Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation

Jasper A. Vrugt,[1,2] Cajo J. F. ter Braak,[3] Martyn P. Clark,[4] James M. Hyman,[5] and Bruce A. Robinson[6]

Received 30 November 2007; revised 3 July 2008; accepted 8 September 2008; published 17 December 2008.

[1] There is increasing consensus in the hydrologic literature that an appropriate framework for streamflow forecasting and simulation should include explicit recognition of forcing and parameter and model structural error. This paper presents a novel Markov chain Monte Carlo (MCMC) sampler, entitled differential evolution adaptive Metropolis (DREAM), that is especially designed to efficiently estimate the posterior probability density function of hydrologic model parameters in complex, high-dimensional sampling problems. This MCMC scheme adaptively updates the scale and orientation of the proposal distribution during sampling and maintains detailed balance and ergodicity. It is then demonstrated how DREAM can be used to analyze forcing data error during watershed model calibration using a five-parameter rainfall-runoff model with streamflow data from two different catchments. Explicit treatment of precipitation error during hydrologic model calibration not only results in prediction uncertainty bounds that are more appropriate but also significantly alters the posterior distribution of the watershed model parameters. This has significant implications for regionalization studies. The approach also provides important new ways to estimate areal average watershed precipitation, information that is of utmost importance for testing hydrologic theory, diagnosing structural errors in models, and appropriately benchmarking rainfall measurement devices.

Citation: Vrugt, J. A., C. J. F. ter Braak, M. P. Clark, J. M. Hyman, and B. A. Robinson (2008), 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.

1. Introduction and Scope

[2] Hydrologic models, no matter how sophisticated and spatially explicit, aggregate at some level of detail complex, spatially distributed vegetation and subsurface properties into much simpler homogeneous storages with transfer functions that describe the flow of water within and between these different compartments. These conceptual storages correspond to physically identifiable control volumes in real space, even though the boundaries of these control volumes are generally not known. A consequence of this aggregation process is that most of the parameters in these models cannot be inferred through direct observation in the field, but can only be meaningfully derived by calibration against an input-output record of the catchment response. In this process the parameters are adjusted in such a way that the model approximates as closely and consistently as possible the response of the catchment over some historical period of time. The parameters estimated in this manner represent effective conceptual representations of spatially and temporally heterogeneous watershed properties.

[3] The traditional approach to watershed model calibration assumes that the uncertainty in the input-output representation of the model is attributable primarily to uncertainty associated with the parameter values. This approach effectively neglects errors in forcing data, and assumes that model structural inadequacies can be described with relatively simple additive error structures. This is not realistic for real world applications, and it is therefore highly desirable to develop an inference methodology that treats all sources of error separately and appropriately. Such a method would help to better understand what is and what is not well understood about the catchments under study, and help provide meaningful uncertainty estimates on model predictions, state variables and parameters. Such an approach should also enhance the prospects of finding useful regionalization relationships between catchment properties and optimized model parameters, something that is desirable, especially

[1] Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

[2] Institute for Biodiversity and Ecosystems Dynamics, University of Amsterdam, Amsterdam, Netherlands.

[3] Biometris, Wageningen University and Research Centre, Wageningen, Netherlands.

[4] NIWA, Christchurch, New Zealand.

[5] Mathematical Modeling and Analysis, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

[6] Civilian Nuclear Program Office, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.

0043-1397/08/2007WR006720

within the context of the Predictions in Ungauged Basins (PUB) initiative [*Sivapalan*, 2003].

[4] In recent years, significant progress has been made toward the development of a systematic framework for uncertainty treatment. While initial methodologies have focused on methods to quantify parameter uncertainty only [*Beven and Binley*, 1992; *Freer et al.*, 1996; *Gupta et al.*, 1998; *Vrugt et al.*, 2003], recent emerging approaches include state space filtering [*Vrugt et al.*, 2005; *Moradkhani et al.*, 2005a, 2005b; *Slater and Clark*, 2006; *Vrugt et al.*, 2006a], multimodel averaging [*Butts et al.*, 2004; *Georgakakos et al.*, 2004; *Vrugt et al.*, 2006b; *Marshall et al.*, 2006; *Ajami et al.*, 2007; *Vrugt and Robinson*, 2007b] and Bayesian approaches [*Kavetski et al.*, 2006a, 2006b; *Kuczera et al.*, 2006; P. Reichert and J. Mieleitner, Analyzing input and structural uncertainty of a hydrological model with stochastic, time-dependent parameters, unpublished manuscript, 2008] to explicitly treat individual error sources, and assess predictive uncertainty distributions. Much progress has also been made in the description of forcing data error [*Clark and Slater*, 2006], development of a formal hierarchical framework to formulate, build and test conceptual watershed models [*Clark et al.*, 2008], and algorithms for efficient sampling of complex distributions [*Vrugt et al.*, 2003; *Vrugt and Robinson*, 2007a; *Vrugt et al.*, 2008a] to derive uncertainty estimates of state variables, parameters and model output predictions.

