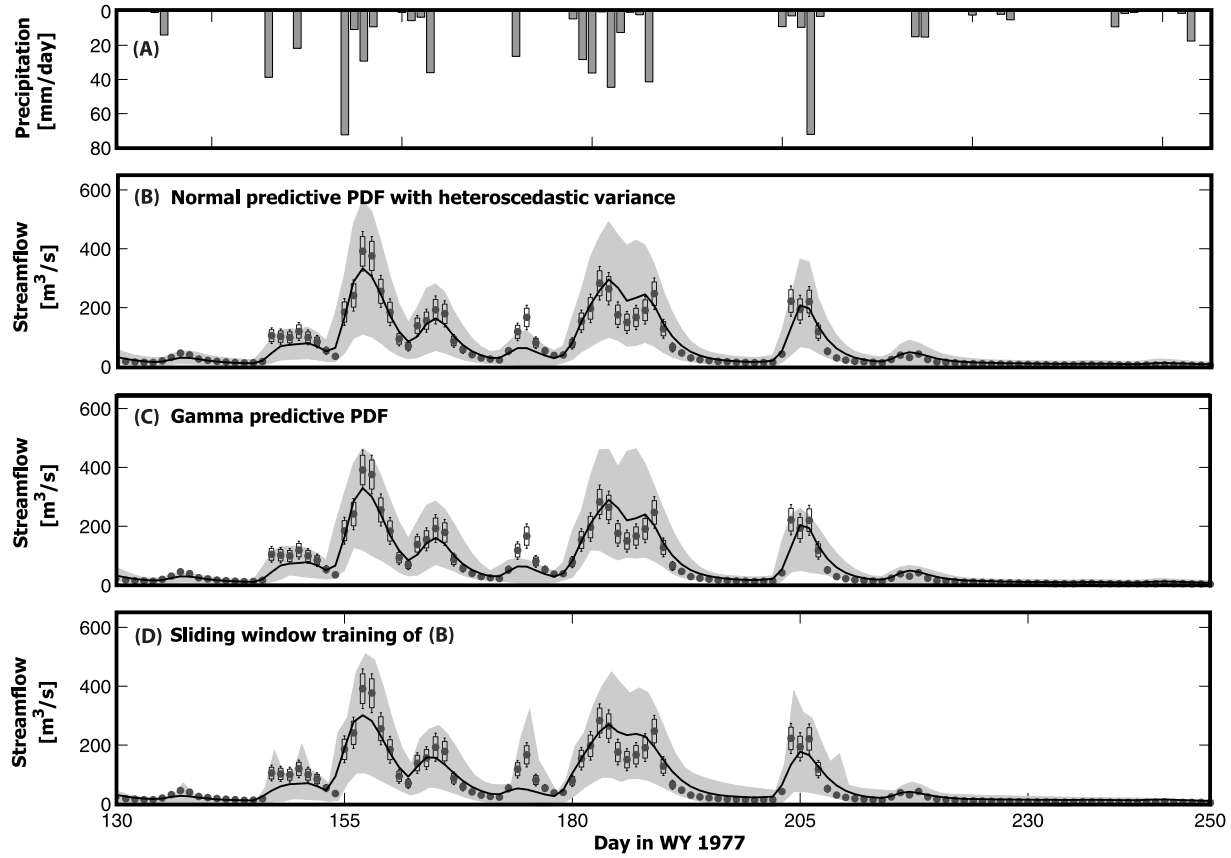

Figure 6. (a) Time series plots of BMA-derived 95% uncertainty prediction intervals of streamflow for a representative portion of the 1977 evaluation year. (b) Normal model-specific distribution in BMA and heteroscedastic variance dependent on flow level. (c) Gamma distribution for BMA predictive distribution. (d) Sliding window training approach using normal predictive distribution with nonconstant (heteroscedastic) BMA variance. The mean ensemble prediction is indicated using a solid line, and the ensemble spread is represented by the shaded region. Streamflow measurement uncertainty is indicated using box plots.

in the ensemble, irrespective of flow level. As discussed earlier, and demonstrated in the literature, this assumption seems particularly inappropriate for daily streamflow data. In the first refinement to the BMA method, we allow for heteroscedasticity of streamflow errors by implementing a linear dependency of the variance in equation (8) with flow level:

$$\sigma_k^2 = b \cdot f_k \tag{13}$$

where b represents the slope relating the original forecast to the BMA variance. The values for w_k, $k = 1, \ldots, K$; and b were estimated by maximum likelihood using the SCEM-UA algorithm and the 8-year (WY 1953–1960) training ensemble. Note, that when using different BMA variances for the individual models in the ensemble (as is done here), we need to modify the last term in equation (10) to be able to compute the variance of the BMA deterministic forecast. The term σ^2 needs to be replaced with $\sum_{k=1}^{K} w_k \sigma_k^2$ to reflect a weighted combination of the individual optimized σ_k^2s.

