Portland State University

PDXScholar

Civil and Environmental Engineering Faculty
Publications and Presentations

Civil and Environmental Engineering

3-2012

Toward Reduction of Model Uncertainty: Integration of Bayesian Model Averaging and Data Assimilation

Mark A. Parrish
Portland State University

Hamid Moradkhani
Portland State University

Caleb Matthew DeChant
Portland State University

Follow this and additional works at: https://pdxscholar.library.pdx.edu/cengin_fac

Part of the [Hydraulic Engineering Commons](#)

Let us know how access to this document benefits you.

Citation Details

Parrish, M. A., H. Moradkhani, and C. M. DeChant (2012), Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, *Water Resour. Res.*, 48, W03519.

This Article is brought to you for free and open access. It has been accepted for inclusion in Civil and Environmental Engineering Faculty Publications and Presentations by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: pdxscholar@pdx.edu.

Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation

Mark A. Parrish,¹ Hamid Moradkhani,¹ and Caleb M. DeChant¹

Received 5 July 2011; revised 19 January 2012; accepted 6 February 2012; published 20 March 2012.

[1] Multimodeling in hydrologic forecasting has proved to improve upon the systematic bias and general limitations of a single model. This is typically done by establishing a new model as a linear combination or a weighted average of several models with weights on the basis of individual model performance in previous time steps. The most commonly used multimodeling method, Bayesian model averaging (BMA), assumes a fixed probability distribution around individual models' forecast in establishing the prior and uses a calibration period to determine static weights for each individual model. More recent work has focused on a sequential Bayesian model selection technique with weights that are adjusted at each time step in an attempt to accentuate the dynamics of an individual model's performance with respect to the system's response. However, these approaches still assume a fixed distribution around the individual models' forecast. A new sequential Bayesian model-averaging technique is developed incorporating a sliding window of individual model performance around the forecast. Additionally, this new technique relaxes the fixed distribution assumption in establishing the prior utilizing a particle filter data assimilation method that reflects both the performance dynamics of the models' forecasts along with their uncertainty. A comparative analysis of the different BMA strategies is performed across different rates of change in the hydrograph. Results show that methods employing the particle filter show higher probabilistic skill in high ranges of volatility but are overconfident in medium and low ranges of volatility.

Citation: Parrish, M. A., H. Moradkhani, and C. M. DeChant (2012), Toward reduction of model uncertainty: Integration of Bayesian model averaging and data assimilation, *Water Resour. Res.*, 48, W03519, doi:10.1029/2011WR011116.

1. Introduction

[2] Model simulations/predictions are subject to various uncertainties and sources of forecast errors. Uncertainties may stem from model initialization because of incomplete data coverage, observation errors, or improper data assimilation procedures. Other sources of uncertainty in prediction are associated with model input (i.e., forcing data) and imperfections of the model structure itself, mainly due to parameterization and spatiotemporal discretization of physical processes. Model structure uncertainty is associated with the assumptions reflected in model conceptualization and mathematical structure. An unfortunate truth in model development is that no matter how many resources are invested in developing a particular model, there remain conditions and situations in which the model is unsuitable to give an accurate forecast. Reliance on a single model typically overestimates the confidence and increases the statistical bias of the forecast.

[3] Model-averaging techniques look to overcome the limitations of a single model by linearly combining a number

of competing models into a single new model forecast. This method dates back to work by *Bates and Granger* [1969] who utilized model averaging in economic forecasting, and showed that a pooled forecast of competing models outperformed any single model's forecast.

[4] The early applications of model-averaging for hydrological systems resulted in a point forecast [*Shamseldin et al.*, 1997]. Techniques such as equal weight, Granger-Ramanathan averaging, and Bates-Granger averaging [*Granger and Ramanathan*, 1984, *Diks and Vrugt*, 2010] linearly combine the deterministic model outputs into another single-point deterministic forecast. An extension of these approaches was performed by *Doblas Reyes et al.* [2005] using a multiple linear regression model to compute model weights while assuming a model Gaussian distribution, which allows for a probabilistic performance valuation. Although these techniques outperform any single model's predictions, *Hoetting et al.* [1999] and *Rafterty et al.* [2005] argued that these weights are unintuitive and do not necessarily reflect the strength of a particular model's performance. Therefore, as an alternative approach to overcome these criticisms, *Hoetting et al.* [1999] considered the idea of Bayesian model averaging (BMA), a technique that weights a model by its performance and likelihood of predicting the observation, resulting in a probabilistic forecast. The model weights are all nonnegative with the total sum equaling 1, therefore acting as a probability measure of

¹Department of Civil and Environmental Engineering, Portland State University, Portland, Oregon, USA.

a model's likelihood of success. *Rafterty et al.* [2005] incorporated these techniques on an ensemble of meteorological models, while others [*Duan et al.*, 2007; *Vrugt and Robinson*, 2007; *Rojas et al.*, 2008] used these techniques on hydrological models.

[5] *Duan et al.* [2007] showed the BMA techniques outperform, or are comparable to, an individual model forecast using a variety of pointwise performance measures on a number of conceptual rainfall-runoff (CRR) models. These results were obtained by using a single training period to determine the weights of each model. Marked improvements were noted by splitting the training periods into specific flow regimes, therefore accentuating particular models strengths by making their weights larger during that flow sequence. Others [*Rafterty et al.*, 2005; *Vrugt and Robinson*, 2007] used a sliding window of training periods to estimate BMA weights. Another Bayesian formulation approach, the hierarchical mixtures of experts is performed by *Marshall et al.* [2007], which dynamically adjusts the model weights based on the watersheds predictor variables, such as system storage. Recently, the sequential Bayesian combination approach was introduced by *Hsu et al.* [2009] to act as an alternative to BMA by using sequential Bayes' law in recursively updating the posterior probability of a model likelihood function given new observations. The posterior probability acts as weights for the multimodel averaging.

[6] To apply the BMA techniques to deterministic models, some measure of the uncertainty surrounding the model forecast is required. This uncertainty is caused by a mixture of both estimates in the observations and state descriptions. Current techniques assume a fixed probability distribution function (PDF) that represents the uncertainty in the forecast. The parameters of this PDF are either estimated by error in the model prediction, with respect to the observation during a calibration period [*Hsu et al.*, 2009], or by using optimization techniques such as the expectation maximization algorithm [e.g., *Rafterty et al.*, 2005; *Ajami et al.*, 2007; *Duan et al.*, 2007]. Sequential data assimilation techniques, however, can provide a more direct way of accounting for this uncertainty, while having the potential to reduce these uncertainties.

[7] Sequential data assimilation techniques merge the observations, as they become available, with the dynamic model to update/correct the model states forecast. The ensemble Kalman filter (EnKF), the most commonly used ensemble data assimilation method in hydrometeorologic forecasting, relies on Gaussian distributions to represent the model error and observation uncertainties by randomly sampling from each distribution. Assuming that the model states are linearly correlated with the prediction, the assimilation process corrects the uncertain model states to build a posterior distribution [*Moradkhani and Sorooshian*, 2008]. Another data assimilation technique, the particle filter (PF), was developed to relax these assumptions, allowing non-Gaussian error distributions and creating a complete representation of the forecast density [e.g., *Moradkhani et al.*, 2005a, 2006; *Weerts and El Serafy*, 2006; *Matgen et al.*, 2010]. *Moradkhani et al.* [2005a] extended this approach to dual state-parameter estimation, where the sequential state and parameter estimation can be done concurrently accounting for interdependencies among the state variables and parameters during model simulation. *Leisenring and Moradkhani* [2010]

extended the application of PF to a snow water equivalent estimation. In a parallel effort, *DeChant and Moradkhani* [2011a] assimilated the brightness temperature from an AMSR-E satellite to a National Weather Service (NWS) SNOW-17 model within a particle filtering framework. The resulting snowmelt fluxes were used as forcing data for the Sacramento Soil Moisture Accounting Model (SAC-SMA) to improve the streamflow forecasting. *Montzka et al.* [2011] used the PF to update both hydraulic parameters and state variables to explore the potential of using surface soil moisture measurements from different satellite platforms in retrieving soil moisture profiles and soil hydraulic properties using the HYDRUS-1D model. More recently, *DeChant and Moradkhani* [2011b] combined the PF data assimilation with an ensemble streamflow prediction (ESP) model of National Weather Service (NWS) to more accurately and reliably characterize the initial condition uncertainty in generating the ensemble of streamflow forecast.

[8] Studies on the application of BMA strategies to CRR models to date have primarily focused on the performance over multiple years and hydrologic situations. In a more theoretical study for temperature forecast, *Weigel et al.* [2008] showed that prediction skill would only improve with multimodel averaging schemes when individual models were overconfident. Since the confidence in individual CRR models is highly correlated with precipitation pulses, this suggests that the performance of different BMA strategies applied to CRR models might also change with different stages on the hydrograph.

