Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments

Joerg Rings,[1,2] Jasper A. Vrugt,[3,4] Gerrit Schoups,[5] Johan A. Huisman,[2] and Harry Vereecken[2]

Received 8 November 2011; revised 27 February 2012; accepted 18 March 2012; published 15 May 2012.

[1] Bayesian model averaging (BMA) is a standard method for combining predictive distributions from different models. In recent years, this method has enjoyed widespread application and use in many fields of study to improve the spread-skill relationship of forecast ensembles. The BMA predictive probability density function (pdf) of any quantity of interest is a weighted average of pdfs centered around the individual (possibly bias-corrected) forecasts, where the weights are equal to posterior probabilities of the models generating the forecasts, and reflect the individual models skill over a training (calibration) period. The original BMA approach presented by Raftery et al. (2005) assumes that the conditional pdf of each individual model is adequately described with a rather standard Gaussian or Gamma statistical distribution, possibly with a heteroscedastic variance. Here we analyze the advantages of using BMA with a flexible representation of the conditional pdf. A joint particle filtering and Gaussian mixture modeling framework is presented to derive analytically, as closely and consistently as possible, the evolving forecast density (conditional pdf) of each constituent ensemble member. The median forecasts and evolving conditional pdfs of the constituent models are subsequently combined using BMA to derive one overall predictive distribution. This paper introduces the theory and concepts of this new ensemble postprocessing method, and demonstrates its usefulness and applicability by numerical simulation of the rainfall-runoff transformation using discharge data from three different catchments in the contiguous United States. The revised BMA method receives significantly lower-prediction errors than the original default BMA method (due to filtering) with predictive uncertainty intervals that are substantially smaller but still statistically coherent (due to the use of a time-variant conditional pdf).

Citation: Rings, J., J. A. Vrugt, G. Schoups, J. A. Huisman, and H. Vereecken (2012), Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments, *Water Resour. Res.*, *48*, W05520, doi:10.1029/2011WR011607.

1. Introduction

[2] During the last decade, multimodel ensemble prediction systems have become the basis for probabilistic weather and climate forecasts at many operational centers throughout the world [*Molteni et al.*, 1996; *Grimitt and Mass*, 2002; *Barnston et al.*, 2003; *Palmer et al.*, 2004]. Multimodel ensemble predictions aim to capture several sources of uncertainty in numerical weather forecasts, including uncertainty about the initial conditions, lateral boundary conditions, and model physics, and have convincingly demonstrated improvements to numerical weather and climate forecasts and the production of more skillful estimates of forecast probability density functions (pdf) [*Krishnamurti et al.*, 1999; *Rajagopalan et al.*, 2002; *Doblas-Reyes et al.*, 2005; *Gneiting et al.*, 2005; *Min and Hense*, 2006, among others]. However, because the current generation of ensemble systems do not explicitly account for all sources of forecast uncertainty, some form of postprocessing is necessary to provide predictive ensemble pdfs that are meaningful, and can be used to provide accurate forecasts [*Hamill and Colucci*, 1997; *Richardson*, 2001; *Raftery et al.*, 2005; *Gneiting et al.*, 2005].

[3] Ensemble Bayesian model averaging (BMA) has been proposed by *Raftery et al.* [2005] as a formal statistical method for the postprocessing of forecast ensembles. Our concern is to find the predictive probability function $p(\tilde{\mathbf{Y}}_n|f_{1n}, \ldots, f_{Kn})$ of some quantity of interest $\tilde{\mathbf{Y}}_n = \{\tilde{y}_t; t = 1, \ldots, n\}$, which in our case is streamflow, but could also be temperature or sea level pressure as in the work of *Raftery et al.* [2005]. If $\{f_{1t}, \ldots, f_{Kt}\}$ denotes an ensemble of $t = 1, \ldots, n$ individual predictions obtained from K different models, then BMA approximates the

[1]Department of Land, Air, and Water Resources, University of California, Davis, California, USA.
[2]Agrosphere, IBG-3, Forschungszenstrum Jülich, Jülich, Germany.
[3]Department of Civil and Environmental Engineering, University of California, Davis, California, USA.
[4]Institute for Biodiversity and Ecosystem Dynamics (IBED), University of Amsterdam, Amsterdam, Netherlands.
[5]Department of Water Management, Delft University of Technology, Delft, Netherlands.

0043-1397/12/2011WR011607

predictive probability density function as a weighted average of the conditional pdfs, $g_k(\tilde{\mathbf{Y}}_n|f_{kn})$ of the individual predictors of the ensemble, or in mathematical notation

$$p(\tilde{\mathbf{Y}}_n|f_{1n}, \ldots, f_{Kn}) = \sum_{k=1}^{K} w_k\, g_k(\tilde{\mathbf{Y}}_n|f_{kn}). \quad (1)$$

The weights are derived from a training period, and reflect the forecasting skill of each individual model over a training period. To ensure that $p(\tilde{y}_t|f_{1t}, \ldots, f_{Kt})$ represents a proper distribution, the BMA weights are restricted to the simplex, $\Delta^{K-1} = \{\mathbf{w}|w_i \geq 0, i = 1, \ldots, K\}$ and assumed to add up to 1, $\sum_{k=1}^{K} w_k = 1$. Note that this assumption has been relaxed by *Vrugt and Robinson* [2007] and *Diks and Vrugt* [2010].

[4] The original BMA method described by *Raftery et al.* [2005] assumes that the conditional pdf, $g_k(\cdot)$ of the different ensemble members can be approximated by a normal distribution centered at a linear function of the original forecast, $a_k + b_k f_{kt}$ and standard deviation σ, that essentially conveys the predictive uncertainty of each individual forecast

$$\tilde{y}_t|f_{kt} \sim N(a_k + b_k f_{kt}, \sigma^2). \quad (2)$$

The values for a_k and b_k are bias-correction terms that are derived by simple linear regression of $\tilde{\mathbf{Y}}_n$ on f_{kn} for each of the K ensemble members. This (global) forecast correction removes long-term prediction bias, and is necessary to receive adequate performance. The BMA predictive mean at any given time t can be computed as

$$E[\tilde{y}_t|f_{1t}, \ldots, f_{Kt}] = \sum_{k=1}^{K} w_k(a_k + b_k f_{kt}), \quad (3)$$

which is a deterministic forecast whose predictive performance can be compared with the individual forecasts in the ensemble, or with the ensemble mean. The variance of this prediction, $\mathrm{var}[\tilde{y}_t|f_{1t}, \ldots, f_{Kt}]$, is derived from

$$\mathrm{var}[\tilde{y}_t|f_{1t}, \ldots, f_{Kt}] = \sum_{k=1}^{K} w_k \left((a_k + b_k f_{kt}) - \sum_{l=1}^{K} w_l(a_l + b_l f_l)\right)^2 + \sigma^2. \quad (4)$$

This prediction uncertainty is made up of two separate terms, the first representing the ensemble spread, and the second representing the within-ensemble forecast variance.

[5] The developments considered thus far have assumed that each ensemble member has a similar variance, irrespective of forecast skill. This seems rather difficult to justify in practice. An alternative, and perhaps more appealing approach, would therefore be to vary σ^2 among the different members of the ensemble. This requires some minor modifications to the BMA methodology, the most important of which is that the last term on the right-hand side of equation (4), needs to be replaced with $\sum_{k=1}^{K} w_k \sigma_k^2$. We will later show that using this method has only a minor effect on the BMA results.

[6] The assumption of a Gaussian conditional distribution of the individual ensemble members works well for variables whose conditional distribution is well approximated with a normal pdf. Examples of this include variables such as temperature and sea level pressure considered by *Raftery et al.* [2005]. Yet, this approach seems inappropriate for other variables such as wind speed and discharge, which are naturally bounded by zero. Indeed, this has inspired *Vrugt and Robinson* [2007] and *Sloughter et al.* [2010] to consider alternative formulations for $g_k(\cdot)$, but their Gamma pdf only marginally improved the results.

[7] We hypothesize that further improvements to the BMA method can be made if we relax the assumption of a preconceived and time-invariant form of $g_k(\cdot)$ considered hitherto in favor of a flexible, time-varying description of the conditional pdf. Arguably, this should further enhance the BMA results. This paper introduces the theory and concepts of this alternative BMA method, and demonstrates its usefulness and applicability by numerical simulation of the rainfall-runoff transformation using discharge data from three different catchments in the contiguous United States.