[5] This paper has two main contributions. First, a novel adaptive Markov chain Monte Carlo (MCMC) algorithm is introduced for efficiently estimating the posterior probability density function of parameters within a Bayesian framework. This method, entitled differential evolution adaptive Metropolis (DREAM), runs multiple chains simultaneously for global exploration, and automatically tunes the scale and orientation of the proposal distribution during the evolution to the posterior distribution. The DREAM scheme is an adaptation of the shuffled complex evolution Metropolis (SCEM-UA) [*Vrugt et al.*, 2003] global optimization algorithm that has the advantage of maintaining detailed balance and ergodicity while showing good efficiency on complex, highly nonlinear, and multimodal target distributions [*Vrugt et al.*, 2008a]. Second, the applicability of DREAM is demonstrated for analyzing forcing error during watershed model calibration. *Vrugt et al.* [2008b] extended the work presented in this paper to include model structural error as well through the use of a first-order autoregressive scheme of the error residuals.

[6] The framework presented herein has various elements in common with the Bayesian total error analysis (BATEA) approach of *Kavetski et al.* [2006a, 2006b], but uses a different inference methodology to estimate the model parameters and rainfall multipliers that characterize and describe forcing data error. In addition, this method generalizes the ‘‘do hydrology backward’’ approach introduced by *Kirchner* [2008] to second- and higher-order nonlinear dynamical catchment systems, and simultaneously provides uncertainty estimates of rainfall, model parameters and streamflow predictions. This approach is key to understanding how much information can be extracted from the observed discharge data, and quantifying the uncertainty associated with the inferred records of whole-catchment precipitation.

[7] The paper is organized as follows. Section 2 briefly discusses the general model calibration problem, and highlights the need for explicit treatment of forcing data error. Section 3 describes a parsimonious framework for describing forcing data error that is very similar to the methodology described by *Kavetski et al.* [2002]. Successful implementation of this method requires the availability of an efficient and robust parameter estimation method. Section 4 introduces the differential evolution adaptive Metropolis (DREAM) algorithm, which satisfies this requirement. Then section 5 demonstrates how DREAM can help to provide fundamental insights into rainfall uncertainty, and its effect on streamflow prediction uncertainty and the optimized values of the hydrologic model parameters. A summary with conclusions is presented in section 6.

2. General Model Calibration Problem

[8] For a model to be useful in prediction, the values of the parameters need to accurately reflect the invariant properties of the components of the underlying system they represent. Unfortunately, in watershed hydrology many of the parameters can generally not be measured directly, but can only be meaningfully derived through calibration against a historical record of streamflow data. Figure 1 provides a schematic overview of the resulting model calibration problem. In this plot, the symbol t denotes time, and the circled plus represents observations of the forcing (rainfall) and streamflow response that are subject to measurement errors and uncertainty, and therefore may be different than the true values. Similarly, the boxed f represents the watershed model with functional response to indicate that the model is at best only an approximation of the underlying catchment. The label ‘‘output’’ on the y axis of the plot on the right hand side can represent any time series of data; here this is considered to be the streamflow response.

[9] Using a priori values of the parameters derived through either regionalization relationships, pedotransfer functions or some independent in situ or remote sensing data, the predictions of the model (indicated with grey line) are behaviorally consistent with the observations (dotted line), but demonstrate a significant bias toward lower streamflow values. The common approach is to ascribe this mismatch between model and data to parameter uncertainty, without considering forcing and structural model uncertainty as potential sources of error. The goal of model calibration then becomes one of finding those values of the parameters that provide the best possible fit to the observed behavior using either manual or computerized methods. A model calibrated by such means can be used for the simulation or prediction of hydrologic events outside of the historical record used for model calibration, provided that it can be reasonably assumed that the physical characteristics of the watershed and the hydrologic/climate conditions remain similar.

[10] Mathematically, the model calibration problem depicted in Figure 1 can be formulated as follows. Let $\tilde{\mathbf{S}} = f(\theta, \mathbf{P})$ denote the streamflow predictions $\tilde{\mathbf{S}} = \{\tilde{s}_1, \ldots, \tilde{s}_n\}$ of the model f with observed forcing $\mathbf{P}$ (rainfall, and potential evapotranspiration), and watershed model parameters θ. Let $\mathbf{S} = \{s_1, \ldots, s_n\}$ represent a vector with n observed streamflow values. The difference between the model-predicted