[9] In this paper, we conduct a comparison of several BMA strategies applied to CRR models and evaluate their forecast skills over a range of stages of the hydrograph. In section 2, a brief description of the different BMA strategies considered in this study is provided. In section 3, we outline the experimental design, including the individual models used and a brief description of the study area. In section 4, the performance measures used in our analysis are described and employed over the entire validation period followed by an analysis over different ranges of volatility in the hydrograph. Section 5 contains the summary and conclusions of our analysis.

2. Bayesian Model Averaging Strategies

[10] In section 2 we describe the BMA strategies implemented in this study. We first introduce the common methodology for all of the strategies, and then explore the various interpretations of the approach, which leads naturally to the different strategies.

2.1. General BMA Methodology

[11] Consider a quantity, y , to be forecasted, such as the magnitude of a river flow at a particular location and time. Assume we have k models, $M = [M_1, M_2, \dots, M_k]$ giving us an independent model forecast, $Y^f = [y^{f1}, y^{f2}, \dots, y^{fk}]$ for this quantity for time steps 1 through T , where $y^{fi} = [y_1^{fi}, y_2^{fi}, \dots, y_T^{fi}]$. In general, the BMA procedure seeks to compute a new forecast density as a weighted average of the competing models forecasts with weights that correspond to the comparative performance of the models over some training period of observations $Y = [y_1, y_2, \dots, y_T]$.

[12] First, the BMA methodology assumes that the model forecasts are unbiased; that is, the $E[Y - y^f] = 0$ for each model i . Although there are numerous bias-correction methods, in this paper we incorporate a linear regression of Y on $E[y^f]$. That is,

$$Y = a_i * E[y^f] + b_i. \quad (1)$$

Unique coefficients a_i and b_i for each model are determined using a least squares approximation, with the observations in the training period as the dependent variable and the forecast as the explanatory variable. These coefficients are then applied to all future model forecasts. All future references to model forecasts are assumed to be unbiased. Different application strategies for this technique, however, are later discussed in section 2.3 on the Bayesian modeling averaging with a sliding window.

[13] The forecast density for y conditioned on the models forecast, M_i , and training period of observations, Y , can be expressed according to the law of total probability as,

$$P(y_t | M_1, M_2, \dots, M_k, Y) = \sum_{i=1}^k P(y_t | M_i, Y) P(M_i | Y), \quad (2)$$

where $P(y_t | M_i, Y)$ is defined as the posterior distribution of y based only on model M_i and the training data Y . $P(M_i | Y)$ is defined as the posterior probability or the relative likelihood of model M_i being correct given the training data Y .

[14] As an illustrative example of how this process produces a multimodel forecast PDF, Figure 1 has been prepared. In this illustration, three models are considered. Figure 1A of the illustration shows the posterior distribution of y for each model. Figure 1B shows the weight defining the models relative likelihood of being the best model. The product of these weights with the distribution from Figure 1A displays the relative contribution of the models forecast to the eventual PDF. Finally, Figure 1D shows the

summation and the eventual forecast PDF (i.e., multimodel posterior distribution) for the quantity D .

[15] The various strategies explored in this paper are based on different methods for computing these posterior distributions. For example, a characteristic of the BMA methodology is that a model forecast does not necessarily need to be probabilistic. For deterministic models, this opens up the interpretation of how the posterior distribution, $P(y_t | M_i, Y)$ might be defined. Previous applications have assumed that the $P(y_t | M_i, Y) \sim g(y_t | y_t^f, \sigma_i^2)$, where σ_i^2 is somehow associated with uncertainty within an individual model and g represents a normal distribution [Duan *et al.*, 2007]. However, it is possible to relax this assumption using data assimilation techniques such as a PF, whose forecast is a distribution and can act directly as the posterior distribution of y given the past model predictions and observations.

2.2. Static Bayesian Model Averaging (BMA_Static)

[16] The first approach considered is the standard BMA method implemented by Duan *et al.* [2007]. This approach seeks to find static parameters (weights and model variances) from a fixed training period with a set of observations Y . Assuming that the conditional posterior distribution of y_t , based only on model M_i , $P(y_t | M_i, Y)$, is normally distributed and denoted by $g(y_t | y_t^f, \sigma_i^2)$, equation (2) can be written to describe the likelihood that a particular observation is predicted.

$$P(y_t | M_1, M_2, \dots, M_k, Y) = \sum_{i=1}^k g(y_t | y_t^f, \sigma_i^2) * w_i. \quad (3)$$

For numerical stability and algebraic simplicity, the log of (3) is considered:

$$\text{Log}(P[y_t | M_1, M_2, \dots, M_k, Y]) = \text{Log} \left(\sum_{i=1}^k g[y_t | y_t^f, \sigma_i^2] * w_i \right). \quad (4)$$

Figure 1. Implementation of Bayesian model averaging on three models: (A) posterior distribution of q for each model, (B) normalized likelihood of model giving the correct response, (C) model forecasts weighted by normalized likelihood, and (D) weighted forecast summed to form multimodel density.

Summing this over the training period an objective function can be maximized and written as

$$\begin{aligned} & \text{MAX} \sum_{t=1}^T \text{Log}(P[y_t | M_1, M_2, \dots, M_k, Y]) \\ & = \text{MAX} \sum_{t=1}^T \text{Log} \left(\sum_{i=1}^k (g[y_t | y_t^{f_i}, \sigma_i^2]) * w_i \right). \end{aligned} \quad (5)$$

The problem then becomes finding the optimal weights, w_i , and variance, σ_i^2 , for each model to maximize the objective function. Unfortunately, the objective function cannot be maximized analytically and requires a numerical technique to solve. A common procedure to obtain the above parameters is by using the expectation maximization (EM) algorithm [Raftery *et al.*, 2005], although other global optimization methods such as SCE-UA and its variants [Duan *et al.*, 1992; Vrugt and Robinson, 2007; Ebtehaj *et al.*, 2010] can be used. In this study we decided to use the EM procedure.

[17] The EM algorithm can be used to solve a finite mixture problem, such as the BMA methodology, by setting the optimization problem in terms of a latent variable. Following Duan *et al.* [2007], we introduce an unobserved variable $z_k(t)$ defined as

$$z_k(t) = \begin{cases} 1 & \text{best forecast at time } t \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

In other words, for each time step only one $z_k(t)$ is equal to 1 for each model $1, \dots, k$. The EM algorithm iterates between two steps until a convergence of the likelihood function is reached. First, the values for the parameters are estimated as shown in algorithm 1 allowing the calculation of an initial likelihood. The EM algorithm then estimates the value $z_k(t)$ for each time step in the expectation step as a relative comparison of the conditional likelihood $g(y_t | y_t^{f_i}, \sigma_i^2)$ using the parameters from the previous iteration. The maximization step is then executed, computing the estimate for w_k as the mean of the $z_k(t)$ for each model step. The variance σ_k is computed as a weighted average of the square differences weighted by $z_k(t)$ for each time step. Once a convergence is reached, the result is a single static weight and variance for each model.

[18] Algorithm 1. Expectation maximization algorithm: Model weights and variance are determined in an optimization framework.

Initialize:

$$\begin{aligned} \sigma_k^{20} &= \frac{1}{KT} \sum_{t=1}^T \sum_{k=1}^K (y_t - y_t^{f_k})^2, \\ w_k^0 &= 1/K, \\ l(\theta^0) &= \sum_{t=1}^T \log \left[\sum_{k=1}^K w_k^0 g(y_t | y_t^{f_k}, \sigma_k^0) \right]. \end{aligned}$$

Main:

Do $j = j + 1$.

For $t = 1:T$,

$$z_{k,t}^j = \frac{g(y_t | y_t^{f_k}, \sigma_k^{j-1})}{\sum_{k=1}^K g(y_t | y_t^{f_k}, \sigma_k^{j-1})} \quad \text{Expectation}$$

End.

$$\begin{aligned} w_k^j &= \frac{1}{T} \sum_{t=1}^T z_{k,t}^j, \\ \sigma_k^{2j} &= \frac{\sum_{t=1}^T z_{k,t}^j * (y_t - y_t^{f_k})^2}{\sum_{t=1}^T z_{k,t}^j} \quad \text{Maximization} \\ l(\theta^j) &= \sum_{t=1}^T \log \left[\sum_{k=1}^K w_k^j g(y_t | y_t^{f_k}, \sigma_k^j) \right], \end{aligned}$$

while $|l(\theta^j) - l(\theta^{j-1})| < \varepsilon$.