[8] The remainder of this paper is organized as follows: In section 2, we introduce the underlying theory and concepts of our approach. This is followed in section 3 with a detailed description of the numerical experiments, calibration data, and hydrologic model. In section 4, we compare the results of the original and proposed BMA method. Here we are especially concerned with forecast skill, and the average spread and statistical coherency of the 95% prediction uncertainty intervals. Finally, a summary with conclusions is presented in section 5.

2. Bayesian Model Averaging

2.1. Normal Conditional Distribution

[9] The standard BMA approach assumes that the conditional pdf, $g_k(\cdot)$ of each ensemble member, $k = 1, \ldots, K$ is time (space)-invariant, and adequately described with a normal distribution, $g_k(\cdot) \sim N(\cdot)$. The values of w_k, $k = 1, \ldots, K$ and σ^2 can then be derived by maximization of the following log-likelihood function, $\ell(\cdot)$

$$\ell(w_1, \ldots, w_K, \sigma^2|a_1 + b_1 f_{1n}, \ldots, a_k + b_k f_{Kn}, \tilde{\mathbf{Y}}_n) = \sum_{t=1}^{n} \log\left(\sum_{k=1}^{K} w_k g_k(\tilde{y}_t|a_k + b_k f_{kt})\right), \quad (5)$$

where n signifies the total number of measurements in the training data set. In the absence of a closed-form analytical solution that conveniently maximizes this equation, we resort to an iterative solution of $w_k, k = 1, \ldots, K$; and σ^2 using a Markov chain Monte Carlo (MCMC) simulation with the DiffeRential Evolution Adaptive Metropolis (DREAM) algorithm [*Vrugt et al.*, 2008b, 2009]. Explicit details of this approach within the context of BMA can be found in the work of *Vrugt et al.* [2008a] using numerical experiments with multimodel ensembles of surface temperature, sea level pressure, and streamflow forecasts. Note that equation (5) is easily extended to accommodate

a different variance, $\sigma_k, k = \{1, \ldots, K\}$ for each of the different individual predictors.

2.2. Flexible, Time-Varying Conditional Distribution

[10] The assumption of a time-invariant normal distribution of the conditional distribution, $g_k(\cdot)$ of the different ensemble members is statistically convenient, but often not borne out of the actual predictive uncertainty of the individual forecasts. The variance of this prediction uncertainty is typically larger, sometimes dramatically, than the calibrated BMA variance, σ^2 derived from equation (5). A second and, from the view of this paper, perhaps more important problem is that the actual prediction uncertainty of each ensemble member typically varies dynamically from one time step to the next, and most often deviates considerably from a normal distribution. We therefore posit that considerable improvements to the BMA method can be made if we allow the functional shape and size of $g_k(\cdot)$ to vary dynamically from observation to observation.

[11] One possible refinement of the BMA method introduced by *Vrugt and Robinson* [2007] is to allow for heteroscedasticity of the variance of $g_k(\cdot)$ using a simple linear dependency between the BMA variance, σ_k^2 and the actual biased-corrected forecast of each individual ensemble member

$$\sigma_{kt}^2 = c_k(a_k + b_k f_{kt}), \tag{6}$$

where the parameter c_k signifies the slope of this relationship. This approach does not require a direct estimation of σ_k^2. Instead, the BMA variance of each ensemble member is estimated indirectly from the calibration of $c_k; k = 1, \ldots, K$ using a MCMC simulation with DREAM. Simulation experiments presented by *Vrugt and Robinson* [2007] using daily discharge data demonstrated that a heteroscedastic BMA error variance not only reduced the l-d-ahead forecast error of the BMA mean, but also increased the sharpness of the 95% prediction uncertainty intervals. Yet, the conditional pdf, $g_{kt}(\cdot)$ (now time-variant due to its dependence on f_{kt}) is still assumed to be adequately described with a normal distribution. To relax this assumption, we follow a recommendation made in our previous work [*Vrugt and Robinson*, 2007] and derive $g_{kt}(\cdot)$ using particle filtering and Gaussian mixture modeling. We summarize this approach in sections 2.2.1 and 2.2.2.

2.2.1. Particle Filtering Using Particle-DREAM

[12] If we assume that the K forecasts of the ensemble are generated with a dynamical model, then we can derive $g_{kt}(\cdot); t = 1, \ldots, n$ directly using particle filtering (for further illustration, see also the tutorials by A. Doucet and A. M. Johansen (A tutorial on particle filtering and smoothing: Fifteen years later, 2008, available at http://www.cs.ubc.ca/~arnaud/doucet_johansen_tutorialPF.pdf) and applications in an environmental and hydrological context by, e.g., *Moradkhani et al.* [2005]; *van Leeuwen* [2009]; *Rings et al.* [2010]; *Vrugt et al.* [2012]). To help explain this procedure, lets write the underlying (nonlinear) model in a state-space formulation

$$\mathbf{x}_{t+1} = \Phi(\mathbf{x}_t, \theta, \mathbf{u}_t) + \mathbf{q}_{t+1}, \tag{7}$$

where $\Phi(\cdot)$ is the (nonlinear) model operator expressing the state transition in response to forcing data $\mathbf{u}_t$, model parameters, θ and state variables, $\mathbf{x}_t$. In the remainder of this paper, we assume that θ is fixed and consists of d parameter values, $\theta \in \mathbb{R}^d$ and that the state space $\{\mathbf{x}_t; t = 1, \ldots, n\}$ is of fixed dimension Ω, $\mathbf{x}_t \in \mathbb{R}^\Omega$. The variable $\mathbf{q}_t \in \mathbb{R}^\Omega$ represents errors in the model formulation, which is typically ignored in classical model calibration studies.

[13] The measurement operator, $\hbar(\cdot)$, defines the observation process and projects the model states, $\mathbf{x}_{t+1}$ to the model output, f_{t+1},

$$f_{t+1} = \hbar(\mathbf{x}_{t+1}, \phi) + v_t, \tag{8}$$

where $v_t \in \mathbb{R}^1$ denotes the measurement error, $v_t \sim N(0, \sigma_v)$, and any additional measurement variables are stored in ϕ.

[14] If we assume that the prior state pdf $p(\mathbf{x}_0)$ is available, then we can use the Chapman-Kolmogorov equation [*Jazwinski*, 1970] to derive the evolving state distribution, $p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1})$ at time t,

$$p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1}) = \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\tilde{y}_1, \ldots, \tilde{y}_{t-1}) \mathrm{d}\mathbf{x}_{t-1}, \tag{9}$$

where $p(\mathbf{x}|\mathbf{x}_{t-1})$ denotes the time evolution of the model states computed using the nonlinear model operator of equation (7). From this forecast density we can simply derive $p(f_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1})$ using the measurement operator $\hbar(\cdot)$. This is the distribution we are actually interested in, more of which will be discussed later.

[15] After this prediction step, the observation $\tilde{y}_t$ becomes available, and the forecast density can be updated via Bayes rule:

$$p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_t) = \frac{p(\tilde{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1})}{p(\tilde{y}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1})}. \tag{10}$$

This analysis density is conditioned on the current observation, and hence differs from the original forecast density. The first term in the numerator, $p(\tilde{y}_t|\mathbf{x}_t^i)$ measures how well $\mathbf{x}_t$ predicts the next observation $\tilde{y}_t$:

$$p(\tilde{y}_t|\mathbf{x}_t) = \frac{1}{\sqrt{2\pi\sigma_v^2}} \exp\left[-\frac{1}{2}\sigma_v^{-2}(\hbar(\mathbf{x}_t, \phi) - \tilde{y}_t)^2\right], \tag{11}$$

where $\hbar(\cdot)$ is the measurement operator of equation (8). The normalizing constant in the denominator of equation (10) follows from,

$$p(\tilde{y}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1}) = \int_{\mathbf{x}_t} p(\tilde{y}_t|\mathbf{x}_t) p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_{t-1}) \mathrm{d}\mathbf{x}_t. \tag{12}$$

If we substitute equations (9) and (12) in equation (10) we derive the following expression for the new posterior, $p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_t)$ after assimilating observation $\tilde{y}_t$:

$$p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_t) = \frac{p(\tilde{y}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\tilde{y}_1, \ldots, \tilde{y}_{t-1}) \mathrm{d}\mathbf{x}_{t-1}}{\int_{\mathbf{x}_t} (p(\tilde{y}_t|\mathbf{x}_t) \int_{\mathbf{x}_{t-1}} p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\tilde{y}_1, \ldots, \tilde{y}_{t-1}) \mathrm{d}\mathbf{x}_{t-1}) \mathrm{d}\mathbf{x}_t}. \tag{13}$$

[16] A key task that remains is to implement equations (9)–(12) on a digital computer. If the functions $\Phi(\cdot)$ and $h(\cdot)$ are linear, and $\mathbf{q}_t$ and v_t are Gaussian, the Kalman filter [*Kalman*, 1960] finds the exact filtering distribution. For nonlinear and non-Gaussian problems, the multidimensional integration in the denominator of equation (13) cannot be computed analytically, rendering an exact solution of $p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_t)$ impossible. For such cases, the extended Kalman filter (EKF) can be used, but this approach is quite unstable if the model operator is strongly nonlinear [*Evensen*, 1994; *Miller et al.*, 1994].