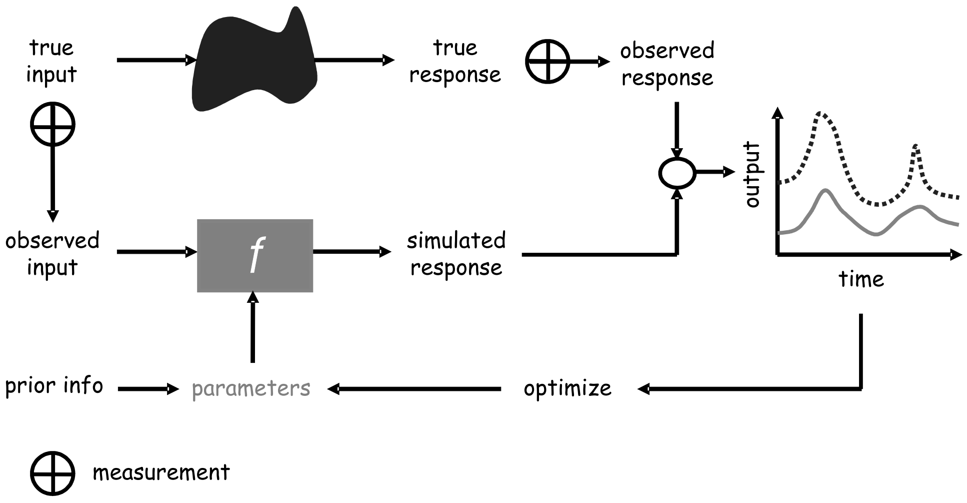

Figure 1. Schematic overview of the model calibration problem. The model parameters are iteratively adjusted so that the predictions of the model, *f* (represented with the solid line), approximate as closely and consistently as possible the observed response (indicated with the dotted line).

streamflow and measured discharge can be represented by the residual vector or objective function E:

$$E(\theta) = \left\{\tilde{\mathbf{S}} - \mathbf{S}\right\} = \{\tilde{s}_1 - s_1, \ldots, \tilde{s}_n - s_n\} = \{e_1(\theta), \ldots, e_n(\theta)\} \quad (1)$$

Traditionally, we are seeking to have a minimal discrepancy between our model predictions and observations. This can be done by minimizing the following additive simple least squares (SLS) objective function with respect to θ:

$$F_{SLS}(\theta) = \sum_{i=1}^{n} e_i(\theta)^2 \quad (2)$$

Significant advances have been made in the last few decades by posing the hydrologic model calibration problem within this SLS framework.

[11] Recent contributions to the literature have questioned the validity of this classical model calibration paradigm when confronted with significant errors and uncertainty in model forcing, **P** and model structure, *f*. These error sources need to be explicitly considered to be able to advance the field of watershed hydrology, and to help draw appropriate conclusions about parameter, model predictive and state uncertainty. In principle, one could hypothesize more appropriate statistical error models for forcing data and model structural inadequacies, and estimate the unknowns in these models simultaneously with the hydrologic model parameters during model calibration. However, this approach will significantly increase the number of parameters to be estimated. To successfully resolve this problem, we use recent advances in Markov chain Monte Carlo (MCMC) simulation for sampling of high-dimensional posterior distributions. Specifically, we use a new algorithm called DREAM and exploit the advantages that this algorithm possesses when implemented on a distributed computer network.

[12] This paper focuses on rainfall forcing error only, because these errors typically dominate in many catchments because of the significant spatial and temporal variability of rainfall fields. However, the inference methodology presented herein can easily be extended to include additional errors such as potential evapotranspiration or temperature. These quantities will primarily affect the streamflow response during drying conditions of the watershed.

3. Description of Rainfall Forcing Data Error

[13] There are various ways in which rainfall forcing error can be included in the parameter estimation problem in watershed model calibration. In principle, one could make every rainfall observation an independent, latent variable, and augment the vector of watershed model parameters with these additional variables. Unfortunately, this approach is infeasible, as the dimensionality of the parameter estimation problem would grow manifold, and the statistical significance of the inferred parameters would be subject to question. For instance, if daily rainfall observations are used for simulation purposes, about 1,100 additional latent variables would be necessary if using 3 years of streamflow data for calibration purposes. With so many latent variables, the predictive value of the hydrologic model would become very low. Moreover, this approach is also susceptible to overparameterization, deteriorating the forecasting capabilities of the watershed model.

[14] An alternative implementation used in this paper is to use a single rainfall multiplier for each storm event. This is an attractive and parsimonious alternative that has been successfully applied by *Kavetski et al.* [2002, 2006b]. By allowing these multipliers to vary between hydrologically reasonable ranges, systematic errors in rainfall forcing can be corrected, and parameter inference and streamflow predictions can be improved. This method is computationally feasible and has the advantage of being somewhat scale-independent. The only limitation is that observed rainfall depths of zero are not corrected.