2.3. Bayesian Model Averaging With Sliding Window (BMA-SW)

[19] A major critique of the static BMA technique is that the model weights and the assumed error distribution do not change with respect to the hydrograph [Marshall *et al.*, 2007]. Understanding that different models have strengths in capturing different parts of the hydrograph, various adaptations of BMA in the literature have been performed. Duan *et al.* [2007] split the training period into different flow intervals and computed separate weights for each interval. Raftery *et al.* [2005] chose to create a sliding window basing the weights on some defined length of historical record.

[20] The BMA sliding window approach acts identical to the BMA static approach except that the time series of observations and model estimates used in the expectation maximization algorithm is limited to a shorter sliding window surrounding the forecast, instead of an isolated calibration time period. This approach is supported by an autocorrelation analysis on the observed flows for a calibration time period on the Leaf River Basin, Collins, Mississippi. This analysis illustrates that the flows are only correlated for a maximum of 40 previous days, but the most significant correlation happens within the first few days.

[21] For the BMA sliding window technique, two approaches can be taken to unbiased the model forecast using equation (1). The approach taken by Raftery *et al.* [2005] was to bias correct the data independently for each forecast, therefore each forecast time step would have different coefficients a and b . As an alternative to this approach, we used a separate calibration period to bias correct our forecast. Static coefficients for equation (1) are then used for all future forecasts. As an example, we calculated the performance measures of this technique versus the previous sliding window approach on 30 yr of data over the Leaf River Basin using the models discussed in section 3.1.

[22] Comparing the two approaches, we determined that using static unbiasing coefficients and smaller window

sizes results in smaller errors, while for dynamic unbiasing coefficients, smaller window sizes correspond to larger errors. Overall, the static unbiasing coefficients outperformed the dynamic coefficients for all point-wise performance measures we consider (Figure 2). Intuitively, this makes sense as the weights determined from the expectation maximization algorithm are more sensitive to the location on the hydrograph, and can give more weight to models that perform better at that particular location.

2.4. Sequential Bayesian Model Combination (SBC)

[23] Instead of setting up (3) as an optimization problem, with the weights and variance as free variables, another approach is to use a recursive definition of Bayes' law to calculate the posterior distribution, $P(M_i|Y)$. Similar to the Bayesian sliding window approach introduced earlier, this approach allows the weights to change over time, although the model variance will remain the same [Hsu et al., 2009].

[24] Starting again with (2),

$$P(y_t|M_1, M_2, \dots, M_k, Y) = \sum_{i=1}^k P(y_t|M_i, Y)P(M_i|Y). \quad (7)$$

The conditional posterior likelihood of a model given the training period $P(M_i|Y)$ can be calculated using the Bayes' equation,

$$P(M_i|Y_T) = \frac{P(Y_T|M_i) * P(M_i)}{\sum_{i=1}^k P(Y_T|M_i) * P(M_i)}, \quad (8)$$

Figure 2. Performance measures for different sliding window approaches.

letting

$$P_t(M_i) = P(M_i|Y_{T-1}) = \frac{P(Y_{T-1}|M_i) * P_{t-1}(M_i)}{\sum_{i=1}^k P(Y_{T-1}|M_i) * P_{t-1}(M_i)}. \quad (9)$$

This implies that the posterior distribution is a weighted average of its current forecast performance weighted by the previous time steps conditional probabilities.

[25] The calculation of the posterior density still requires the computation of $P(y_t|M_i, Y_T)$. A similar assumption that this conditional distribution is Gaussian can be made to implement this approach. However, a single variance is calculated as an average sum of square differences from a calibration time period, where the bias correction is also performed. Our preliminary analysis showed that this static weight performed worse than the static BMA scheme for probabilistic performance measures; thus, it was not considered in our analysis, although a modification of this technique (section 2.7) using the PF is presented.

2.5. Bayesian Model Averaging and Sequential Data Assimilation

[26] Sequential data assimilation estimates the observational and state uncertainty as a PDF around the optimal estimation of the system state. To introduce this approach, we first consider a general state-space formulation for any stochastic hydrological model.

[27] Let

$$x_t^f = f(x_{t-1}^u, \theta, u_t) + \omega_t, \quad (10)$$

$$y_t = h(x_t^f) + v_t, \quad (11)$$

where, x_t^f is a n -dimensional vector describing the system states forecast. $f(\bullet, \bullet)$ is the forward model that propagates the forcing data u_t into the system with updated states x_{t-1}^u from the previous time step, θ is the model parameter, and ω_t represents the process noise. y_t is a scalar forecast for an observation that is related to the system state through the operator $h(\bullet)$ and some observational noise v_t .

[28] To make the structure of the formulation applicable to the current discussion, it is useful to consider how the general formulation might apply to a rainfall-runoff model. In this case, the system state represents the volume of storage in each of the model catchments. The current system state is a function of the previous state and forcing data: precipitation and evaporation. The uncertainty associated with the state forecast is a result of the model structure and forcing errors. The observation model is a function of the current state of the system, and represents flow or storage volume exceeding a maximum system capacity. The observation noise can be associated with both errors in the rating curve or the gage accuracy.

[29] The goal of sequential data assimilation is to calculate an optimal estimation and confidence of the system state, x_t , given some scalar observations $Y_T = [y_1, y_2, \dots, y_T]$, illustrated by the posterior density $P(x_t^+|Y_T)$. Although it is possible to derive an analytical form of the posterior density using a recursive form of

Bayes' theorem, the multidimensional integration required to actually compute the density proves difficult for most hydrological models [Moradkhani *et al.*, 2005a]. To overcome this difficulty of calculating the analytical solution for the posterior density, sequential Monte Carlo methods are employed [Moradkhani *et al.*, 2005a].

[30] Sequential data assimilation uses the operators f and h in equations (10) and (11) to develop prior uncertainty estimates by assuming distributions for the process and observational noise. If the operators f and h are linear, the exact state posterior distribution can be determined using the standard Kalman filter [Kalman, 1960]; and if w are v are assumed to be Gaussian the ensemble Kalman filter can be used. However, a PF works with any operator and noise distribution. In general, hydrological models are highly nonlinear with non-Gaussian error distributions, suggesting the PF is better suited to characterize the system uncertainty [Moradkhani *et al.*, 2005a; Moradkhani, 2008]. In a recent study, DeChant and Moradkhani [2012] showed that the PF provides more robust and consistent results in the presence of highly nonlinear observation model, as compared to the ensemble Kalman filter. Therefore, in this study we chose the PF as the data assimilation procedure.

2.6. Particle Filter Algorithm (PF)

[31] The particle filter uses a sequential Monte Carlo simulation method to numerically integrate the posterior distribution $P(x_t|Y)$. This algorithm (algorithm 2) replicates a large number of random realizations (i.e., particles) of the system state to represent the true density function. This finite set of particles is propagated forward in time as the model is integrated with uncertain forcing data, and further perturbed with process noise to account for model structural error. After generating the replicates of model states through the forward model, here the hydrologic model, an observation is assimilated into the system and weights are associated with the particles based on the likelihood of each particle's simulated observation proximity to the real observation. The posterior density is then approximated by sum of discrete measures:

$$P(x_t|Y) = \sum_{i=1}^N w_t^i \delta(x_t - x_t^i), \quad (12)$$

where N is the number of particles.

[32] The particle weights are determined using the principle of sequential importance sampling (SIS), which defines the particle weights at time t , which are directly related to the particle weights at the previous time step through recursive Bayes' law. A problem with this approach, however, is the degeneration of particles weights. Weight degeneration refers to only a small number of particles having significant weights, leaving too few particles to represent the state PDF. This can be overcome by resampling when the effective sampling size becomes less than some fixed value [Arulampalam *et al.*, 2002]. Another approach implemented in this study is sequential importance resampling (SIR), which resamples the particles at every time step [Moradkhani *et al.*, 2005a]. All resampling methods replace particles with insignificant weight by higher weighted particles to more effectively build the state PDF.