[17] We therefore resort to a Monte Carlo simulation, and approximate the evolving state distribution using an ensemble of different trajectories. The idea is to represent $p(\mathbf{x}_t|\tilde{y}_1, \ldots, \tilde{y}_t)$ in equation (13) by a set of P different trajectories, also called particles. Many contributions to the statistical and modeling literature have demonstrated unequivocally that particle filters are prone to sample degeneracy in which an increasing number of particles are exploring unproductive parts of the state space and assigned a negligible (zero) weight. A recent paper by *Vrugt et al.* [2012] presented the theory and simulation results of an alternative Bayesian filter that maintains adequate particle diversity. This Particle-DREAM filter is inspired by recent developments in particle Markov chain Monte Carlo (MCMC) sampling [*Andrieu et al.*, 2010] and combines the strengths of sequential Monte Carlo sampling and MCMC simulation with DREAM [*Vrugt et al.*, 2008b, 2009] to continuously relinquish bad trajectories and avoid sample impoverishment. Numerical experiments using the Lorenz attractor, the Lorenz96 model, and a rainfall-runoff model have shown that Particle-DREAM requires relatively few particles to work well in practice and provides important insights into the information content of the calibration data and nonstationarity of model parameters.

[18] The underlying premise of this paper is that the forecast density, $p(f_{kt}|\tilde{y}_1, \ldots, \tilde{y}_{t-1})$ of the k-th ensemble member derived with Particle-DREAM would be a desirable choice for $g_{kt}(\cdot)$ in the BMA methodology. This forecast distribution not only appropriately summarizes predictive uncertainty, but also removes the need to a priori specify the functional shape of $g_{kt}(\cdot)$. Indeed, $p(f_{kt}|\tilde{y}_1, \ldots, \tilde{y}_{t-1})$ might be multivariate, non-Gaussian, and multimodal, and hence deviate considerably from any traditional distribution currently used in BMA. Yet, if we adopt this approach and use the forecast density as an approximation to $g_{kt}(\cdot)$, then one hurdle remains, and that is that we only have (discrete) samples of $g_{kt}(\cdot)$. Without a continuous distribution it becomes rather cumbersome to derive the associated BMA weights of the individual ensemble members. We therefore postprocess the n forecast densities of each k-th ensemble member, $p(f_{kt}|\tilde{y}_1, \ldots, \tilde{y}_{t-1}); t = 1, \ldots, n$ by fitting a different Gaussian mixture to each of the individual (marginal) histograms. Section 2.2.2 explains this approach in more detail.

2.2.2. Gaussian Mixture Modeling to Provide the Distributional Form of $g_{kt}(\cdot)$

[19] The cross-entropy (CE) algorithm of *Rubinstein and Kroese* [2004] is used to create a mixture of Gaussian distributions [*Botev and Kroese*, 2004] for each ensemble member and each time step. We start with a single ($J = 1$) normal distribution, and estimate the mean μ^{mix} and standard deviation σ^{mix} of this distribution from the $P = 250$ discrete samples created with Particle-DREAM using a standard likelihood function. Then we sequentially add another component (normal distribution), and (re-)estimate the mean and standard deviation of each individual Gaussian distribution. We continue this process until the relative improvement in fit falls below 1%. We denote this final mixture distribution of ensemble member k at time t with $\pi_{kt} = \sum_{j=1}^{J} p_{ktj} N(\mu_{ktj}^{\text{mix}}, \sigma_{ktj}^{\text{mix}})$, where p_{ktj} denotes the normalized posterior probability (or weight) of the j-th distribution within the mixture, $\sum_{j=1}^{J} p_{ktj} = 1$. Note that the total number of constituent Gaussian distributions, J, can vary dynamically with time and also between different ensemble members.

[20] To illustrate our joint particle filtering and Gaussian mixture modeling framework, please consider Figure 1 that presents histograms of the Particle-DREAM-derived forecast density at days 214, 215, 323, and 403 for one of the models considered in our ensemble. The quantity of interest is river discharge, details of which will be presented in section 3. In Figure 1, the red lines represent the corresponding fit of the mixture distribution. Note that the forecast density not only varies dynamically between the different days, but also deviates considerably from a normal distribution. This provides evidence for our claim that a time-invariant Gaussian conditional pdf seems inappropriate to fully capture the discharge dynamics. A better compliance of the BMA

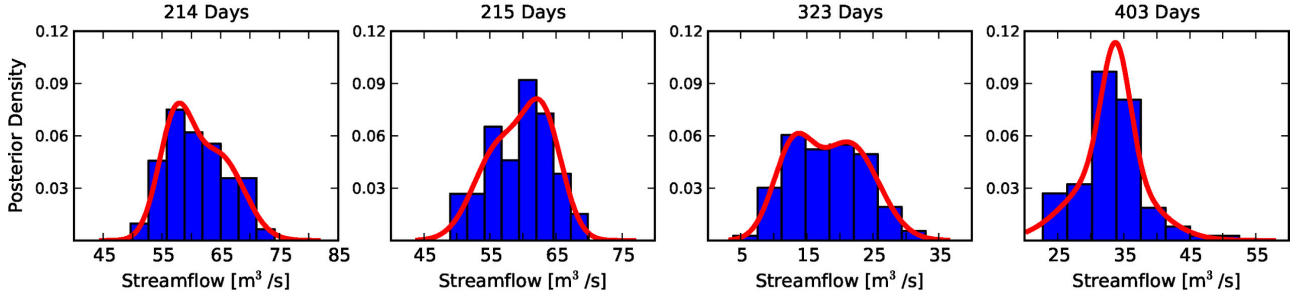

Figure 1. Leaf River: Histograms of the predictive densities derived with Particle-DREAM at days 214, 215, 323, and 403 of the training data period. The fit of the Gaussian mixture distribution is illustrated with the solid red line.

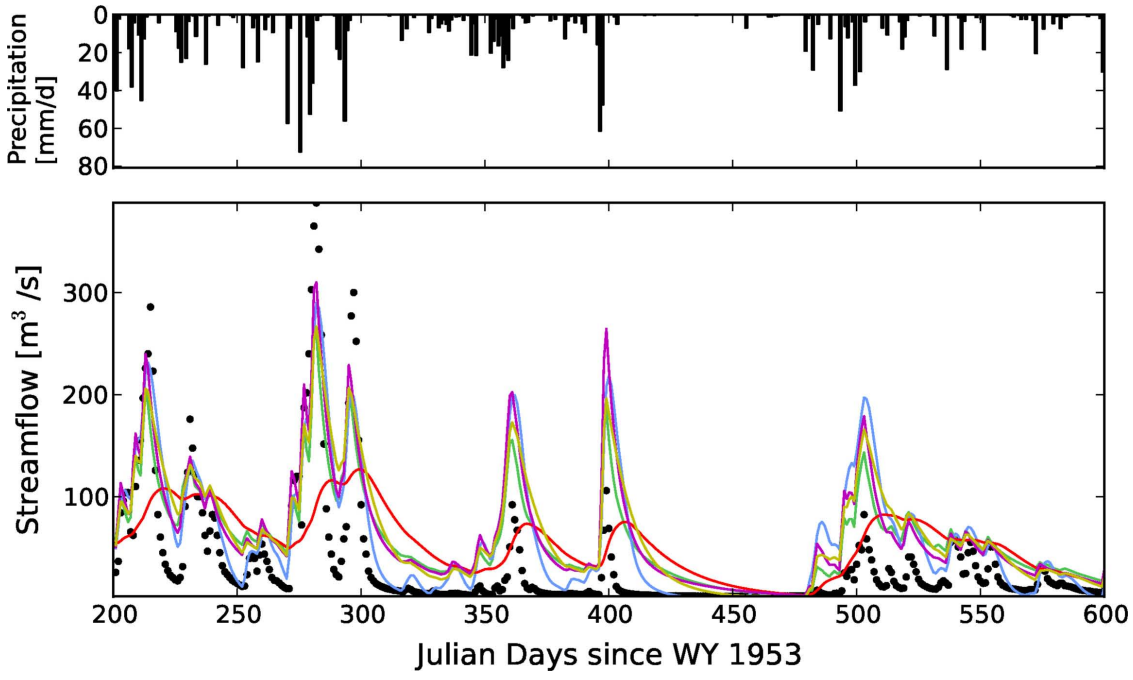

Figure 2. Leaf River: The five deterministic model forecasts (color coded) for a selected time period of the training period. The different lines represent the individual model forecasts and the black dots refer to the observed streamflow values.

model and actual data should be achievable if we use a flexible, time variable, description of $g_{kt}(\cdot)$.