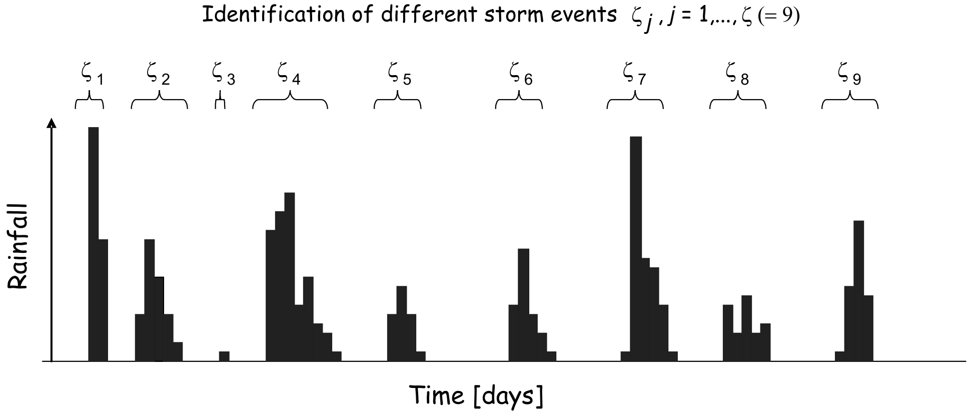

Figure 2. Illustrative example of how rainfall multipliers are assigned to individual storm events. The values of these multipliers are estimated simultaneously with the hydrologic model parameters by minimizing the mismatch between observed and simulated catchment response.

[15] Prior to calibration, individual storm events are identified from the measured hyetograph and hydrograph. A simple example of this approach is illustrated in Figure 2. Each storm, $j = 1,\ldots,\zeta$ is assigned a different rainfall multiplier β_j, and these values are added to the vector of model parameters θ to be optimized; hence $\theta = [\theta; \beta]$. Note that the individual storms are clearly separated in time in the hypothetical example considered in Figure 2. This makes the assignment of the multipliers straightforward. In practice, the distinction between different storms is typically not that simple, and therefore information from the measured hyetograph and discharge data must be combined to identify different rainfall events.

[16] It is desirable to develop an inference method that not only estimates the most likely value of θ, but simultaneously also estimates its underlying posterior probability distribution. This approach should provide useful information about the uncertainty associated with the model parameters and storm multipliers, and help generate predictive uncertainty distributions. The next section discusses the Bayesian approach used in this study to estimate θ using observations of catchment streamflow response and rainfall data.

4. Bayesian Statistics and Markov Chain Monte Carlo Simulation

[17] In the last decade, Bayesian statistics have increasingly found use in the field of hydrology for statistical inference of parameters, state variables, and model output prediction [*Kuczera and Parent*, 1998; *Bates and Campbell*, 2001; *Engeland and Gottschalk*, 2002; *Vrugt et al.*, 2003; *Marshall et al.*, 2004; *Liu and Gupta*, 2007]. The Bayesian paradigm provides a simple way to combine multiple probability distributions using Bayes theorem. In a hydrologic context, this method is suited to systematically address and quantify the various error sources within a single cohesive, integrated, and hierarchical method.

[18] To successfully implement the Bayesian paradigm, sampling methods are needed that can efficiently summarize the posterior probability density function (pdf). This distribution combines the data li od with a prior distribution using Bayes theorem, and contains all the desired information to make statistically sound inferences about the uncertainty of the individual components in the model. Unfortunately, for most practical hydrologic problems this posterior distribution cannot be obtained by analytical means or by analytical approximation. We therefore resort to iterative approximation methods such as Markov chain Monte Carlo (MCMC) sampling to generate a sample from the posterior pdf.

4.1. Random Walk Metropolis Algorithm

[19] The basis of the MCMC method is a Markov chain that generates a random walk through the search space with stable frequency stemming from a fixed probability distribution. To visit configurations with a stable frequency, an MCMC algorithm generates trial moves from the current ("*old*") position of the Markov chain θ_{t-1} to a new state ϑ. The earliest and most general MCMC approach is the random walk Metropolis (RWM) algorithm. Assuming that a random walk has already sampled points $\{\theta_0, \ldots, \theta_{t-1}\}$, this algorithm proceeds in the following three steps. First, a candidate point ϑ is sampled from a proposal distribution q that is symmetric, $q(\theta_{t-1}, \vartheta) = q(\vartheta, \theta_{t-1})$ and may depend on the present location, θ_{t-1}. Next, the candidate point is either accepted or rejected using the Metropolis acceptance probability:

$$\alpha(\theta_{t-1}, \vartheta) = \begin{cases} \min\left(\frac{\pi(\vartheta)}{\pi(\theta_{t-1})}, 1\right) & \text{if } \pi(\theta_{t-1}) > 0 \\ 1 & \text{if } \pi(\theta_{t-1}) = 0 \end{cases} \quad (3)$$

where $\pi(\cdot)$ denotes the density of the target distribution. Finally, if the proposal is accepted, the chain moves to ϑ otherwise the chain remains at its current location θ_{t-1}.