Table 5. Summary Statistics of the BMA Predictive PDF for the 28-year Evaluation Period Using Three Different Situations[a]

Statistic	Unit	Previous	Situation A	Situation B	Situation C
RMSE	[m^3/s]	22.29	21.23	21.69	22.32
MAE	[m^3/s]	9.79	9.68	9.76	10.04
Coverage 95%		0.051	0.028	0.031	0.037
Average width 95%	[m^3/s]	57.57	54.57	53.32	52.50
Ignorance Score		4.22	3.25	3.32	2.77

[a]Situation A is normal predictive PDF with nonconstant variance dependent on flow level, situation B is gamma predictive PDF, and situation C is sliding training approach with normal predictive PDF and heteroscedastic BMA variance. In situations A and B the BMA weights and variance were estimated from the 8-year calibration period and not updated. The results of the BMA method with normal predictive PDF but constant variance (as discussed throughout the paper) are listed under the heading previous. For more explanation, refer to the text.

[52] Situation A in Table 5 lists summary statistics of the sharpness and mean of the BMA predictive PDF for the evaluation period. Notice that the RMSE of the BMA deterministic forecast and average width of the 95% prediction uncertainty have decreased with about 5%, yielding improved performance over the deterministic forecast of the best member (SAC-SMA) of the ensemble. Seemingly, a nonconstant variance for the individual members increases the sharpness and predictive capabili-