[21] Now that we have the mixture distribution of each individual forecast at each time, we can derive the associated BMA weights of each individual ensemble member. We estimate $w_k; k = 1, \ldots, K$ by maximizing the following log likelihood function, $\ell(\cdot)$,

$$\begin{aligned}\ell(w_1, \ldots, w_K) &= \sum_{t=1}^{n} \log\left(\sum_{k=1}^{K}[w_k \pi_{kt}]\right) \\ &= \sum_{t=1}^{n} \log\left(\sum_{k=1}^{K}\left(w_k \sum_{j=1}^{J}[p_{ktj} N(\mu_{ktj}^{\text{mix}}, \sigma_{ktj}^{\text{mix}})]\right)\right),\end{aligned} \tag{14}$$

using MCMC simulation with DREAM.

3. Model, Historical Data, and Ensemble Generation

[22] We illustrate the usefulness and applicability of the joint particle filtering, Gaussian mixture modeling, and BMA framework, by application to rainfall-discharge modeling using historical data from the Leaf River, French Broad, and Guadalupe catchments in the United States. Five years of daily discharge data from each of the three catchments were used for BMA training, followed by a similar length data set to test the performance of the original and revised BMA method during an independent evaluation period. The watershed modeling toolbox of *Schoups et al.* [2010] was used to create a single watershed model, $\Phi(\cdot)$ that describes the precipitation-discharge transformation using four state variables for three different fast flow and one slow flow reservoir. An ensemble of five different members was created by randomly sampling the eight calibration parameters from their prior distribution. This approach differs somewhat from previous publications in that we used a single model for ensemble generation. This simplifies the analysis somewhat, but is sufficient to illustrate the main findings of this paper. Numerical experiments with structurally different watershed models provided very similar results (not shown herein).

[23] Note also that we purposely use an ensemble of uncalibrated model predictions. This not only best highlights the advantages of the proposed BMA approach, but also illustrates the gains that can be achieved with implementation of the theory and concepts presented herein in operational forecasting systems, many of which work with parameter-rich and CPU intensive simulation models that are computationally too demanding to calibrate directly against available observations. Examples include weather, hydrogeological, and global-scale hydrologic models. Our previous work [*Vrugt and Robinson*, 2007] used a calibrated ensemble of eight different watershed models with thr main conclusion that the default BMA approach cannot achieve a performance matching that of the ensemble Kalman filter.

[24] Before applying the linear bias correction outlined above, the measured discharge data, $\tilde{\mathbf{Y}}_n = \{\tilde{y}_1, \ldots, \tilde{y}_n\}$ and corresponding ensemble forecasts, $f = \{f_{1t}, \ldots, f_{kt}; t = 1, \ldots, n\}$ of each individual watershed were preprocessed using a Box-Cox power transformation [*Box and Cox*, 1964],

$$z_t^{BC} = \begin{cases} [z_t^{\lambda} - 1]/[\lambda(\text{GM}(\tilde{\mathbf{Y}}_n))^{\lambda-1}] & \text{if } \lambda \neq 0 \\ \text{GM}(\tilde{\mathbf{Y}}_n)\log(z_t) & \text{if } \lambda = 0 \end{cases}, \tag{15}$$

where z_t^{BC} denotes the transformed observation (model prediction) at time t, z_t is either the measured discharge $\tilde{y}_t$ or corresponding ensemble forecast $f_{kt}; k = \{1, \ldots, K\}$, $\mathrm{GM}(\mathbf{Y}_n)$ denotes the geometrical mean of the measured data, and λ signifies a transformation exponent, separately derived for each individual catchment using a MCMC simulation with DREAM. The maximum likelihood values of λ varied from 0.075 (Leaf River) to 0.16 (French Broad) and 0.21 (Guadalupe). This transformation was deemed necessary to remove heteroscedasticity [*Sorooshian and Dracup*, 1980]. The normal quantile transform (NQT) constitutes an alternative approach to enforce the normality of the data and forecasts [*Moran*, 1970; *Kelly and Krzysztofowicz*, 1997; *Montanari and Brath*, 2004]. Yet this approach posed some problems with extrapolation beyond the maximum measured discharge, which was necessary when calculating the error statistics of the particle filter in the original streamflow space.

[25] Section 4 discusses the main results of this paper. We present the results for three different BMA cases, each using the Box-Cox transformed discharge data. The first case uses the original transformed and bias-corrected model forecasts with a normal, time-invariant conditional pdf of the different ensemble members. This summarizes the results of the default BMA approach, "$\mathrm{BMA}_{(\mathrm{D})}$." The second case, hereafter referred to as "$\mathrm{BMA}_{(\mathrm{PF})}$," uses the median filter predictions derived with Particle-DREAM as the individual BMA forecasts and a normal (time-invariant) conditional distribution for each of the five different ensemble members. The forecasts of this ensemble are derived using sequential state updating, and hence in theory should exhibit a better predictive performance as the original unfiltered forecasts. Finally, the last approach, "$\mathrm{BMA}_{(\mathrm{PFM})}$" is similar to the "$\mathrm{BMA}_{(\mathrm{PF})}$," but uses the fitted Gaussian mixture distribution of the forecast density as a conditional distribution. This results in a time-varying, and free-form description of $g_{kt}(\cdot)$.

[26] In all of our calculations with Particle-DREAM, the measurement error, σ_v was assumed to be 10% of the actual measured discharge. This value is consistent with expert knowledge, and consistent with results of a nonparametric measurement error estimator [*Vrugt et al.*, 2005]. The model error, $\mathbf{q}_t$ was assumed to be adequately described with a normal distribution with error standard deviation of the states automatically tuned so that the filter receives an adequate performance. Details of this are beyond the scope of the current paper, and can be found in the work of *Vrugt et al.* [2005].

4. Results and Discussion

[27] To provide insights into the properties of the ensemble, Figure 2 presents the deterministic forecasts of the five different ensemble members for the Leaf River watershed for a representative period in 1953. The black dots represent the measured discharge data and the color-coded lines denote the corresponding predictions of the five different watershed model parameterizations. The original model ensemble does not properly track the observational data, and shows significant prediction bias. Peak and low flow are particularly poorly described. Indeed, a linear bias-correction (after the Box-Cox transformation) is warranted to improve the predictive performance of each of the ensemble members.

[28] The respective plots for the French Broad and Guadalupe watersheds are shown in Figures 3 and 4. The model predictions are more realistic for these two rivers, and the ensemble spread tends to better capture the measured discharge dynamics. Note that the Guadalupe basin is rather dry, and characterized with a few heavy precipitation and flash flooding events that are difficult to model in practice. Direct runoff (overland flow) is only poorly described in the watershed model.

[29] We now illustrate the results of the particle filter in the original discharge space. The results of this are presented in Figures 5 (Leaf River), 6 (French Broad), and 7 (Guadalupe). Each individual panel separately presents the results of each individual ensemble member. The impact of sequential state updating becomes immediately apparent. The median predictions of the forecast densities derived with Particle-DREAM (color-coded lines) track much better the observed discharge (solid dots) for each individual watershed, with 95% prediction uncertainty intervals (color-coded regions) that appear, perhaps, rather large but appropriately encapsulate the measured data. Even the flash-flooding events in the Guadalupe basin are now reasonably predicted. The ability of the particle filter to continuously update the state variables, allows for a much better compliance between the measured and predicted discharge values, even with parameter values that are randomly sampled from their prior distributions and deemed inadequate. We conclude that the particle filter is a useful preprocessing step prior to BMA. Yet this requires significant computational efforts, and perhaps is not easy to implement for some models, particularly those that resolve spatially distributed processes, and are hence computationally demanding.