[20] The original RWM scheme was constructed to maintain detailed balance with respect to $\pi(\cdot)$ at each step in the chain:

$$p(\theta_{t-1})p(\theta_{t-1} \rightarrow \vartheta) = p(\vartheta)p(\vartheta \rightarrow \theta_{t-1}) \quad (4)$$

where $p(\theta_{t-1})$ ($p(\vartheta)$) denotes the probability of finding the system in state $\theta_{t-1}(\vartheta)$, and $p(\theta_{t-1} \rightarrow \vartheta)$ ($p(\vartheta \rightarrow \theta_{t-1})$)

denotes the conditional probability of performing a trial move from θ_{t-1} to ϑ (ϑ to θ_{t-1}). The result is a Markov chain which, under certain regularity conditions, has a unique stationary distribution with pdf $\pi(\cdot)$. In practice, this means that if one looks at the values of θ generated by the RWM that are sufficiently far from the starting value, the successively generated parameter combinations will be distributed with stable frequencies stemming from the underlying posterior pdf of θ, $\pi(\cdot)$. Hastings extended equation (4) to include nonsymmetrical proposal distributions, i.e., $q(\theta_{t-1}, \vartheta) \neq q(\vartheta, \theta_{t-1})$, in which a proposal jump to ϑ and the reverse jump do not have equal probability. This extension is called the Metropolis Hastings algorithm (MH), and has become the basic building block of many existing MCMC sampling schemes.

[21] Existing theory and experiments prove convergence of well-constructed MCMC schemes to the appropriate limiting distribution under a variety of different conditions. In practice, this convergence is often observed to be impractically slow. This deficiency is frequently caused by an inappropriate selection of the proposal distribution used to generate trial moves in the Markov chain. To improve the search efficiency of MCMC methods, it seems natural to tune the orientation and scale of the proposal distribution during the evolution of the sampler to the posterior target distribution, using the information from past states. This information is stored in the sample paths of the Markov chain.

[22] An adaptive MCMC algorithm that has become popular in the field of hydrology is the shuffled complex evolution Metropolis (SCEM-UA) global optimization algorithm developed by *Vrugt et al.* [2003]. This method is a modified version of the original SCE-UA global optimization algorithm [*Duan et al.*, 1992] and runs multiple chains in parallel to provide a robust exploration of the search space. These chains communicate with each other through an external population of points, which are used to continuously update the size and shape of the proposal distribution in each chain. The MCMC evolution is repeated until the $\hat{R}$ statistic of *Gelman and Rubin* [1992] indicates convergence to a stationary posterior distribution. This statistic compares the between and within variance of the different parallel chains.

[23] Numerous studies have demonstrated the usefulness of the SCEM-UA algorithm for estimating (nonlinear) parameter uncertainty. However, the method does not maintain detailed balance at every single step in the chain, casting doubt on whether the algorithm will appropriately sample the underlying pdf. Although various benchmark studies have reported very good sampling efficiencies and convergence properties of the SCEM-UA algorithm, violating detailed balance is a reason for at least some researchers and practitioners not to use this method for posterior inference. An adaptive MCMC algorithm that is efficient in hydrologic applications, and maintains detailed balance and ergodicity therefore remains desirable.

4.2. Differential Evolution Adaptive Metropolis (DREAM)

[24] *Vrugt et al.* [2008a] recently introduced the differential evolution adaptive Metropolis (DREAM) algorithm. This algorithm uses diff al evolution as genetic algorithm for population evolution, with a Metropolis selection rule to decide whether candidate points should replace their respective parents or not. DREAM is a follow up on the DE-MC method of *ter Braak* [2006], but contains several extensions to increase search efficiency and acceptance rate for complex and multimodal response surfaces with numerous local optimal solutions. Such surfaces are frequently encountered in hydrologic modeling. The method is presented below.

[25] 1. Draw an initial population Θ of size N, typically $N = d$ or $2d$, using the specified prior distribution.

[26] 2. Compute the density $\pi(\theta^i)$ of each point of Θ, $i = 1,\ldots,N$.

FOR $i \leftarrow 1, \ldots, N$ DO (CHAIN EVOLUTION)

[27] 3. Generate a candidate point, ϑ^i in chain i,

$$\vartheta^i = \theta^i + \gamma(\delta) \cdot \sum_{j=1}^{\delta} \theta^{r_1(j)} - \gamma(\delta) \cdot \sum_{n=1}^{\delta} \theta^{r_2(n)} + \mathbf{e} \tag{5}$$

where δ signifies the number of pairs used to generate the proposal, and $r_1(j), r_2(n) \in \{1,\ldots,N\}$; $r_1(j) \neq r_2(n) \neq i$ for $j = 1,\ldots,\delta$, and $n = 1,\ldots\delta$. The value of $\mathbf{e} \sim N_d(0, b)$ is drawn from a symmetric distribution with small b, and the value of γ depends on the number of pairs used to create the proposal. By comparison with RWM, a good choice for $\gamma = 2.38/\sqrt{2\delta d_{eff}}$ [*Roberts and Rosenthal*, 2001; *Ter Braak*, 2006], with $d_{eff} = d$, but potentially decreased in the next step. This choice is expected to yield an acceptance probability of 0.44 for $d = 1$, 0.28 for $d = 5$ and 0.23 for large d.