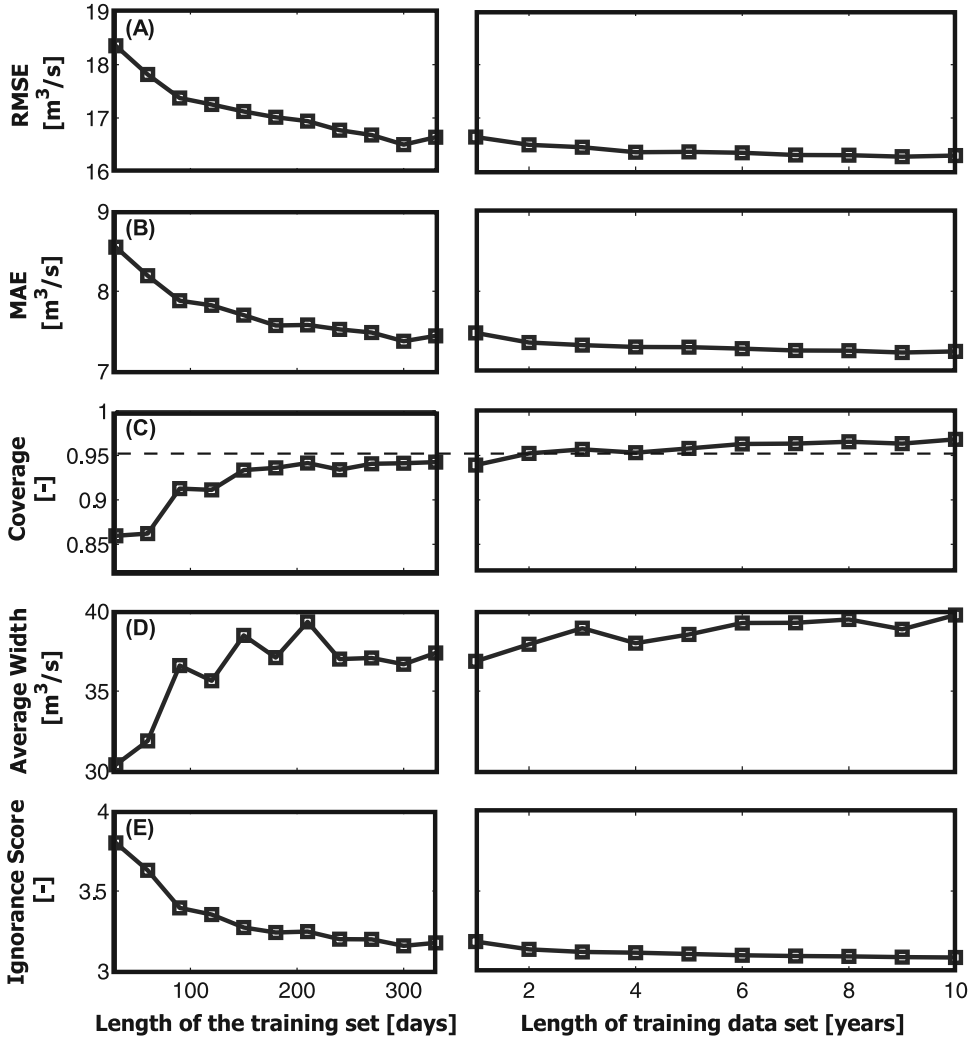

Figure 7. Comparison of training period lengths for the Leaf River streamflow data set: (a) root-mean-square error (RMSE), (b) mean average error (MAE), (c) coverage of 95% prediction intervals, (d) average width of 95% prediction interval, and (e) ignorance score.

ties of the BMA model. The ignorance score (average of the negative of the logarithm of the BMA predictive PDF evaluated at the observations; smaller scores are preferred) shows similar conclusions. Despite this improvement, the quadratic forecast errors and prediction uncertainty bounds derived with the BMA method are still significantly larger than those obtained using the EnKF implementation for the SAC-SMA model, similar to what was illustrated previously in Figure 5c.

5.2. Gamma Distribution

[53] As discussed previously, and illustrated in Figure 2, one would not expect streamflow data to be normally distributed. Thus a version of BMA that assumes normal predictive distributions may be inappropriate in the present context. A normal distribution cannot represent the highly skewed predictive distribution of streamflow. To test the BMA method with a more representative distribution, we modify the conditional PDF $g_k(\Delta|f_k)$ in equation (7) by implementing the gamma distribution, as previously illustrated in [*Sloughter et al.*, 2006] with shape parameter α and scale parameter β. The PDF is given by:

$$\Delta|f_k \sim \frac{1}{\beta\Gamma(\alpha)}\Delta^{\alpha-1}\exp(-\Delta/\beta) \tag{14}$$

for $\Delta > 0$ and, $g_k(\Delta|f_k) = 0$ for $\Delta \leq 0$. The mean of this distribution is $\mu = \alpha\beta$, and it variance is $\sigma^2 = \alpha\beta^2$.

[54] We derive $\alpha_k = \mu_k^2/\sigma_k^2$ and $\beta_k = \sigma_k^2/\mu_k$ of the gamma distribution from the original forecast, f_k, of the individual ensemble members through the following relationships:

$$\mu_k = f_k \tag{15}$$

and

$$\sigma_k^2 = b_k f_k + c_1 \tag{16}$$

As before, we estimate the maximum likelihood values for w_k, $k = 1,\ldots, K$; b_k; and c_1 using the SCEM-UA algorithm and the 8-year training ensemble.

[55] Situation B in Table 5 presents the results of this analysis in terms of summary statistics of the BMA model and associated PDF over the evaluation period. Compared to our previous implementation, the performance of the BMA model has slightly deteriorated, even though the average width of the 95% prediction uncertainty interval has further decreased, with an approximate same coverage. Thus a version of BMA that assumes normal distributions with a heteroscedastic variance for the model specific predictive distributions works surprisingly well for streamflow forecasting, and the incorporation of a gamma distribution does not improve the BMA method.

5.3. Sliding Window Training Approach

[56] Up until now, the BMA weights and variance have been estimated from a fixed training set of data consisting of the first 8 years of the historical record, without updating. However, it may more productive to recursively update the BMA predictive model when new observations become available, and the hydrologic regime and relative performance of the various ensemble members changes. Following *Raftery et al.* [2005], our final implementation therefore considers a sliding window training approach using the normal distribution with nonconstant variance discussed previously in this paper. By allowing the weights and variance to vary via a recursive training approach, we hope to endow the BMA method with an ability to evolve in a manner analogous to the dynamic state variable updating employed in the EnKF method.