[30] Tables 1, 2, and 3 summarize the BMA weights for each of the five ensemble members for the three different watersheds considered herein. The column headings have been defined previously. The heading $\mathrm{BMA}_{(\mathrm{D})}$ lists the results of the default BMA approach using the original Box-Cox transformed and bias-corrected discharge forecasts with a normal conditional distribution. The other two columns, $\mathrm{BMA}_{(\mathrm{PF})}$ and $\mathrm{BMA}_{(\mathrm{PFM})}$ summarize the BMA results with median forecasts of the particle filter, but differ in that $\mathrm{BMA}_{(\mathrm{PF})}$ uses a normal (time-invariant) conditional distribution and $\mathrm{BMA}_{(\mathrm{PFM})}$ uses the Gaussian mixture distribution of the forecast density as approximation to $g_{kt}(\cdot)$.

[31] The original BMA results ($\mathrm{BMA}_{(\mathrm{D})}$) tends to place the weights on just two or three members of the ensemble. The other forecasts of the ensemble receive a nearly zero weight and do not play any role in the BMA model. This is different when the forecasts are derived from the particle filter. The weights are more homogeneously distributed among the different ensemble members (except for the Leaf River), simply because the individual forecasts closely track the observed discharge data and exhibit a similar predictive performance. This is perhaps a desirable finding, as perhaps each constituent member brings along additional information and detail about the rainfall-runoff process.

[32] The results presented thus far do not convey any information about the predictive performance of the

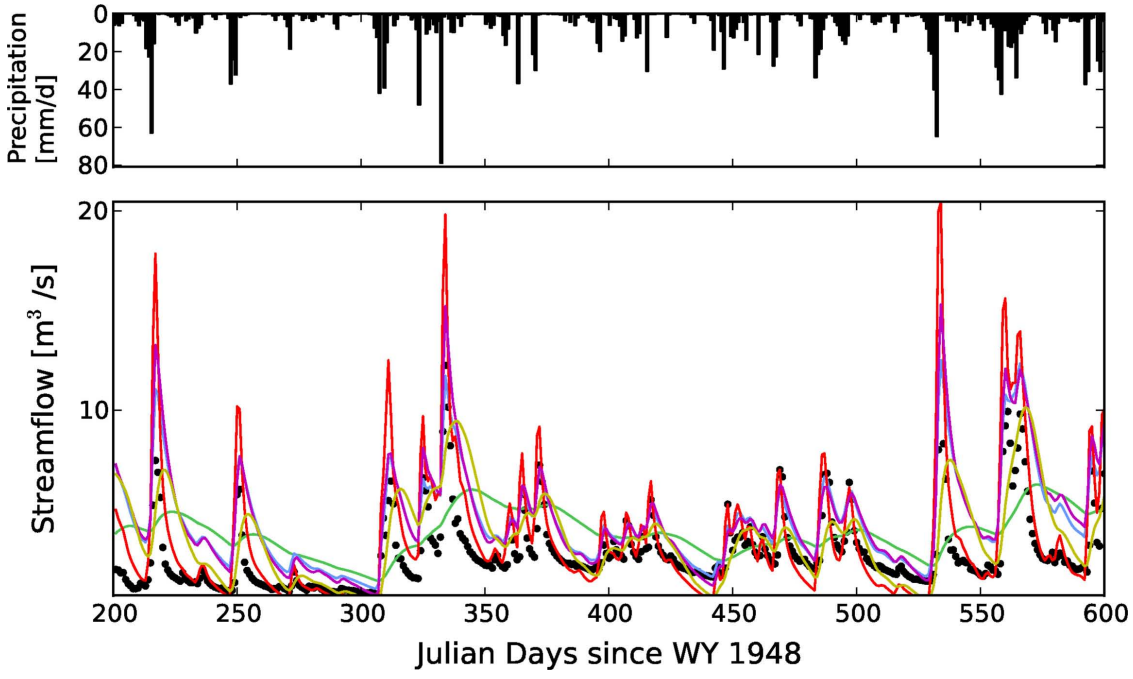

Figure 3. French Broad: The five deterministic model forecasts (color coded) for a selected time period of the training period. The different lines represent the individual model forecasts and the black dots refer to the observed streamflow values.

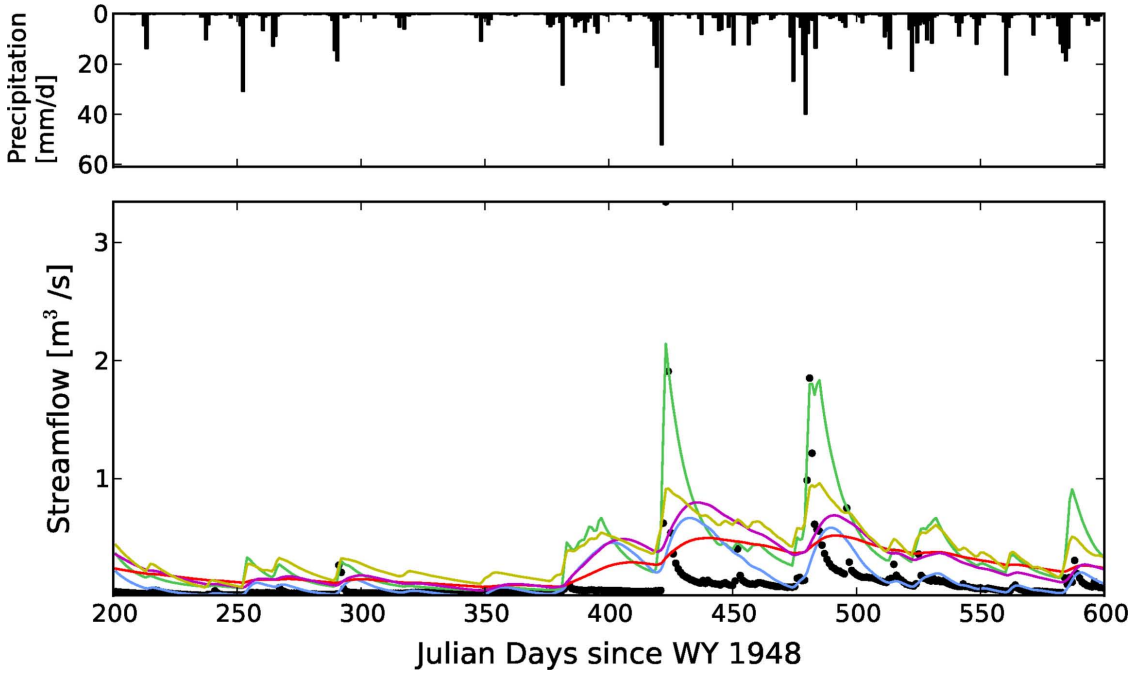

Figure 4. Guadalupe: The five deterministic model forecasts (color coded) for a selected time period of the training period. The different lines represent the individual model forecasts and the black dots refer to the observed streamflow values.