[33] Algorithm 2. Particle filter algorithm using a SIR technique. For more detail, see Moradkhani *et al.* [2005a].

```

Set:  $N$  = number of particles ] Initialize state estimate
 $X_0 = (x_0^1, x_0^2, x_0^3, \dots, x_0^n)$ 

For  $t = 1$  to  $T$ 
  For  $i = 1$  to  $n$ 
    SET:  $w_{t-1}^i = 1/n$ 
     $x_t^{i-} = f(x_{t-1}^{i-}, u_k) + \omega_t \quad \omega_t \sim N(0, Q_t)$  ] Forward model:
    ] State forecast
     $y_t^i = h(x_t^{i-}) + v_t \quad v_t \sim N(0, R_t)$  ] Observation simulation

  Next  $i$ 
  Read Observation:  $y_t$ 

   $L(y_t|x_t^{i-}) = \frac{1}{\sqrt{2\pi R_t}} e^{\left(\frac{-(y_t - y_t^i)^2}{2R_t}\right)}$ 
   $P(y_t|x_t^{i-}) = \frac{L(y_t|x_t^{i-})}{\sum_{i=1}^N L(y_t|x_t^{i-})}$ 
   $w_t^{i+} = \frac{w_{t-1}^i P(y_t|x_t^{i-})}{\sum_{i=1}^N P(y_t|x_t^{i-})}$  ] Determine particle weights

   $x_{t+1}^+ = (x_{t+1}^1, x_{t+1}^2, x_{t+1}^3, \dots, x_{t+1}^n)$  ] Resample

Next  $t$ .
```

2.7. Bayesian Model Averaging With Data Assimilation (BMA_PF, SBC_PF)

[34] In the previous discussion on Bayesian model averaging the strategies determined the posterior density, $P(M_i|Y_T)$, by assuming a normal distribution for the posterior probability of y_t given a model M_i , $P(y_t|M_i, Y_T)$. In section 2.7, we relax that assumption by allowing the observation simulation, developed in the state-space formulation, to approximate this value. The advantage of using a PF is that it can identify the multimodality or skew in state estimation, therefore allowing the simulated observation to be multimodal or skewed.

[35] To understand this approach, we consider the PF discussed in algorithm 2. Let us assume we have just resampled a new ensemble of state forecasts for time t . As we discussed earlier, we can approximate the posterior probability using the following equations:

$$X_t = (x_t^1, x_t^2, x_t^3, \dots, x_t^n) \rightarrow P(x_t^+|y_t) \sim \sum_{i=1}^N w_t^i \delta(x_t - x_t^{i+}). \quad (13)$$

[36] In the particle filter approach, the replicates of model states forecasts are generated forward in time using the hydrologic model. The set of simulated observations are also generated forward in time using the observation equation, and then used to compute the posterior distribution, $P(y_t|M_i, Y_{T-1})$ (Figure 3).

Figure 3. The $P(y_t | M_i, Y_{T-1})$ can be estimated as the ensemble of simulated observations output from the rainfall-runoff models.

[37] We employ a kernel density smoother [Wilks, 2006], to create the probability density function of the model ensemble created from the PF. Like a histogram, the kernel density smoother divides the ensemble into intervals and calculates a probability for the center of this interval. Unlike a histogram, however, these intervals can overlap and act more like a sliding window moving across the ensemble members creating a more continuous curve. The other difference between a density smoother and a histogram is that the probability at the center of the interval is based on a weighted average of a kernel, a value from 0 to 1 that acts like density surrounding the particles in the interval, rather than just a tally of ensemble members.

[38] Calculating the posterior density using a particle filter has few advantages. First, since the particle filter produces almost the complete probability distribution function, multimodal density functions can be incorporated in the analysis scheme that better reflects the true uncertainty. Second, computationally for the BMA_SW there is no need to estimate the variance using historical records or the expectation maximization algorithm. Without the need to compute an optimal variance, the latent variable z can simply be calculated in one step as a relative ratio of the likelihood. For the sequential Bayesian combination (SBC) approach the distribution can be substituted directly into equation (9).

2.8. BMA Particle Filter and Model Selection With Minimum Weight Thresholds (BMA_PF_Threshold)

[39] Another basic assumption inherent in the BMA methods discussed above is that the subset of models contributing to the eventual forecast density remains constant. In this regard, the weights or contribution of the

strongest performing models are high, creating the central tendency of the forecast density while the poorer performing models are given less weight adding to the uncertainty. However, in general, individual models perform well over only a subset of the hydrograph. For example, an individual model may be calibrated to a high flow, while other models may be calibrated to low flows. Including all of the models, even if they are not performing well at a particular stage in the hydrograph, this may dampen the effectiveness of the BMA scheme by skimming the weights of the better performing models and therefore reducing the sharpness of the forecast density. The question remains as to how much this increased sharpness may affect the overall reliability of the forecast. As a way to explore the effectiveness of this assumption, we develop a final BMA strategy that allows the subset of models to be averaged to a change in time.

[40] The sequential nature of the BMA_PF strategy, with a sliding window of one time step, yields itself naturally to test this assumption. The BMA with a PF and model selection (BMA_PF_Threshold) strategy is performed using the following algorithm. Initially, the BMA_PF strategy is implemented including the entire set of models. The model with the minimum value below a predefined threshold is then removed from the set of models. The BMA_PF strategy is performed again on the reduced set of models. This process continues until the weights of all included models are above the minimum threshold. For this study we consider threshold values of 0.01, 0.10, 0.20, 0.3, 0.4, and 0.5, chosen to represent a monotonic decreasing function from the maximum number of models to the single best model. Figure 4 shows the average number of models used for each threshold level on the 10 yr Leaf River calibration period.

Figure 4. Average number of models used for a 10 yr calibration period using the BMA_PF_Threshold strategy with varying thresholds.

3. Case Study

[41] For this study we consider the historic records of the Leaf River Basin near Collins, Mississippi. The Leaf River Basin drains at $\sim 1950 \text{ km}^2$. The available data contained daily average precipitation and evapotranspiration, and mean daily stream discharge rates. The time period 1953–1963 was defined as the calibration for the BMA strategies while 1980–1988 was defined as the validation time period.

[42] As shown in Figure 5, the Leaf River Basin displays an annual cycle of six wet months, December–May, followed by six dry months, June–October. The variance of the recorded flow records peaks around February and is at a minimum from September through October. Statistics show a mean flow rate of 27.11 cm and maximum and minimum values of 1313.12 cm and 1.55 cm, respectively.

3.1. Model Structure

[43] In this study, we utilize two conceptual rainfall-runoff models, the Sacramento Soil Moisture Accounting

Figure 5. Seasonality of Leaf River for calibration time period.

(SAC-SMA) and HYMOD models. SAC-SMA is a lumped rainfall-runoff model with 16 model parameters developed by *Burnash et al.* [1973]. It remains widely used by the National Weather Service (NWS) in predicting streamflow at different time scales. HYMOD is a parsimonious model which is an extension of simple lumped storage models developed in the 1960s with only five parameters and five state variables initially developed for research purposes at the University of Arizona [*Boyle et al.*, 2001]. It has its origins in the probability distributed moisture model (PDM) [*Moore*, 1985] and has been used in several other studies (*Wagener et al.* [2001], *Moradkhani et al.* [2005b], among others). Both of these models have previously been used in Bayesian model averaging studies [*Duan et al.*, 2007; *Vrugt and Robinson*, 2007].

3.2. Model Calibration

[44] To calibrate the models, the shuffle complex evolution algorithm (University of Arizona [SCE-UA]) [*Duan et al.*, 1993] was employed. SCE-UA has been used extensively and reported to be an efficient global optimization method for the calibration of conceptual hydrologic models [*Muttill and Jayawardena*, 2008]. Recent studies by *Ebtehaj et al.* [2010], however, showed that the robustness of this algorithm can be improved by using a moving block bootstrap resampling method. To address the uncertainty in parameter estimation, we used three distinct objective functions including the root-mean-square error (RMSE), heteroscedastic maximum likelihood estimator (HMLE), and the absolute bias. The RMSE is an appropriate measure when the measurement errors are known to be uncorrelated and homoscedastic, or when the properties of the measurement errors are unknown [*Gupta et al.*, 1998]. On the other hand, the HMLE is a goodness-of-fit estimate when the measurement errors are believed to be heteroscedastic [*Sorooshian and Dracup*, 1980]. These objective functions force the hydrologic models to favor different phases of the hydrograph. The RMSE and bias force the models to fit the high and low flows, respectively, while the HMLE places an equal emphasis on all parts of the streamflow hydrograph, which compromises between RMSE and bias [*Duan et al.*, 2007; *Najafi et al.*, 2011]. A total of two conceptual models, each with three distinct parameter sets, are considered to generate six hydrologic models in this study.

4. Forecast Verification

[45] In section 4 we compare the skill of the different Bayesian model averaging schemes using both point-wise and probabilistic performance measures. Table 1 outlines the basic differences of each BMA strategy. This section is broken up into three parts. In the first section, 4.1, we will describe the performance measures used to quantify the skill of the forecast. In sections 4.2 and 4.3 we evaluate the skill of the different Bayesian model averaging schemes on the Leaf River data set described previously, first, evaluating the skill on the complete hydrograph and then on subsets of the hydrograph with varying characteristics.