[57] To understand how many years of streamflow data are needed for accurate BMA training consider Figure 7, which shows the evolution of the BMA derived RMSE, MAE, average width of 95% prediction interval, and ignorance score as function of the number of hydrologic years in the training data set. The results depicted in Figure 7 are qualitatively similar to those presented by *Raftery et al.* [2005] and highlight several important observations. First, the mean absolute error (MAE) and RMSE of the BMA deterministic forecasts and ignorance score decrease sharply at first, but then remain stable with increasing length of the training period. Second, the coverage increases with increasing length of the training set, hitting the correct 95% at about 2 years, and increasing beyond that. Finally, notice that shorter training periods generally result in sharper forecasts with smaller uncertainty bounds. On the basis of these results, we select a training period of 2 years, as increases beyond this appear to have little effect on the forecast error and sharpness of the BMA predictive PDF.

[58] Because the size of our sliding window is significant, we decided not to update the BMA weights after processing each daily streamflow observation, but instead to retrain the model after processing 2 months of data. With this approach, the data changes within the window are sufficient to potentially affect the optimal values of the BMA weights and variance. The results of our analysis are presented in situation C of Table 5 and Figure 6d.

[59] Unfortunately, the use of a sliding window updating approach does not further improve the mean of the BMA predictive model. The summary statistics of the 1-day-ahead streamflow error are on the same order as those obtained using the batch approaches without recursive weight and variance updating. However, the sharpness of the BMA predictive PDF has significantly improved, with smaller uncertainty bounds, and an ignorance score is obtained that is substantially closer to zero. These uncertainty bounds appear to better characterize simulation uncertainty, as they now bracket the streamflow data around day 175. The prediction uncertainty for this particular event was poorly characterized using the two previous BMA implementations. Despite these modifications, the BMA model is still unable to approach the predictive performance of the EnKF for the best model of the ensemble (SAC-SMA).

[60] In summary, it seems that for the class of models considered in this paper, sequential data assimilation using an individual model offers important advantages over the BMA method. The EnKF has the flexibility to explicitly account for all sources of uncertainty, and has dynamic updating built in, which improves predictive performance. However, the BMA method can generate useful probabilistic estimates of streamflow uncertainty using deterministic predictions of individually calibrated models. Moreover, we posit that BMA has significant advantages over sequential data assimilation methods in cases in which the covariance among observations and model states is low.

6. Conclusions

[61] The purpose of the present study was to compare ensemble Kalman filtering (EnKF), based on a single model, and the more recently proposed ensemble Bayesian model averaging (BMA) for probabilistic streamflow forecasting. The results show that in its current implementation, BMA cannot achieve a performance matching that of the EnKF method for the best model in the ensemble, either in terms of forecast accuracy or precision.

[62] While this conclusion favors the use of sequential data assimilation for probabilistic ensemble streamflow forecasting, there are certainly numerous examples in hydrology for which the BMA approach is worth pursuing further. The flexibility of the BMA approach for addressing conceptual model uncertainty is a significant potential advantage. There are no restrictions on the differences in structure, conceptualization, physics, or numerical formulation that can be accommodated within the BMA framework. The comparisons presented herein merely suggest that, for streamflow forecasting, additional improvements in the BMA method must be made to make it competitive with the EnKF approach. With regard to computational requirements, both the EnKF and BMA method are affordable, and their implementation is relatively easy. So both methods are suited to be included into existing operational flood forecasting codes and software.

[63] Despite this, neither BMA nor EnkF are complete in themselves. Both rely on normal or gamma probability distributions that might work well in practice, but seem inappropriate to accurately characterize input and model uncertainty. It is clear that more research is needed on the development of appropriate error structures for errors in rainfall, discharge, evaporation, model structure and parameters.

[64] Finally, regarding the steps required to improve the BMA method, we note that the skill and performance of the individual members of the ensemble ultimately determine

the success of multimodel forecasting approaches. Research on the methods that can reliably generate a multimodel ensemble is needed and ongoing. Another avenue of research worth pursuing, and suggested by one of the reviewers, is to develop a version of BMA in which the predictive distribution for each individual model at each observation is given by the EnKF, rather than by a normal or gamma distribution.