[28] 4. Replace each element, $j = 1, \ldots, d$ of the proposal ϑ^i_j with θ^i_j using a binomial scheme with crossover probability *CR*,

$$\vartheta^i_j = \begin{cases} \theta^i_j & \text{if } U \leq 1 - CR, \quad d_{eff} = d_{eff} - 1 \\ \vartheta^i_j & \text{otherwise} \end{cases} \quad j = 1, \ldots, d \tag{6}$$

where $U \in [0, 1]$ is a draw from a uniform distribution.

[29] 5. Compute $\pi(\vartheta^i)$ and accept the candidate point with Metropolis acceptance probability, $\alpha(\theta^i, \vartheta^i)$,

$$\alpha\left(\theta^i, \vartheta^i\right) = \begin{cases} \min\left(\frac{\pi\left(\vartheta^i\right)}{\pi\left(\theta^i\right)}, 1\right) & \text{if } \pi\left(\theta^i\right) > 0 \\ 1 & \text{if } \pi\left(\theta^i\right) = 0 \end{cases} \tag{7}$$

[30] 6. If the candidate point is accepted, move the chain, $\theta^i = \vartheta^i$; otherwise remain at the old location, θ^i.

END FOR (CHAIN EVOLUTION)

[31] 7. Remove potential outlier chains using the inter-quartile range (IQR) statistic.

[32] 8. Compute the *Gelman and Rubin* [1992], $\hat{R}$ convergence diagnostic for each dimension $j = 1,\ldots,d$ using the last 50% of the samples in each chain.

[33] 9. If $\hat{R} \leq 1.2$ for all j, stop, otherwise go to CHAIN EVOLUTION.

[34] The DREAM algorithm adaptively updates the scale and orientation of the proposal distribution during the evolution of the individual chains to a limiting distribution. The method starts with an initial population of points to strategically sample the space of potential solutions. The use of a number of individual chains with different starting points enables dealing with multiple regions of highest attraction, and facilitates the use of a powerful array of heuristic tests to judge whether convergence of DREAM has been achieved. If the state of a single chain is given by a single d-dimensional vector θ, then at each generation t, the N chains in DREAM define a population Θ, which corresponds to an $N \times d$ matrix, with each chain as a row. Jumps in each chain are generated by taking a fixed multiple of the difference of randomly other chosen chains. The Metropolis ratio is used to decide whether to accept candidate points or not. At every step, the points in Θ contain the most relevant information about the search, and this population of points is used to globally share information about the progress of the search of the individual chains. This information exchange enhances the survivability of individual chains, and facilitates adaptive updating of the scale and orientation of the proposal distribution. This series of operations results in a MCMC sampler that conducts a robust and efficient search of the parameter space. Because the joint pdf of the N chains factorizes to $\pi(\theta_1) \times \ldots \times \pi(\theta_N)$, the states $\theta_1 \ldots \theta_N$ of the individual chains are independent at any generation after DREAM has become independent of its initial value. After this so-called burn-in period, the convergence of a DREAM run can thus be monitored with the $\hat{R}$ statistic of *Gelman and Rubin* [1992].

[35] Outlier chains can significantly deteriorate the performance of MCMC samplers, and need to be removed to facilitate convergence to a limiting distribution. To detect aberrant trajectories, DREAM stores in Ω the mean of the logarithm of the posterior densities of the last 50% of the samples in each chain. From these, the interquartile range statistic, $IQR = Q_3 - Q_1$ is computed, in which Q_1 and Q_3 denote the lower and upper quartile of the N different chains. Chains with $\Omega < Q_1 - 2\ IQR$ are considered outliers, and are moved to the current best member of Θ. This step does not maintain detailed balance and can therefore only be used during burn in. If an outlier chain is being detected we apply another burn-in period before summarizing the posterior moments.

[36] To speed up convergence to the target distribution, DREAM estimates a distribution of CR values during burn in that maximizes the squared distance, $\triangle = \sum_{i=1}^{N} \sum_{j=1}^{d} (\bar{\theta}_{j,t}^{i} - \bar{\theta}_{j,t-1}^{i})^2$ between two subsequent samples, $\bar{\theta}_t$ and $\bar{\theta}_{t-1}$ of the N chains. The position of the chains is normalized (hence the bar) with the prior distribution so that all d dimensions contribute equally to $\triangle$. A detailed description of this adaptation strategy appears in *Vrugt et al.* [2008a] and so will not be repeated here. Note that self-adaptation within the context of multiple different search algorithms is presented by *Vrugt and Robinson* [2007a], and has shown to significantly enhance the efficiency of population-based evolutionary optimization.