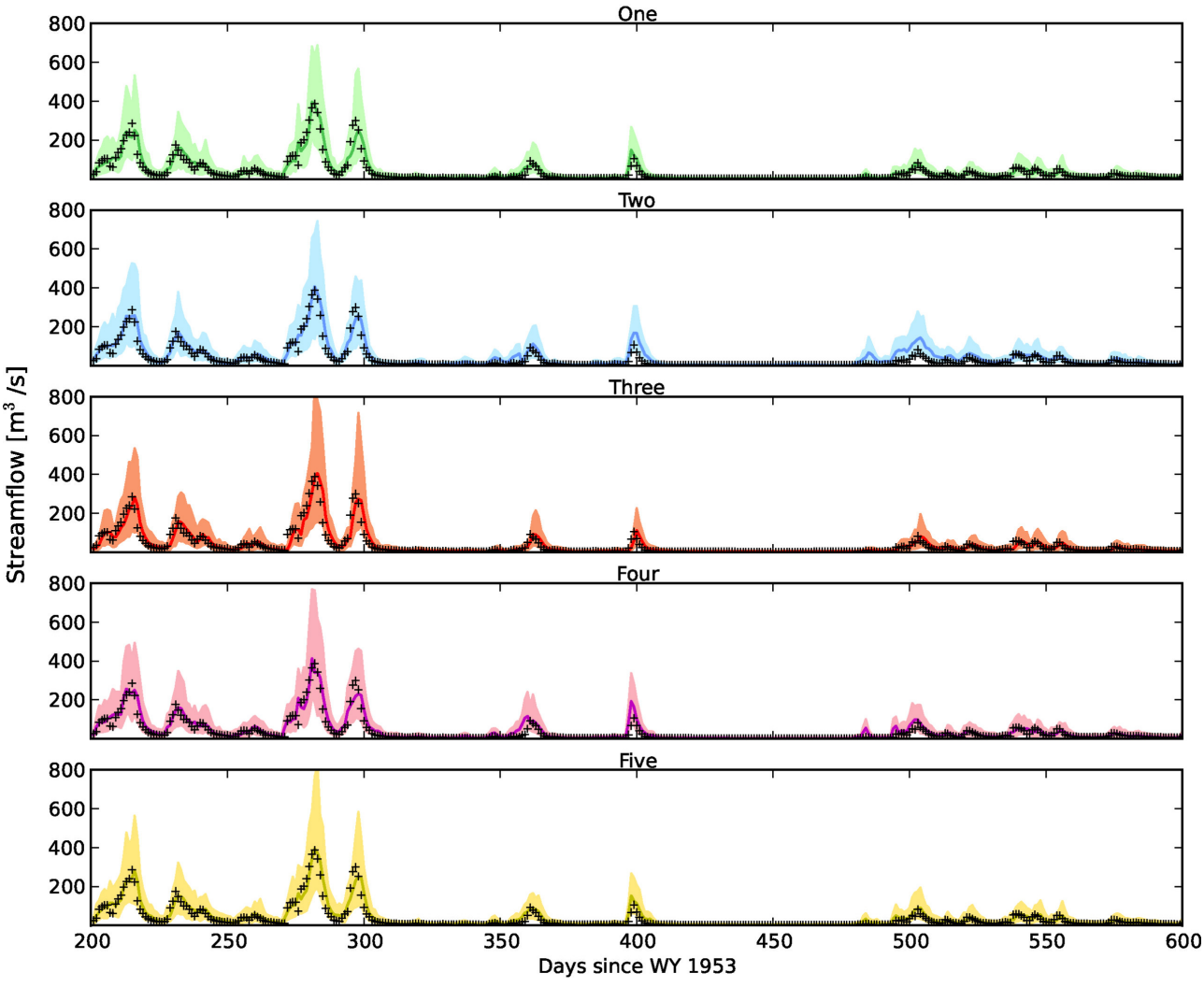

Figure 5. Leaf River: The effect of particle filtering on the discharge forecasts. Each horizontal panel plots the results for a different ensemble member; the + (plus) symbols represent the observed discharge data, the solid line the median Particle-DREAM prediction, and the green, blue, orange, purple, and yellow regions denote the corresponding 95% uncertainty ranges.

different ensemble members and BMA models. Tables 4 (Leaf River), 5 (French Broad), and 6 (Guadalupe) summarize for each different watershed the average l-d-ahead forecast error of the different ensemble members and different BMA approaches for the calibration and evaluation period. We also list the average spread of the 95% prediction uncertainly intervals, and the percentage of discharge observations that are contained within this interval. The most important findings can be summarized as follows.

[33] In the first place, notice that particle filtering has substantially improved the predictive performance of each individual ensemble member. Sequential state updating with Particle-DREAM substantially reduces (with a few exceptions) the average prediction error (root-mean-square error [RMSE]) of each forecast (1→5) of the ensemble. This improvement is most substantial for the Leaf River and French Broad watersheds, and is observed during both the calibration and evaluation period. Note that $BMA_{(PF)}$ and $BMA_{(PFM)}$ differ only in their conditional distribution used to assess the prediction uncertainty of the BMA model, and hence list a similar prediction error of the forecasts of the ensemble. A lower RMSE for the Guadalupe Basin during evaluation is connected to the fact that the RMSE is very sensitive to the fitting of rare rainfall events, so that a lower RMSE can just be related to less-extreme streamflow peaks during the evaluation period.

[34] A second finding is that the forecast error of the BMA model has significantly decreased with particle filtering of the original forecasts. The RMSE of $BMA_{(PF)}$ and $BMA_{(PFM)}$ is significantly lower than the RMSE of default BMA, $BMA_{(D)}$ using the unfiltered forecasts. This is perhaps not surprising and the immediate effect of the state-updating step in $BMA_{(PF)}$ and $BMA_{(PFM)}$. Note that the RMSE of $BMA_{(PFM)}$ is somewhat lower than the RMSE of $BMA_{(PF)}$; but both prediction errors are similar to the RMSE values of the individual constituent ensemble

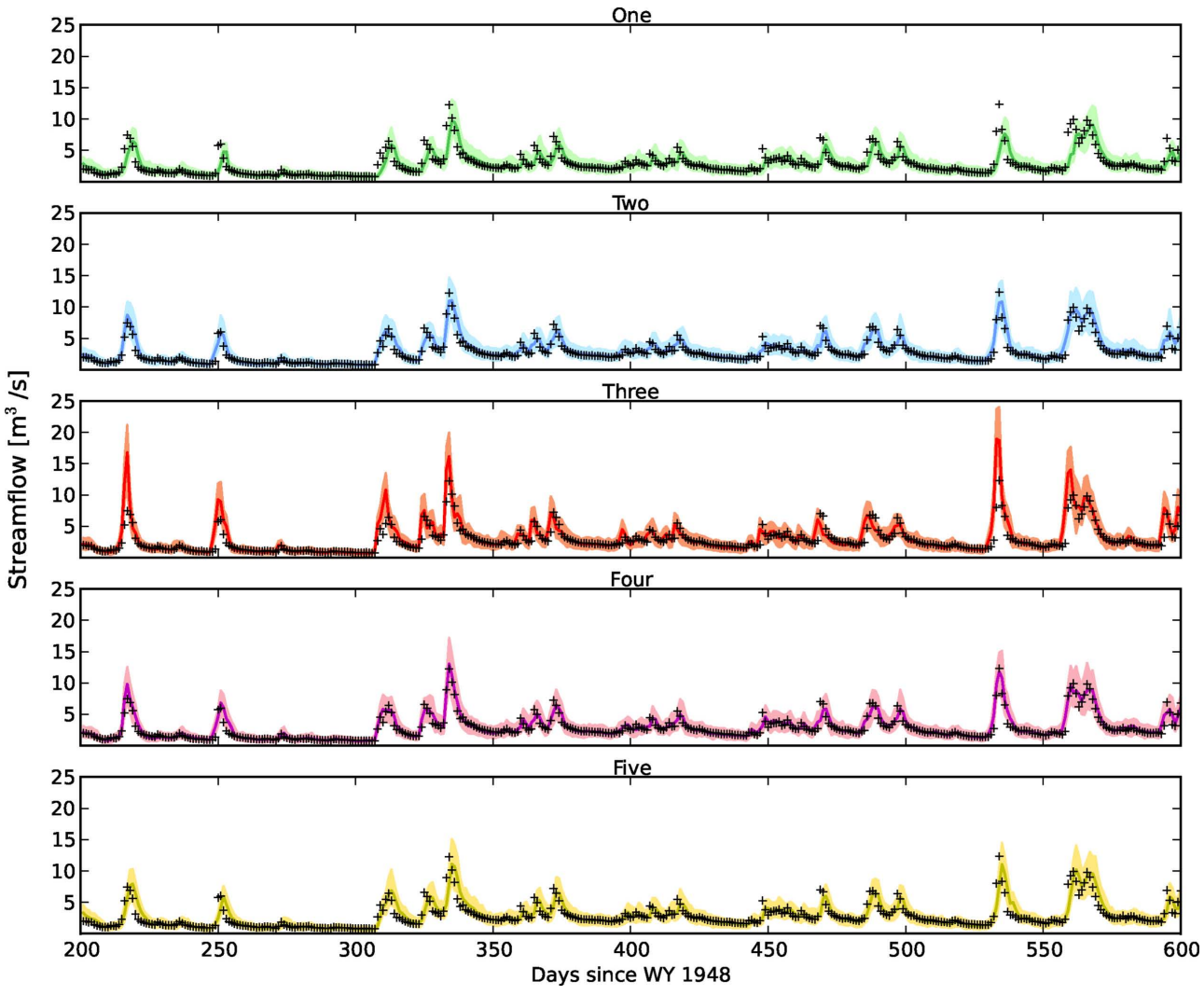

Figure 6. French Broad: The effect of particle filtering on the discharge forecasts. Each horizontal panel plots the results for a different ensemble member; the + (plus) symbols represent the observed discharge data, the solid line the median Particle-DREAM prediction, and the green, blue, orange, purple, and yellow regions denote the corresponding 95% uncertainty ranges.

members. Thus, postprocessing of the filtered forecast ensemble with BMA does not further reduce the average prediction error, irrespective of whether we are using a normal or time-varying and flexible conditional distribution.