[65] **Acknowledgments.** The first author is supported by the LANL Director's Funded Postdoctoral program. We are grateful for help from Ludovic Oudin for implementing the GR4J and TOPMO conceptual watershed models in MATLAB and for receiving the source codes of the AWBM and HBV watershed models from Lucy Marshall and Jan Seibert. Computer support provided by the SARA center for parallel computing at the University of Amsterdam is greatly acknowledged. The many suggestions and comments provided by Martyn Clark and two other anonymous reviewers have greatly improved the paper.

References

Ajami, N. K., Q. Duan, X. Gao, and S. Sorooshian (2005), Multi-model combination techniques for hydrologic forecasting: Application to distributed model intercomparison project results, *J. Hydrometeorol.*, *7*(4), 755–768.

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

Bergström, S. (1995), The HBV model, in *Computer Models of Watershed Hydrology*, edited by V. P. Singh, pp. 443–476, Water Resour. Publ., Highlands Ranch, Colo.

Boughton, W. C. (1993), A hydrograph based model for estimating the water yield of ungauged catchments, *J. Irrig. Drain. Eng.*, *116*, 83–98.

Boyle, D. P., H. V. Gupta, S. Sorooshian, V. Koren, Z. Zhang, and M. Smith (2001), Toward improved streamflow forecast: Value of semidistributed modeling, *Water Resour. Res.*, *37*(11), 2749–2759.

Burnash, R. J. E., R. L. Ferral, and R. A. McQuire (1973), A generalized streamflow simulation system, report, Joint Fed. State River Forecast. Cent., Sacramento, Calif.

Clark, M. P., and A. W. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *J. Hydrometeorol.*, *7*(1), 3–22.

Clyde, M. A. (1999), Bayesian model averaging and model search strategies, in *Bayesian Statistics*, vol. 6, edited by J. M. Bernardo et al., pp. 157–185, Oxford Univ. Press, New York.

Evensen, G. (1994), Sequential data assimilation with a nonlinear quasi-geostrophic model using Monte Carlo methods to forecast error statistics, *J. Geophys. Res.*, *99*, 10,143–10,162.

Fiering, M. B. (1967), *Streamflow Synthetic*, Macmillan, New York.

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

Georgakakos, H., and J. A. Sperflage (1995), Hydrologic forecast system—HFS: A user's manual, *HRC Tech. Note 1*, 17 pp., Hydrol. Res. Cent., San Diego, Calif.

Georgakakos, H., H. Rajaram, and S. G. Li (1988), On improved operational hydrologic forecasting of streamflows, *Rep. 325*, 162 pp., Iowa Inst. of Hydraul. Res., Iowa City.

George, E. I., and R. E. McCulloch (1993), Variable selection via Gibbs sampling, *J. Am. Stat. Assoc.*, *88*(423), 881–889.

Georgekakos, K. P., D. J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004), Characterizing streamflow simulation uncertainty through multi-model ensembles, *J. Hydrol.*, *298*(1–4), 222–241.

Gordon, N. D., D. Salmond, and A. F. M. Smith (1993), Novel approach to nonlinear and non-Gaussian Bayesian state estimation, *IEE Proc., Part F*, *140*, 107–113.

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian model averaging: A tutorial, *Stat. Sci.*, *14*(4), 382–417.

Jefferys, W., and J. Berger (1992), Ockham's razor and Bayesian analysis, *Am. Sci.*, *80*, 64–72.

Kalman, R. (1960), New approach to linear filtering and prediction problems, *J. Basic Eng.*, *82D*, 35–45.

Kitanidis, P. K., and R. L. Bras (1980a), Real time forecasting with a conceptual hydrologic model: 1. Analysis of uncertainty, *Water Resour. Res.*, *16*(6), 1025–1033.

Kitanidis, P. K., and R. L. Bras (1980b), Real time forecasting with a conceptual hydrologic model: 2. Applications and results, *Water Resour. Res.*, *16*(6), 1034–1044.

Kuczera, G., and E. Parent (1998), Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm, *J. Hydrol.*, *211*, 69–85.

Marshall, L., D. Nott, and A. Sharma (2005), Hydrological model selection: A Bayesian alternative, *Water Resour. Res.*, *41*, W10422, doi:10.1029/2004WR003719.

Moradkhani, H., S. Sorooshian, H. V. Gupta, and P. R. Hauser (2005a), Dual state-parameter estimation of hydrological models using ensemble Kalman filter, *Adv. Water Resour.*, *28*, 135–147.