4.1. Performance Measures

[46] The goal of forecast verification is to summarize the relationship between a predicted value and its corresponding observation, in order to determine the effectiveness of

Table 1. Comparison of BMA Strategies Evaluated

Technique	Name	Dynamic Weights	Dynamic Variance	Error Distribution	Average of Models
Static Bayesian model averaging	BMA_Static	No	No	Gaussian	6
Sequential Bayesian combination	SBC	Yes	No	Gaussian	6
Bayesian model averaging with a sliding window	BMA_SW	Yes	Yes	Gaussian	6
Sequential Bayesian	SBC_PF	Yes	Yes	Particle filter	6
Bayesian model averaging with a particle filter	BMA_PF	Yes	Yes	Particle filter	6
Sequential Bayesian combination with a particle filter	SBC_PF	Yes	Yes	Particle filter	6
Bayesian model averaging with model selection	BMA_PF_Threshold	Yes	Yes	Particle filter	Varies on threshold value

a forecasting technique across a variety of hydrological conditions and with respect to other forecasting techniques. It is evident that a single performance measure on a single hydrologic condition is not sufficient in answering all of those questions. In this study, we calculate three performance measures associated with accuracy and skill. More details of this performance measures can be found in the work of Wilks [2006] and the NWS River Forecast Verification Plan [National Weather Service (NWS), 2006; Demargne et al., 2010].

[47] Accuracy is a measure of the error between a prediction and a corresponding observation, which we assume as a true value. For this category, we consider the deterministic forecast of the different BMA strategies utilizing the expected value as the single-value representation of the forecast density. A common metric, the percent bias (PBIAS), is analyzed.

[48] Forecast skill compares the forecasts' performance to some reference forecast, often historical climatology. The goal of forecast skill is to assign the percent increase or decrease in the performance of a forecast relative to some benchmark technique. The forecast skill of a particular forecast can be calculated relative to any of the performance measures. In this analysis, we calculate two common measures: the Nash-Sutcliffe efficiency (NSE) and the ranked probability skill score (RPSS).

[49] The NSE is a point-wise performance skill measure. It uses the mean square error (MSE) as its score and the observation mean as the reference forecast. The RPSS relies on a probabilistic performance measure, the ranked probability score as its score with the climatology discussed below as its reference forecast. The ranked probability score (RPS) breaks the forecast density into multiple bins, providing a way to measure both the skill of predicting a certain event occurring along with calculating the distance a forecast density may be off from the observation. In a mathematical form, RPS is the sum of the squared error of the cumulative probability forecasts averaged over multiple events. In streamflow forecasting, the probability forecast is usually expressed using a nonexceedance probability forecast within prespecified categories calculated from historical observations. (i.e., 5%, 10%, 25%, 50%, 75%, 90%, 95%, and 99% nonexceedance). The observed value for a given threshold (forecast category) takes on the value of 1 if the observed flow value is greater than the threshold for that category. Otherwise, the observed value is 0. The discrete expression of RPS is given as

$$RPS_t = \sum_{i=1}^J [F_i^t - O_i^t]^2, \quad (14)$$

where F_i^t is the forecast probability at time t given by p (forecast < thresh _{i}) and O_i^t is the observed probability

given by p (observed < thresh _{i}), where i is the probability category.

[50] Verification statistics such as root-mean-square error and RPS are less meaningful when used in absolute terms. Therefore, forecasters prefer to calculate the relative scores and obtain skill scores which will range between 0 and 1 [Wilks, 2006]. Skill scores, such as the rank probability skill score (RPSS) are usually computed as the percentage improvement over a reference score (e.g., climatology):

$$RPSS = \left(1 - \frac{RPS}{RPS_{ref}}\right) \times 100 = \left(1 - \frac{RPS}{RPS_{climatology}}\right) \times 100. \quad (15)$$

For this study, we define the $RPS_{climatology}$ using the calibration time period. This is accomplished by creating a probability of nonexceedance curve using the historical observations. The threshold flow value for each bin is calculated for each of the selected nonexceedance probabilities discussed above using the corresponding flow from this curve. The climatology forecast then directly corresponds to the selected nonexceedance probabilities (Figure 6).

[51] For this study we also consider two other probabilistic performance measures, the width of the 95% prediction interval and the percent of observations that fall within this interval. The width of the 95% prediction interval quantifies the amount of uncertainty that surrounds the forecast, while the percent of observations that fall within the

Figure 6. Climatology for the Leaf River near Collins, MS. The diamond-shapes represent the 5%, 10%, 25%, 50%, 75%, 90%, and 95% nonexceedance probabilities.

prediction interval measures reliability of the uncertainty quantification. A forecast with a narrow prediction interval and low percentage of observations falling within this interval are overconfident. A forecast with a large prediction interval may be underconfident, overestimating the uncertainty of the forecast. The optimal forecast would have as small of a 95% prediction interval as possible while still capturing 95% of the observations. In this regard, these two performance metrics should be viewed together.

[52] Additional probabilistic forecast measures can be used including normalized root-mean-square error ratio (NRR) [Moradkhani *et al.*, 2005b; Moradkhani and Meskele, 2009], Q-Q plots [Thyer *et al.*, 2009], reliability diagrams, and different decompositions of the Brier score [Clark and Slater, 2006]. However, because of the quantity of methods being compared, the analysis is limited to the three probabilistic measures described above, to concisely demonstrate the forecast skill across all of the BMA strategies.

4.2. Performance Measures on Complete Hydrograph

[53] We begin this analysis by first evaluating the pointwise performance measures across the entire 8-yr validation time period for the Leaf River Basin. Figure 7 illustrates the results for all of the different BMA strategies outlined previously in the paper, and Table 2 shows the performance results of the individual models. There are a few noteworthy results to be explored in this analysis. First, while the BMA_PF_Threshold with a threshold of 0.40 produced the higher NSE and lower bias when compared to the other averaging strategies, it was incapable of outperforming all of the individual model filtering experiments. The individual PF HYMOD_PBIAS model produced the lowest mean bias but had a significantly lower NSE than the BMA_PF_Threshold with a threshold of 0.40. While on the other hand, the SAC_RMSE individual PF produced the highest NSE, but it had a significantly higher mean bias compared

Table 2. Pointwise Performance Measures for Individual Models

	Models	Mean Bias	NSE
Deterministic model	Hymod_HMLE	6.6813	0.70652
	Hymod_PBIAS	1.3251	0.74774
	Hymod_RMSE	7.0171	0.78921
	SAC_HMLE	0.21771	0.81588
	SAC_PBIAS	1.6585	0.76208
	SAC_RMSE	4.5353	0.86809
Particle filter	Hymod_HMLE	3.2453	0.81672
	Hymod_PBIAS	−0.05295	0.78383
	Hymod_RMSE	3.144	0.83224
	SAC_HMLE	−0.28105	0.83674
	SAC_PBIAS	1.1422	0.77471
	SAC_RMSE	3.4449	0.89354

to the BMA_PF_Threshold with a threshold of 0.40. Performance by the model averaging strategies show less polarizing results between the two metrics. This suggests that although the model averaging schemes may not be able to outperform the single best model on a particular performance metric, the averaging can balance out the weaknesses of single calibration models, producing better overall performance across both NSE and bias. Considering the combination of both metrics, SAC_PF_HMLE is the most competitive to the averaged model with a reduced mean bias and competitive NSE when compared to the BMA_PF_Threshold equal to 0.40.

[54] The other noteworthy results are related to how the performance metrics change with the BMA averaging strategy parameters, such as the assumed error distribution window size and the threshold value. From Table 2, the individual PF models outperformed the individual deterministic models for all point-wise performance metrics. Therefore, it is not surprising that overall, the model averaging schemes using a PF outperformed the averaged deterministic models assuming a Gaussian distribution. Also, for both

Figure 7. Pointwise performance measure for the complete hydrograph.

Figure 8. RPSS values for the complete hydrograph.

measures, a decrease in window size led to improved performance metrics. Finally, there seemed little change in the NSE metric with a reduction in the number of models due to an increase in the threshold, however, averaging a smaller number of models did seem to positively affect the mean bias metric.

[55] As discussed by *Doblas Reyes et al.* [2005] and *Palmer et al.* [2004], the strength of multimodel averaging is most evident when comparing probabilistic measures, such as the RPSS.

[56] Figure 8 shows the results of a RPSS performance measure across the entire hydrograph. For this metric, the Gaussian distribution strategy with a sliding window of one time step outperformed all other strategies. Following the trends seen in the point-wise metrics, an increase in

window size corresponds to lower performance. Interestingly, opposite of the point-wise metrics, the RPSS values decreased with an increase in the threshold value. This seems to relate to an increase in the width of a 95% prediction interval, when more models are averaged as shown in Figure 9 as previously shown in the work of *Voisin et al.* [2010].

[57] Looking further into why the Gaussian distribution outperformed the PF strategies in the RPSS values, we consider the other two descriptive measures of the distributions created from the model averaging strategies: the width of the 95% prediction interval and the percent of observations falling within this interval.

[58] Overall, the width of the prediction intervals using the Gaussian distribution was nearly doubled (showing

Figure 9. Percent of observations that fall within the forecasted 95% prediction interval for the complete hydrograph.

higher uncertainty) when compared to the PF. However, more than 95% of the observations fell into the prediction intervals for all BMA strategies assuming a Gaussian distribution and a window size greater than five. On the other hand, at most, only 85% of the observations fell into the prediction intervals using the PF.