[35] A third finding is that the average spread of the 95% prediction uncertainty ranges of the BMA model is substantially reduced (in most cases) when using the filtered forecasts. The sharpness of the predictive pdf has substantially increased with $BMA_{(PF)}$ and $BMA_{(PFM)}$. But $BMA_{(PF)}$ somewhat underestimates the actual prediction uncertainty, as the intervals do not contain the desired 95% of the discharge data. The overall best results are obtained with $BMA_{(PFM)}$. This method not only exhibits the best predictive performance from all three BMA approaches, but also adequately captures the expected percentage of observations at the 95% prediction uncertainty interval. We thus conclude that a flexible, time-varying conditional pdf of each of the model forecasts in the BMA method has desirable advantages. It is important to realize, however, that this approach can only be applied sequentially, and hence each time requires actual data of the quantity of interest to forecast into the future. The default BMA method on the contrary, receives rather poor performance, but can be used without hesitation once the BMA weights and variances have been determined from a training period. This has several practical advantages.

[36] Finally, we test whether the performance of $BMA_{(D)}$ and $BMA_{(PF)}$ could be further improved if we allow for individual variances of the normal conditional pdf for each ensemble member. The results are presented in Table 7. We limit our results to the Leaf River watershed as similar findings were observed for the other two basins. The weights of the different forecasts not only differ from their current values if we use individual variances for each of the conditional distributions of the ensemble, but the weights

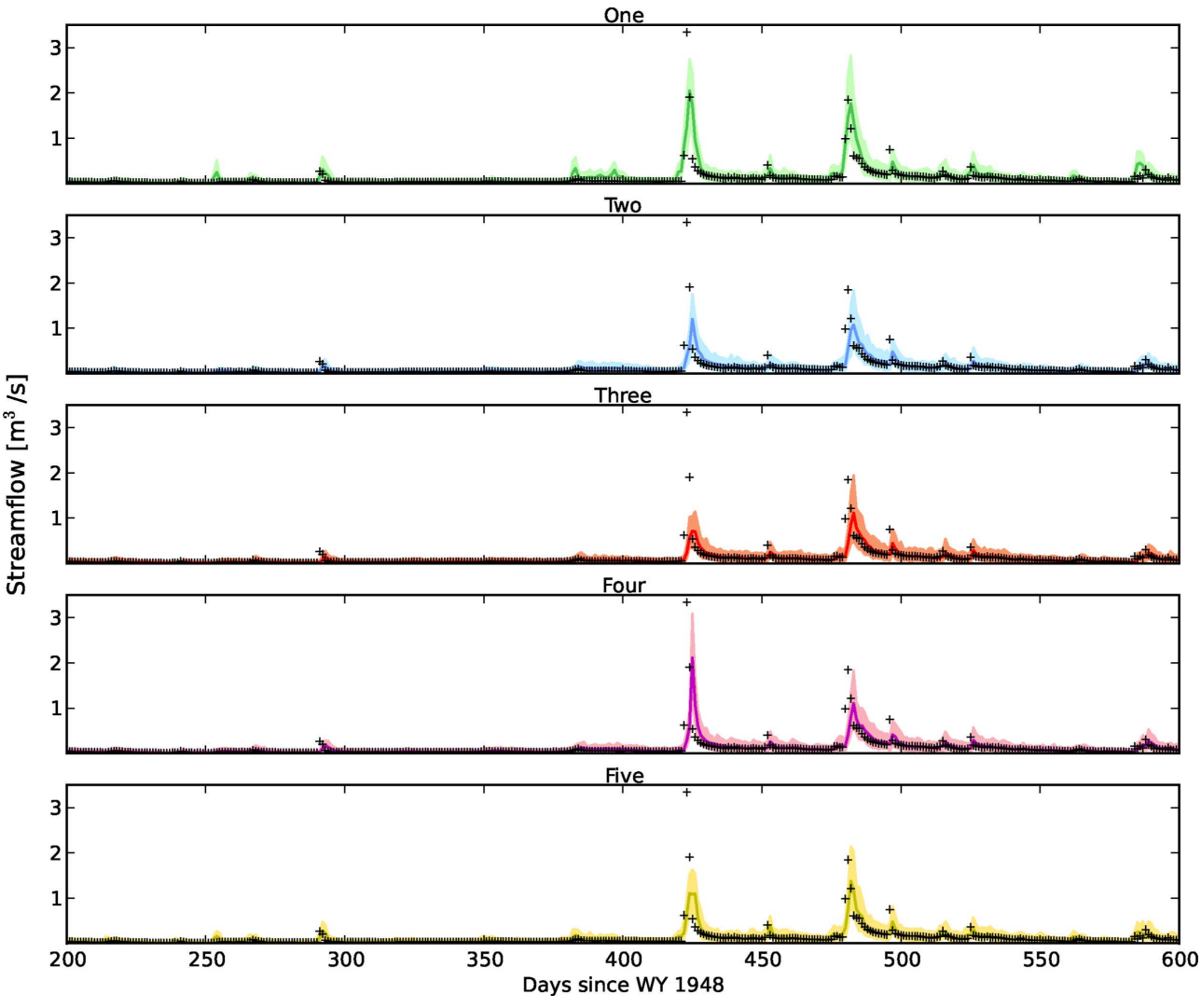

Figure 7. Guadalupe: The effect of particle filtering on the discharge forecasts. Each horizontal panel plots the results for a different ensemble member; the + (plus) symbols represent the observed discharge data, the solid line the median Particle-DREAM prediction, and the green, blue, orange, purple, and yellow regions denote the corresponding 95% uncertainty ranges.

are also distributed more evenly among the different ensemble members. This is most apparent for $BMA_{(PF)}$. Yet this finding does not appear to really affect the overall predictive performance of $BMA_{(D)}$ and $BMA_{(PF)}$. An individual variance for each constituent ensemble member in $BMA_{(D)}$ and $BMA_{(PF)}$ receives very similar discharge prediction errors and 95% uncertainty ranges as those derived previously with both methods using a single variance. This

Table 1. Leaf River: BMA Model Weights Derived for the Calibration Data Period for Each of the Three Different Cases Considered

	BMA Weight		
Model	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$
1	0.5189	0.4642	0.5335
2	0.0001	0.0003	0.0002
3	0.0037	0.4722	0.4643
4	0.4770	0.0632	0.0015
5	0.0004	0.0000	0.0004

Table 2. French Broad: BMA Model Weights Derived for the Calibration Data Period for Each of the Three Different Cases Considered

	BMA Weight		
Model	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$
1	0.1728	0.1976	0.1321
2	0.3471	0.2040	0.1829
3	0.4794	0.1950	0.0942
4	0.0005	0.3779	0.5808
5	0.0002	0.0255	0.0100

Table 3. Guadalupe: BMA Model Weights Derived for the Calibration Data Period for Each of the Three Different Cases Considered

	BMA Weight		
Model	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$
1	0.3582	0.1612	0.0444
2	0.0001	0.0645	0.1576
3	0.0027	0.6921	0.0162
4	0.0004	0.0152	0.5744
5	0.6385	0.0670	0.2074

Table 4. Leaf River: Summary Statistics of the Performance of the Individual Ensemble Members and BMA Model for the Different Cases Considered[a]

	Calibration			Evaluation		
Metric	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$
RMSE 1	39.48	20.55		54.26	33.19	
RMSE 2	44.39	23.09		60.02	37.53	
RMSE 3	47.95	25.71		72.99	43.84	
RMSE 4	41.98	20.72		55.29	34.06	
RMSE 5	45.45	21.06		61.14	35.57	
BMA RMSE	38.69	23.87	23.75	62.20	40.58	40.22
Spread	40.34	17.46	30.66	56.67	28.75	50.78
% contained	93.4	88.8	96.3	89.5	86.6	95.7

[a]Data is for the 5-yr calibration and evaluation period. The RMSE summarizes the average one-day-ahead discharge prediction error, and the spread and percent of observations contained refer to the average width of the 95% uncertainty ranges, and the percentage of discharge observations contained in this interval.