Moradkhani, H., K. Hsu, H. V. Gupta, and S. Sorooshian (2005b), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using particle filter, *Water Resour. Res.*, *41*, W05012, doi:10.1029/2004WR003604.

Neuman, S. P. (2003), Maximum likelihood Bayesian averaging of uncertain model predictions, *Stochastic Environ. Res. Risk Assess.*, *17*, 291–305, doi:10.1007/800477-003-0151-7.

Nielsen, S. A., and E. Hansen (1973), Numerical simulation of the rainfall runoff processes on a daily basis, *Nord. Hydrol.*, *4*, 171–190.

Oudin, L., C. Perrin, T. Mathevet, V. Andréassian, and C. Michel (2005), Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models, *J. Hydrol.*, *320*, 62–83, doi:10.1016/j.jhydrol.2005.07.016.

Perrin, C., C. Michel, and V. Andréassian (2003), Improvement of a parsimonious model for streamflow simulation, *J. Hydrol.*, *279*, 275–289.

Raftery, A. E., and Y. Zheng (2003), Long-run performance of Bayesian model averaging, *J. Am. Stat. Assoc.*, *98*, 931–938.

Raftery, A. E., D. Madigan, and J. A. Hoeting (1997), Bayesian model averaging for linear regression models, *J. Am. Stat. Assoc.*, *92*, 179–191.

Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski (2003), Using Bayesian model averaging to calibrate forecast ensembles, *Tech. Rep. 440*, Dep. of Stat., Univ. of Wash., Seattle.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*, 1155–1174.

Refsgaard, J. C. (1998), Validation and intercomparison of different updating procedures for real-time forecasting, *Nord. Hydrol.*, *28*, 65–84.

Seo, D. J., V. Koren, and N. Cajina (2003), Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting, *J. Hydrometeorol.*, *4*, 627–641.

Shamseldin, A. Y., and K. M. O'Connor (1999), A real-time combination method for the outputs of different rainfall-runoff models, *Hydrol. Sci. J.*, *44*(6), 895–912.

Shamseldin, A. Y, K. M. O'Connor, and G. C. Liang (1997), Methods for combining the outputs of different rainfall-runoff model, *J. Hydrol.*, *197*, 203–229.

Sloughter, J. M., A. E. Raftery, and T. Gneiting (2006), Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Tech. Rep. 496*, Dep. of Stat., Univ. of Wash., Seattle.

Sorooshian, S., and J. A. Dracup (1980), Stochastic parameter estimation procedures for hydrologic rainfall-runoff models: Correlated and heteroscedastic error cases, *Water Resour. Res.*, *16*(2), 430–442.

Vrugt, J. A., W. Bouten, H. V. Gupta, and S. Sorooshian (2002), Toward improved identifiability of hydrologic model parameters: The information content of experimental data, *Water Resour. Res.*, *38*(12), 1312, doi:10.1029/2001WR001118.

Vrugt, J. A., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*, W01017, doi:10.1029/2004WR003059.

Vrugt, J. A., H. V. Gupta, B.Ó. Nualláin, and W. Bouten (2006a), Real-time data assimilation for operation ensemble streamflow forecasting, *J. Hydrometeorol.*, *7*(3), 548–565, doi:10.1175/JHM504.1.

Vrugt, J. A., H. V. Gupta, S. C. Dekker, S. Sorooshian, T. Wagener, and W. Bouten (2006b), Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting Model, *J. Hydrol.*, *325*(1–4), 288–307, doi:10.1016/j.hydrol.2005.10.041.

Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Duan, and B. A. Robinson (2006c), Multi-objective calibration of forecast ensembles using Bayesian model averaging, *Geophys. Res. Lett.*, *33*, L19817, doi:10.1029/2006GL027126.

Yapo, P., H. V. Gupta, and S. Sorooshian (1996), Automatic calibration of conceptual rainfall-runoff models: Sensitivity to calibration data, *J. Hydrol.*, *181*, 23–48.

Ye, M., S. P. Neuman, and P. D. Meyer (2004), Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff, *Water Resour. Res.*, *40*, W05113, doi:10.1029/2003WR002557.

B. A. Robinson and J. A. Vrugt, Earth and Environmental Sciences Division, Los Alamos National Laboratory, MS T003, Los Alamos, NM 87545, USA. (vrugt@lanl.gov)