[59] Since the individual PF models produce a distribution, we can compare the probabilistic skill of these models to the BMA averaged strategies. As shown in Figure 10, the BMA strategy with a sliding window of one time step significantly improved the RPSS values of the individual PF models. Both the Sacramento model and the HYMOD model were overconfident in their forecast with only ~40% of the observations falling within this 95% prediction interval. This illustrates the power of BMA in combination with a PF. While constructing a 95% prediction width similar to the individual PF models, the averaged scheme pushed the number of observations falling into the 95% prediction interval to more than 60%.

[60] Over the entire hydrograph, the performance measures presented here suggest that the Gaussian BMA produced a more reliable 95% predictive interval, but at the expense of lower accuracy and significantly increased uncertainty, in comparison to the BMA-PF. While reliability remains the major performance metric when compared to minimizing the implied uncertainty, the increased accuracy of the BMA-PF suggests that it may be preferred over the Gaussian BMA for certain regions of the hydrograph. From a river forecasting perspective, the question to ask is what magnitude of observations is the BMA-PF missing and where are the benefits of higher accuracy observed? This question is important because a mischaracterization of low flows is nearly inconsequential in comparison to that of poorly estimated high flows. This begs for an analysis of the performance of each method under portions of the hydrograph with differing dynamics. Section 4.3 examines the hydrograph in different rates of volatility to address the consequences of assuming a Gaussian distribution or relying on the PF when using BMA.

Figure 10. Comparison of RPSS values off individual particle filter models to a Bayesian model averaging scheme with a sliding window of 1.

4.3. Performance Measures on Various Rates of Volatility

[61] One way to consider how the BMA schemes perform across different regions of the observed hydrograph is to separate the hydrograph into different flow values, for example, high, medium, and low flows. The sliding window schemes we are analyzing, however, not only address the volume of flow, but also the potential of the scheme to quickly adapt to rapid changes in the hydrograph. In this regard, we analyze how the BMA strategies perform across different parts of the hydrograph using three separate 200-d periods. This separation is defined to distinguish regions in the hydrograph that reflect both different flow values and different rates of volatility, (average daily rate of change) as illustrated in Figure 11. Table 3 outlines the leading performers for the point-wise metrics.

[62] For each range of volatility, different Bayesian model averaging schemes performed the best. In low volatility ranges, the Gaussian distribution with a sliding window of 1 outperformed the other strategies. However, for medium and high volatility stages, the PF strategies performed the best. A reduction in the number of models averaged seemed to increase the NSE for high volatility stages, while including all of the models seemed to increase the NSE for medium volatility ranges (Figure 12). This effect can be explained by further looking into the results where the individual PF-based forecasts showed more spread in high volatility ranges, and that any average of a number of competing models may shift the expected value significantly away from any one model forecast. Limiting the number of models averaged by using a threshold reduced the overall bias of the averaged forecast in all situations.

[63] Coinciding with the different BMA strategies performing best in different volatility ranges, different individual models also performed best across the volatility ranges (Table 3). In low volatility ranges, the PF-SAC-HMLE model performed the best in both NSE and mean bias metrics, competing well with or outperforming all BMA strategies. For medium volatility ranges, both the individual HYMOD and PF-PBIAS calibrated models performed best compared to the other individual models. In this volatility range, the individual models gave similar results to the BMA strategies in terms of the NSE but significantly underperformed the leading BMA strategies in terms of mean bias. Finally, for the high volatility range the PF-HYMOD-PBIAS and the PF-SAC-RMSE outperformed the BMA strategies in terms of mean bias and NSE, respectively. However, the results are similar to the analyses over the complete hydrograph, where an improved performance in one performance metric corresponds to a poor performance in other metrics. The BMA strategies again seemed to balance those extremes.

[64] Turning toward an evaluation of the probabilistic skill for each BMA strategy across the different rates of volatility, we consider how these strategies change with respect to three different variables: sliding window size, number of models averaged, and type of distribution. Our performance evaluation is based on the three metrics discussed above: RPSS, a width of 95% prediction interval, and the number of observations falling into the 95% prediction interval.

Leaf River Hydrograph (Oct 1,1980-Sep 30,1988)

Figure 11. Volatility stages for Leaf River Basin during the validation period of 1 October 1980–30 September 1988.

[65] For all distribution assumptions, strategies with smaller window sizes correspond to higher RPSS values (Figure 13). As the window sizes increase, averaging weights are balanced out among all of the competing models. This

Table 3. Pointwise Performance Measures Across Different Stages in the Hydrograph^a

Low Volatility

Strategy		NSE	Mean Bias
Gaussian	BMA_STATIC	−0.39	2.89
	Sliding window = 1	0.59	0.94
Particle filter	Sliding window = 1	0.45	0.92
	Threshold = 0.03	0.36	0.82
	SBC_PF	0.39	0.88
Individual models	PF_SAC_HMLE	0.58	−0.60

Medium Volatility

Strategy		NSE	Mean Bias
Gaussian	BMA_STATIC	0.63	7.20
	Sliding window = 1	0.82	3.35
Particle filter	Sliding window = 1	0.85	0.85
	Sliding window = 2	0.87	1.25
	Threshold = 0.4	0.84	0.58
	SBC_PF	0.80	3.08
Individual models	Hymod_PBIAS	0.75	3.08
	PF_SAC_PBIAS	0.85	3.44

High Volatility

Strategy		NSE	Mean Bias
Gaussian	BMA_STATIC	0.81	13.89
	Sliding window = 1	0.82	10.80
Particle filter	Sliding window = 5	0.82	3.34
	Threshold = 0.5	0.83	2.56
	SBC_PF	0.81	5.80
	PF_HYMOD_PBIAS	0.72	−2.35
Individual models	PF_SAC_RMSE	0.88	13.50

^aLeading performers are listed for each BMA strategy and individual models.

spreads out the width of the 95% prediction interval (Figure 14). Surprisingly, this is true even for stages of low volatility, where it might be assumed that the best performing models remained consistent throughout the period. For the PF, the increase in the size of the prediction interval allows a higher percentage of observations to fall into that interval, but also corresponds to a decrease in the RPSS across all ranges of volatility.

[66] Varying the minimum weight thresholds did not significantly modify the RPSS for any volatility range, although a small decrease in RPSS value is evident as the threshold values increased (smaller set of models averaged). The most obvious consequence of averaging a smaller number of models is the decrease in the width of the 95% prediction interval, and therefore the percent of observations found in that interval.

[67] For RPSS values, the PF scheme with a sliding window of 1 outperformed the Gaussian distribution in both

Figure 12. Effects of threshold level on the RMSE performance measure.

Figure 13. RPSS values for various rates of change on the hydrograph.

the medium and high volatility ranges; however, the Gaussian distribution outperformed the PF in the low volatility range. In all volatility ranges, the actual RPSS values are very similar for both error distribution assumptions. The main difference between these two approaches lies in the width of the 95% prediction intervals and number of observations found in these intervals. The width of the 95% prediction interval for the Gaussian distribution is almost double the width of the PF for all volatility ranges.

However, the Gaussian distribution comes close in all ranges of volatility to having 95% of the observations in this interval. The PFs on the other hand show only 70% of observations in the prediction interval when using a BMA strategy with a sliding window of 1 for the low and high volatility ranges, and close to 80% in the medium volatility range. However, when lengthening the window size, the BMA PF strategy was able to capture close to 95% of the observations, while reducing the prediction interval by

Figure 14. Percent of observations that fall within the forecasted 95% prediction interval for various rates of change in the hydrograph.

Figure 15. Comparison between individual particle filters and BMA strategies in the percent of observations that fall within the forecasted 95% prediction interval and width of the prediction interval.

nearly 40% compared to the highest performing strategies that assume a Gaussian distribution.

[68] A comparison between the individual PFs and the BMA-PF-1 strategy is similar to results over the complete hydrograph. First, the analysis showed an increase in RPSS with the BMA-PF-1 strategy for each range of volatility, which is analogous to Figure 10, over the individual models. Most interesting, however, was the ability of the BMA-PF-1 strategy to increase the percentage of observations in the 95% prediction interval without dramatically increasing the width of the prediction interval as shown in Figure 15.

5. Conclusion

[69] The basic assumption of model averaging is that there is no one best choice in model structure. From this assumption, it may further be hypothesized that there is no best Bayesian model averaging scheme. The analysis shows that the strength of each of the different averaging techniques varies across performance metrics and volatility ranges of the hydrograph. Although this may be true, there exist some notable trends and observations. In section 5, we consider three major points: the strength of the averaged models compared to the individual models, the effect of volatility on model averaging performance, and finally, a discussion on the strengths and weaknesses of the PF error distribution applied to BMA. We conclude this section with a discussion on the limitations of the analysis and areas for future research.