Table 5. French Broad: Summary Results of the Ensemble and BMA Models for the Calibration and Evaluation Period[a]

	Calibration			Evaluation		
Metric	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$
RMSE 1	1.77	1.04		1.80	0.92	
RMSE 2	1.78	0.64		1.90	0.59	
RMSE 3	1.65	1.08		1.68	1.07	
RMSE 4	1.94	0.62		2.01	0.56	
RMSE 5	1.67	0.93		1.77	0.92	
BMA RMSE	1.07	0.64	0.62	0.97	0.55	0.53
Spread	2.68	0.98	1.57	2.50	0.78	1.24
% contained	93.5	88.7	95.3	85.6	87.3	95.6

[a]For a description of each metric, please refer to Table 4.

Table 6. Guadalupe: Summary Results of the Ensemble and BMA Models for the Calibration and Evaluation Period[a]

	Calibration			Evaluation		
Metric	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$	$BMA_{(D)}$	$BMA_{(PF)}$	$BMA_{(PFM)}$
RMSE 1	0.80	0.56		0.50	0.34	
RMSE 2	0.82	0.77		0.44	0.38	
RMSE 3	0.82	0.80		0.46	0.39	
RMSE 4	0.86	0.79		0.49	0.37	
RMSE 5	0.81	0.75		0.45	0.36	
BMA RMSE	0.79	0.80	0.67	0.42	0.37	0.34
Spread	0.20	0.08	0.09	0.19	0.08	0.08
% contained	94.4	95.3	95.7	78.1	84.0	85.8

[a]For a description of each metric, please refer to Table 4.

Table 7. Leaf River: Comparison of $BMA_{(D)}$ and $BMA_{(PF)}$ With Single or Multiple Difference Variances? Unclear of the Conditional Distribution of the Individual Ensemble Members

	BMA weight			
	$BMA_{(D)}$		$BMA_{(PF)}$	
Model/Metric	Single	Individual	Single	Individual
1	0.5189	0.7966	0.4642	0.3739
2	0.0001	0.0003	0.0003	0.0000
3	0.0037	0.0173	0.4722	0.4201
4	0.4770	0.1857	0.0632	0.0547
5	0.0004	0.0002	0.0000	0.1513
BMA RMSE	38.69	39.07	23.87	24.05
Spread	40.34	37.44	17.46	15.18
% contained	93.4	93.4	88.8	84.9

conclusion is perhaps not surprising, and has been reported earlier for the default BMA approach [*Raftery et al.*, 2005; *Vrugt and Robinson*, 2007]. This concludes our numerical experiments.

5. Conclusions

[37] Bayesian model averaging has found widespread application and use for postprocessing of forecast ensembles of environmental system models. The standard BMA method assumes a normal, time-invariant distribution of $g_k(\cdot)$, the conditional pdf of the individual forecasts of the ensemble. In this paper, we relax this assumption and have introduced theory and concepts of a joint particle filtering and Gaussian mixture modeling framework to provide a flexible, time-variable description of $g_{kt}(\cdot)$. Simulation experiments using observed discharge data from the Leaf River, French Broad, and Guadalupe watersheds in the contiguous United States have demonstrated that this revised BMA method exhibits better predictive performance than the original default BMA method, with a spread of the 95% prediction uncertainty intervals that appropriately captures the desired percentage of observations. The Particle-DREAM and DREAM MCMC simulation codes used herein can be obtained from the second author upon request.

[38] **Acknowledgment.** Jasper A. Vrugt would like to acknowledge financial support from the LDRD project "Multilevel Adaptive Sampling for Multiscale Inverse Problems" of the Los Alamos National Laboratory.

References

Andrieu, C., A. Doucet, and R. Holenstein (2010), Particle Markov chain Monte Carlo methods, *J. R. Stat. Soc., Ser. B*, *72*, 269–342.

Barnston, A. G., S. J. Mason, L. Goddard, D. F. DeWitt, and S. E. Zebiak (2003), Multimodel ensembling in seasonal climate forecasting at IRI, *Bull. Am. Meteorol. Soc.*, *84*, 1783–1796, doi:10.1175/BAMS-84-12-1783.

Botev, Z., and D. P. Kroese (2004), Global likelihood optimization via the cross-entropy method with an application to mixture models, in *Proceedings of the 2004 Winter Simulation Conference*, edited by R. G. Ingalls, et al., pp. 529–535, Am. Stat. Assoc., Washington, D. C.

Box, G. E. P., and D. R. Cox (1964), An analysis of transformations, *J. R. Stat. Soc., Ser. B*, *26*, 211–252.

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*, *57A*, 234–252.

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

Gneiting, T., A. E. Raftery, A. H. Westerveld, and T. Goldman (2005), Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation, *Mon. Weather Rev.*, *133*, 1098–1118.

Grimitt, E. P., and C. F. Mass (2002), Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest, *Weather Forecasting*, *17*, 192–205.

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

Jazwinski, A. H. (1970), *Stochastic Processes and Filtering Theory*, 376 pp., Academic, N. Y.

Kalman, R. E. (1960), A new approach to linear filtering and prediction problems, *Trans. ASME J. Basic Eng.*, *82*, 35–45.

Kelly, K. S., and R. Krzysztofowicz (1997), A bivariate meta-Gaussian density for use in hydrology, *Stochastic Hydrol. Hydraul.*, *11*, 17–31.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendan (1999), Improved weather and seasonal climate forecasts from multimodel superensembles, *Science*, *258*, 1548–1550.

Miller, R. N., M. Ghil, and F. Gauthiez (1994), Advanced data assimilation in strongly nonlinear dynamical systems, *J. Atmos. Sci.*, *51*, 1037–1056.

Min, S., and A. Hense (2006), A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models, *Geophys. Res. Lett.*, *33*, L08708, doi:10.1029/2006GL025779.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis (1996), The ECMWF ensemble prediction system: Methodology and validation, *Q. J. R. Meteorol. Soc.*, *122*, 73–119.

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

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

Moran, P. A. P. (1970), Simulation and evaluation of complex water systems operations, *Water Resour. Res.*, *6*, 1737–1742.

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, doi:10.1175/BAMS-85-6-853.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*, 1155–1173.

Rajagopalan, B., U. Lall, and S. E. Zebiak (2002), Categorical climate forecasts through regularization and optimal combination of multiple GCM ensembles, *Mon. Weather Rev.*, *130*, 1792–1811.

Richardson, D. S. (2001), Measure of skill and value of ensemble prediction systems, their interrelationship and the effect of sample size, *Q. J. R. Meteorol. Soc.*, *127*, 2473–2489.

Rings, J., J. A. Huisman, and H. Vereecken (2010), Coupled hydrogeophysical parameter estimation using a sequential Bayesian approach, *Hydrol. Earth Syst. Sci.*, *14*, 545–556, doi:10.5194/hess-14-545-2010.

Rubinstein, R. Y., and D. P. Kroese (2004), *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, 320 pp., Springer, N. Y.

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

Sloughter, J. M., T. Gneiting, and A. E. Raftery (2010), Probabilistic wind speed forecasting using ensembles and Bayesian model averaging, *J. Am. Stat. Assoc.*, *105*, 25–35, doi:10.1198/jasa.2009.ap08615.

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

van Leeuwen, P. J. (2009), Particle filtering in geophysical systems, *Mon. Weather Rev.*, *137*(12), 4089–4114, doi:10.1175/2009MWR2835.1.

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.

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., C. G. H. Diks, and M. P. Clark (2008a), Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling, *Environ. Fluid Mech.*, *8*, 579–595 doi:10.1007/s10652-008-9106-3.

Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008b), Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation, *Water Resour. Res.*, *44*, W00B09, doi:10.1029/2007WR006720.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, J. M. Hyman, and D. Higdon (2009), Accelerating Markov chain Monte Carlo simulations by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numerical Simul.*, *10*, 273–290.

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, and G. Schoups (2012), Hydrologic data assimilation using Particle Markov chain Monte Carlo simulation: Theory, concepts and applications, *Adv. Water Resour.*, in press.

J. A. Huisman and H. Vereecken, Agrosphere, IBG-3, Forschungszenstrum Jülich, D-52425 Jülich, Germany.

J. Rings, Department of Land, Air and Water Resources, University of California, Davis, One Shields Ave., Davis, CA 95616, USA.

G. Schoups, Department of Water Management, Delft University of Technology, PO Box 5048, 2600 GA Delft, Netherlands.

J. A. Vrugt, Department of Civil and Environmental Engineering, University of California Irvine, 4130 Engineering Gateway, Irvine, CA 92697, USA. (jasper@uci.edu)