[37] The DREAM scheme is different from the DE-MC method in three important ways. First, DREAM implements a randomized subspace sam ling strategy, and only modifies selected dimensions with crossover probability CR each time a candidate point is generated. This significantly enhances efficiency for higher-dimensional problems, because with increasing dimensions it is often not optimal to change all d elements of θ_i simultaneously. During the burn-in phase, DREAM adaptively chooses the CR values that yield the best mixing properties of the chains. Second, DREAM incorporates a Differential Evolution offspring strategy that also includes higher-order pairs. This increases the diversity of the proposals and thus variability in the population. Third, DREAM explicitly handles and removes chains that are stuck in nonproductive parts of the parameter space. Such outlier chains prohibit convergence to a limiting distribution, and thus significantly deteriorate the performance of MCMC samplers.

4.3. Theorem

[38] The theorem is that DREAM yields a Markov chain that is ergodic with unique stationary distribution with pdf $\pi(\cdot)^N$.

[39] The proof consists of three parts and was presented by *Vrugt et al.* [2008a].

[40] 1. Chains are updated sequentially and conditionally on the other chains. Therefore DREAM is an N-component Metropolis-within-Gibbs algorithm that defines a single Markov chain on the state space [*Robert and Casella*, 2004]. The conditional pdf of each component is $\pi(\cdot)$.

[41] 2. The update of the ith chain uses a mixture of kernels. For $\delta = 1$, there are $\begin{pmatrix} N-1 \\ 2 \end{pmatrix}$ such kernels. This mixture kernel maintains detailed balance with respect to $\pi(\cdot)$, if each of its components does [*Robert and Casella*, 2004], as we show now. For the ith chain, the conditional probability to jump from θ_{t-1}^{i} to ϑ^i, $p(\theta_{t-1}^{i} \rightarrow \vartheta^i)$ is equal to the reverse jump, $p(\vartheta^i \rightarrow \theta_{t-1}^{i})$ as the distribution of $\mathbf{e}$ is symmetric and the pair $(\theta_{t-1}^{r1}, \theta_{t-1}^{r2})$ is as likely as $(\theta_{t-1}^{r2}, \theta_{t-1}^{r1})$. This also holds true for $\delta > 1$, when more than two members of Θ_{t-1} are selected to generate a proposal point. Detailed balance is thus achieved point wise by accepting the proposal with probability $\min(\pi(\vartheta^i)/\pi(\theta_{t-1}^{i}), 1)$. Detailed balance also holds in terms of arbitrary measurable sets, as the Jacobian of the transformation of equation (5) is 1 in absolute value.

[42] 3. As each update maintains conditional detailed balance, the joint stationary distribution associated with DREAM is $\pi(\theta^1, \ldots, \theta^N) = \pi(\theta^1) \times \ldots \times \pi(\theta^N)$ [*Mengersen and Robert*, 2003]. This distribution is unique and must be the limiting distribution, because the chains are aperiodic, positive recurrent (not transient) and irreducible [*Robert and Casella*, 2004]. The first two conditions are satisfied, except for trivial exceptions. The unbounded support of the distribution of $\mathbf{e}$ in equation (5) guarantees the third condition. This concludes the ergodicity proof.

[43] Case studies presented by *Vrugt et al.* [2008a] have demonstrated that DREAM is generally superior to existing MCMC schemes, and can efficiently handle multimodality, high dimensionality and nonlinearity. In that same paper, recommendations have also been given for some of the values of the algorithmic parameters. The only parameter that remains to be specified by the user before the sampler

Table 1. Prior Ranges and Description of the HYMOD Parameters and Rainfall Multipliers

Parameter	Description	Minimum	Maximum	Unit
$C_{\max}$	maximum storage in watershed	1.00	500.00	mm
b_{exp}	spatial variability of soil moisture storage	0.10	2.00	
Alpha	distribution factor between two reservoirs	0.10	0.99	
R_s	residence time slow flow reservoir	0.001	0.10	days
β_j, $j = 1, \ldots, \zeta$	rainfall multipliers	0.25	2.50	
R_q	residence time quick flow reservoir	0.1	0.99	days

can be used for statistical inference is the population size N. We generally recommend using $N \geq d$, although the subspace sampling strategy allows taking $N \ll d$. In each of the case studies presented in this paper, we report the values of N used in DREAM.