5.1. BMA Strategies and Individual Models

[70] Agreeing with the literature [Doblas Reyes *et al.*, 2005] the advantage of the multimodel averaging is most apparent in measuring probabilistic skill, but some advantage is apparent for point-wise metrics. The comparison of the multimodeled averaged forecast with the individual models forecast illustrates that averaged forecast do not always outperform the best individual model for any particular metric (i.e., mean bias, NSE). However, in this study,

individual models, which performed well on a particular point-wise metric, performed poorly on the other. This especially holds true for the analysis over the complete hydrograph and high volatility ranges. The leading multimodel averaged (BMA-PF-Threshold equals 0.40) performer on the other hand balanced the two point-wise metrics, showing competitive values for both NSE and mean bias.

[71] A comparison of the BMA strategies to the individual models probabilistic forecasts confirmed the strength of using model averaging on overconfident individual models [Weigel *et al.*, 2008]. The individual PF models all exhibited extremely overconfident prediction intervals, in some cases capturing less than 50% of the observations. Despite the overconfidence of the individual models, the leading performing average of these models (BMA-PF-SW-1) was able to significantly increase the percentage of observations within the 95% prediction interval to 75%, and nearly doubling the RPSS value for the analysis across the complete hydrograph. Surprisingly, this is accomplished by not increasing the average width of the 95% prediction interval much more than any individual models average prediction interval width. This trend is consistent throughout all of the ranges of volatility, but especially clear in the high range of volatility, where the PF models are most overconfident.

5.2. Sliding Window Size and Minimum Model Thresholds

[72] Two variations of the static BMA considered in this study are to allow for dynamic model weights and to dynamically change the set of models averaged. Both of these variations showed conflicting results when compared across all of the performance measures.

[73] For point-wise metrics and RPSS values, the dynamic model weights and model uncertainty, generated by the sliding window approach, outperformed the static Bayesian model averaging scheme. This dynamic approach allows a necessary flexibility in gaging the confidence in a

model output, with respect to the changes in the hydrograph. Interestingly, the stages of volatility of the hydrograph did not significantly alter the general trend that smaller window sizes outperformed the larger window sizes.

[74] Larger window sizes more evenly spread the weights among all competing models and consequently increase the 95% forecast interval. This is most notable for the BMA schemes that utilize the overconfident PFs. Resulting from the increase in the prediction width, a higher percentage of observations are captured within the prediction intervals. However, the larger window sizes also reduce the RPSS values, suggesting that increasing the prediction interval with larger window sizes reduces the ability of the forecast to accurately represent the modes and skew of the individual PF forecast.

[75] The BMA-PF-Threshold strategy looks to dynamically reduce or increase the cardinality of the set of models averaged based on how the individual models perform relative to a minimum weight threshold. Increasing the minimum threshold (reducing the set of averaged models) produced higher point-wise metric scores for the analysis over the complete hydrograph and high volatility ranges, as the forecast merges toward the single best performing individual model at the previous time step. On the other hand, this takes away the strength of BMA for probabilistic forecasting, reducing the RPSS and the percentage of observations caught in the 95% prediction interval. Unlike the sliding window approach, an increase in the number of models averaged is linked to both an increase in RPSS and an increase in the number of observations found in the 95% prediction interval. This suggests the addition of more models can both increase the reliability of the forecast while retaining the integrity of the mode and skew of the individual forecasts.

5.3. Gaussian and PF Forecast Distributions

[76] One major motivation of this study was to challenge the standard BMA assumption of a Gaussian uncertainty distribution by comparing it to an approach using a PF. With consideration to the two point-wise measures, the BMA strategies using a PF outperformed the strategies assuming a Gaussian distribution for all measures over the complete hydrograph. This trend carries over to all of the volatility ranges except the low volatility range, where a strategy assuming a Gaussian distribution had a slightly higher NSE. This can easily be explained considering Table 2, which shows that the individual PF models outperformed their deterministic counterparts for all point-wise measures. It is not surprising that Bayesian averaging of the best models results in improved performance when compared to Bayesian averaging of models that performed poorly.

[77] Comparing the strengths of these two approaches becomes more muddled when considering the probabilistic forecast. Over the complete hydrograph, a strategy assuming a Gaussian distribution had a slightly higher RPSS than any of the strategies using the PF. Additionally, strategies assuming a Gaussian distribution reliably had prediction intervals that contained 95% of the observations, when compared to the highest performing PF that only captured 75% of the observations over the complete hydrograph.

[78] A more careful analysis of breaking down the validation period into periods of different ranges of volatility reveals that strategies using a PF outperform the Gaussian distribution in medium and high ranges in terms of the probabilistic performance assessment RPSS. The Gaussian distribution, however, still outperformed the averaged PF for the low range of volatility. After calculating the one step volatility described in section 4.3 as $|Q(t) - Q(t-1)|$ for observations in the validation time period, 64% of all observations fall within the low volatility range (Table 4). This suggests the higher RPSS value over the complete hydrograph might be a reflection of the distribution of the validation period that favored low-flow volatility ranges.

[79] Across the ranges of volatility, BMA-PF strategies reduce the prediction interval by nearly 50% when compared to strategies utilizing the Gaussian uncertainty distribution. For low and medium ranges of volatility, this corresponds to an overconfidence in the BMA-PF strategies, capturing fewer than 95% of the observations. However, for high volatility ranges, the BMA PF approach with a sliding window of 30 nearly captured 95% of observations within its prediction interval, while reducing the prediction interval width by 40% in comparison to the best BMA strategy.

[80] In all performance measures, point-wise and probabilistic, BMA strategies utilizing a particle filter outperform the strategies assuming a Gaussian distribution for high volatility. For low volatility and medium ranges, the BMA strategies using a PF remain overconfident. It should be noted, that even though a reduction of reliability (in terms of capturing the observations within 95% confidence) is seen for low and medium volatility ranges, this is with the advantage of higher accuracy and significantly decreased uncertainty as compared with BMA with Gaussian distribution assumption. This suggests that the Gaussian distributions may be under-confident even for these ranges of volatility.

5.4. Limitations to the Analysis

[81] The focus of this analysis is to introduce a new addition to Bayesian model averaging and quantifying the strengths of integrating data assimilation into BMA. Although certain general trends demonstrated in this analysis may persist beyond the scope of the design, a more careful approach is necessary to extend to any generalities. The design limitations reveal important areas for further research.

[82] First, the analysis was limited to a small number of models. As shown in section 4.3, the addition of more models increases the width of the 95% prediction interval without reducing the RPSS, suggesting more models might yield higher probabilistic skill. One question that remains, however, is whether there is some maximum number of models where the marginal cost of an additional model

Table 4. Percent of Daily Observation in the Validation Period Falling Within Different Volatility Ranges

Volatility Range	$ Q(t) - Q(t-1) $	Percent of Observations
Low	≤ 2	0.64
Medium	> 2 and ≤ 10	0.18
High	> 10	0.18

outweighs the benefit of an additional model? In line with this question, another question immediately follows as to what models should be added; more or less complex structures, parameterization sets based on different objective functions, or parameterizations based on different areas of the hydrograph?

[83] Similar to those questions, another area of concern is in the understanding of the benefits with regard to BMA to over- and underconfident models. In this analysis, the PF models appear to be overconfident. This overconfidence is, however, caused by ignoring parameter uncertainty. In a parallel study, applying the PF in a state-parameter estimation experiment yields a much more reliable predictive uncertainty, as demonstrated by *DeChant and Moradkhani* [2011b]. Current work is focused on combining this approach to BMA, and will provide a suitable platform in understanding this area of interest.

[84] Finally, this study is limited to the domain of the observations of the Leaf River Basin. These results may potentially be linked to times of concentration, intensity of precipitation, or even the natural variability in the watershed. To fully understand how this analysis could be generalized would require a careful design across a number of different types of watersheds and hydrological conditions.

[85] **Acknowledgments.** Partial financial support for this research was provided by NOAA-CSTAR, grant NA11NWS4680002, and NOAA-MAPP grant NA11OAR4310140.

References

- Ajami, N. K., Q. Duan, and S. Sorooshian (2007), 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.
- Arulampalam, M. S., S. Maskell, N. Gordon, and T. Clapp (2002), A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking, *IEEE Trans. Signal Process.*, **50**(2), 174–188.
- Bates, J. M., and C. W. J. Granger (1969), The combination of forecast, *Operations Res.*, **20**, 451–468.
- Boyle, D. (2001), Multicriteria calibration of hydrological models, Ph.D. dissertation, Univ. Arizona, Tucson.
- Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), A generalized streamflow simulation system: Conceptual modeling for digital computers, U.S. Dept. of Commerce, National Weather Service and State of California, Dept. of Water Resources, Sacramento, Calif.
- Clark, M. P., and A. G. Slater (2006), Probabilistic quantitative precipitation estimation in complex terrain, *J. Hydrometeor.*, **7**, 3–22.
- DeChant, C., and H. Moradkhani (2011a), Radiance data assimilation for operational snow and streamflow forecasting, *Adv. Water Resour.*, **34**, 351–364.
- DeChant, C., and H. Moradkhani (2011b), Improving the characterization of initial condition for ensemble streamflow prediction using data assimilation, *Hydrol. Earth Syst. Sci.*, **15**, 3399–3410, doi:10.5194/hess-15-3399.
- DeChant, C., and H. Moradkhani (2012), Examining the effectiveness and robustness of sequential data assimilation methods for quantification of uncertainty in hydrologic forecasting, *Water Resour. Res.*, doi:10.1029/2011WR011011, in press.
- Demargne, J., J. D. Brown, Y. Liu, D. J. Seo, L. Wu, Z. Toth, and Y. Zhu (2010), Diagnostic verification of hydrometeorological and hydrologic ensembles, *Atmos. Sci. Lett.*, **11**(2), 114–122.
- Diks, C. G. H., and J. A. Vrugt (2010), Comparison of point forecast accuracy of model averaging methods in hydrologic applications, *Stoch. Environ. Res. Risk Assess.*, **24**, 809–820, doi:10.1007/s00477-010-0378-z.
- Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer (2005), The rationale behind the success of multi-model ensembles in seasonal forecasting: II. Calibration and combination, *Tellus, Ser. A*, **57**(3), 234–252.
- Duan, Q., S. Sorooshian, and V. K. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, **28**(4), 1015–1031, doi:10.1029/91WR02985.
- Duan, Q. Y., V. K. Gupta, and S. Sorooshian (1993), Shuffled complex evolution approach for effective and efficient global minimization, *J. Optim. Theory Appl.*, **76**, 501–521.
- Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian (2007), Multi-model ensemble hydrologic prediction using Bayesian model averaging, *Adv. Water Resour.*, **30**(5), 1371–1386.
- Ebtehaj, M., H. Moradkhani, and H. V. Gupta (2010), Improving robustness of hydrologic parameter estimation by the use of moving block bootstrap resampling, *Water Resour. Res.*, **46**, W07515, doi:10.1029/2009WR007981.
- Granger, C. W. J., and R. Ramanathan (1984), Improved methods of combining forecast accuracy, *J. Forecasting*, **3**, 197–204.
- Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and non-commensurable measures of information, *Water Resour. Res.*, **34**(4), 751–763.
- Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999), Bayesian modeling averaging: A tutorial, *Stat. Sci.*, **14**(4), 382–417.
- Hsu, K., H. Moradkhani, and S. Sorooshian (2009), A sequential Bayesian approach for hydrologic model selection and prediction, *Water Resour. Res.*, **45**, W00B12, doi:10.1029/2008WR006824.
- Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *J. Basic Eng.*, **82D**, 35–45.
- Leisenring, M., and H. Moradkhani (2010), Snow water equivalent estimation using Bayesian data assimilation methods, *Stochastic Environ. Res. Risk Assess.*, **25**(2), 253–270.
- Marshall, L., D. Nott, and A. Sharma (2007), Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework, *Hydrol. Processes*, **21**(7), 847–861.
- Matgen, P., M. Montanari, R. Hostache, L. Pfister, L. Hoffmann, D. Plaza, V. R. N. Pauwels, G. J. M. De Lannoy, R. De Keyser, and H. H. G. Savenije (2010), Towards the sequential assimilation of SAR-derived water stages into hydraulic models using the Particle Filter: Proof of concept, *Hydrol. Earth Syst. Sci.*, **14**, 1773–1785.
- Montzka, C., H. Moradkhani, L. Weihermuller, M. Canty, H. J. Hendricks Franssen, and H. Vereecken (2011), Hydraulic parameter estimation by remotely-sensed top soil moisture observations with the particle filter, *J. Hydrol.*, **399**(3–4), 410–421.
- Moore, R. J. (1985), The probability-distributed principle and runoff production at point and basin scales, *Hydrol. Sci. J.*, **30**, 273–297.
- Moradkhani, H. (2008), Hydrologic remote sensing and land surface data assimilation, *Sensors*, **8**, 2986–3004, doi:10.3390/s8052986.
- Moradkhani, H., and T. Meskele (2009), Probabilistic assessment of the satellite rainfall retrieval error translation to hydrologic response, in *Satellite Applications for Surface Hydrology*, pp. 229–242, Springer, Water Sci. Technol. Library, Netherlands, doi:10.1007/978-90-481-2915-7.
- Moradkhani, H., and S. Sorooshian (2008), General review of rainfall-runoff modeling: model calibration, data assimilation, and uncertainty analysis, in *Hydrological Modeling and Water Cycle, Coupling of the Atmospheric and Hydrological Models*, vol. 63, pp. 1–23, Springer, Water Sci. Technol. Library, doi:10.1007/978-3-540-77843-1.
- Moradkhani, H., K. Hsu, H. Gupta, and S. Sorooshian (2005a), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using particle filter, *Water Resour. Res.*, **41**, W05012, doi:10.1029/2004WR003604.
- Moradkhani, H., S. Sorooshian, H. Gupta, and P. Houser (2005b), Dual state-parameter estimation of hydrological models using ensemble Kalman filter, *Adv. Water Resour.*, **28**(2), 135–147.
- Moradkhani, H., K. Hsu, Y. Hong, and S. Sorooshian (2006), Investigating the impact of remotely sensed precipitation and hydrologic model uncertainties on the ensemble streamflow forecasting, *Geophys. Res. Lett.*, **33**(12), L12401, doi:10.1029/2006GL026855.
- Muttill, N., and A. W. Jayawardena (2008), Shuffled complex evolution model calibrating algorithm: Enhancing its robustness and efficiency, *Hydrol. Processes*, **22**(23), 4628–4638.
- Najafi, M., H. Moradkhani, and I. Jung (2011), Assessing the uncertainties of hydrologic model selection in climate change impact studies, *Hydrol. Processes*, **25**(18), 2814–2826.
- National Weather Service (NWS) River Forecast Verification Plan (2006), Report of the hydrologic verification system requirements team, US Dept. of Commerce, NOAA/NWS, available at http://nws.noaa.gov/oh/rfcddev/docs/Final_Verification_Report.pdf.
- Palmer, T. N., et al. (2004), Development of a European multi-model ensemble system for seasonal to interannual prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, **85**, 853–872.

- Raftery, A. E., F. Balabdaoui, T. Gneiting, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *113*, 1155–1174.
- Rojas, R., L. Feyen, and A. Dassargues (2008), Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging, *Water Resour. Res.*, *44*, W12418, doi:10.1029/2008WR006908.
- Shamseldin, A. Y., K. M. O'Connor, and G. C. Liang (1997), Methods for combining the outputs of different rainfall runoff models, *J. Hydrol.*, *197*(1), 203–229.
- 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.
- Thyer, M., B. Renard, D. Kavetski, G. Kuczera, S. W. Franks, and S. Srikanthan (2009), Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis, *Water Resour. Res.*, *45*, W00B14, doi:10.1029/2008WR006825.
- Voisin, N., J. C. Schaake, and D. P. Lettenmaier (2010), Adaptation of calibration and downscaling methods for quantitative ensemble precipitation forecasts, *Weather Forecasting*, *25*(6), 1603–1627.
- Vrugt, J., and B. 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.
- Wagener, T., D. P. Boyle, M. J. Lees, H. S. Wheater, H. V. Gupta, and S. Sorooshian (2001), A framework for development and application of hydrological models, *Hydrol. Earth Syst. Sci.*, *5*(1), 13–26.
- Weerts, A. H., and G. Y. H. El Serafy (2006), Particle filtering and ensemble Kalman filtering for state updating with hydrological conceptual rainfall-runoff models, *Water Resour. Res.*, *42*(9), W09403, doi:10.1029/2005WR004093.
- Weigel, A. P., M. A. Liniger, and C. Appenzeller (2008), Can multi-model combination really enhance the prediction skill of probabilistic forecasts?, *Q. J. R. Meteorol. Soc.*, *134*, 241–260.
- Wilks, D. S. (2006), *Statistical Methods in the Atmospheric Sciences*, 2nd ed., 327 pp., Academic, N. Y.

C. M. DeChant, H. Moradkhani, and M. A. Parrish, Department of Civil and Environmental Engineering, Portland State University, 1930 SW 4th Ave., Ste. 200, Portland, OR 97201, USA. (hamidm@cecs.pdx.edu)