5. Case Studies

[44] To illustrate the insights that the approach developed in this study can offer with respect to forcing error, we apply our methodology to streamflow forecasting using the parsimonious, five-parameter Hydrologic Model (HYMOD). This model, originally developed by *Boyle* [2000], consists of a relatively simple rainfall excess model, described in detail by *Moore* [1985], connected with two series of linear reservoirs (three identical quick reservoirs, and a single reservoir for the slow response). The model has five parameters: the maximum storage capacity in the catchment $C_{\max}$ (L), the degree of spatial variability of soil moisture capacity within the catchment b_{exp}, the factor distributing the flow between the two series of reservoirs *Alpha*, and the residence times of the linear slow and quick flow reservoirs, R_s (days) and R_q (days), respectively.

[45] In our studies, we use historical data from the Leaf River (1950 km^2) and French Broad (767 km^2) watersheds in the USA. The data consists of mean areal precipitation (mm/d), potential evapotranspiration (mm/d), and streamflow (m^3/s). To illustrate the approach, a period of 5 years of streamflow data was used for model calibration, whereas the remainder of the data was used for evaluation purposes. Various contributions to the hydrologic literature have recommended to use longer time series of streamflow for calibration purposes. Because of computational reasons, however a period of 5 years was deemed appropriate to illustrate our methodology. In this 5 year calibration time series, a total of ζ = 57 and ζ = 59 storm events were identified for the Leaf River (1 October 1953 to 30 September 1958) and French Broad (1 October 1954 to 30 September 1959) watersheds, respectively.

[46] The upper and lower bounds that define the prior uncertainty ranges of the HYMOD model parameters and rainfall multipliers are given in Table 1. A uniform prior distribution is assumed over this multidimensional hypercube, which implies that the storm events are independent, and that the information content of the observed rainfall is limited to pattern only, without useful information about storm depths. *Kavetski et al.* [2006a] do not recommend

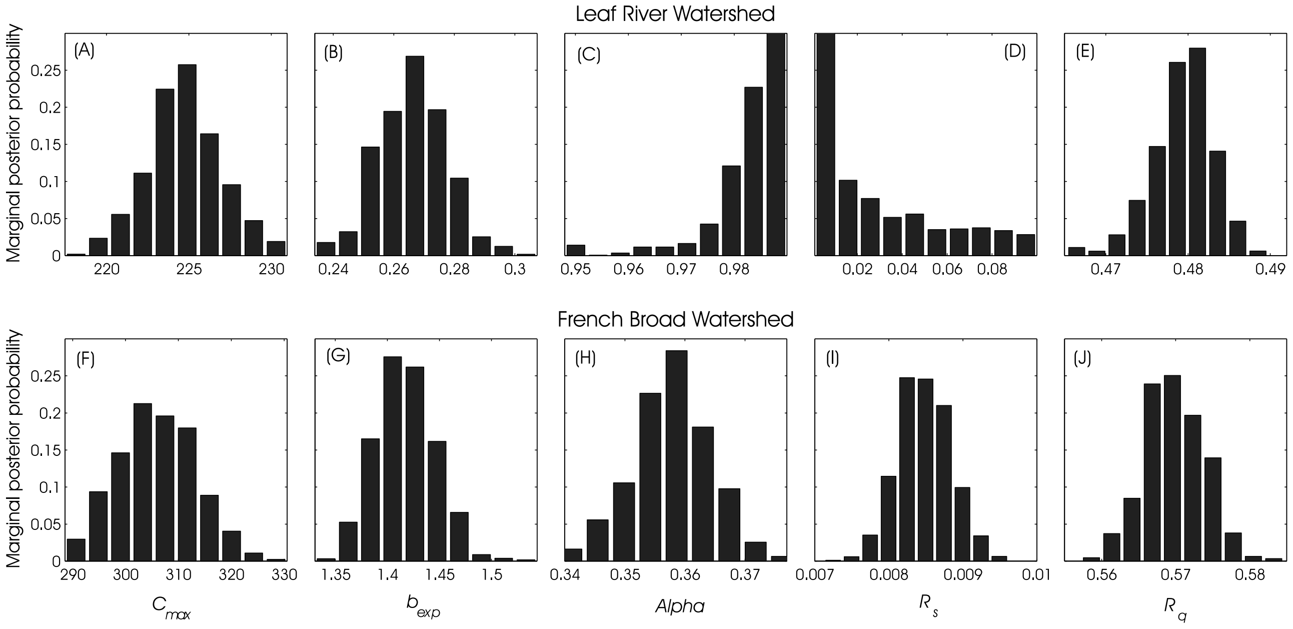

Figure 3. Classical (without explicit assessment of forcing data error) hydrologic model calibration: marginal posterior probability distributions of the HYMOD model parameters $C_{\max}$, b_{exp}, *Alpha*, R_s, and R_q for the (top) Leaf River and (bottom) French Broad watersheds in the United States. The histograms were constructe g the last 10,000 samples generated with DREAM after convergence to a limiting distribution.

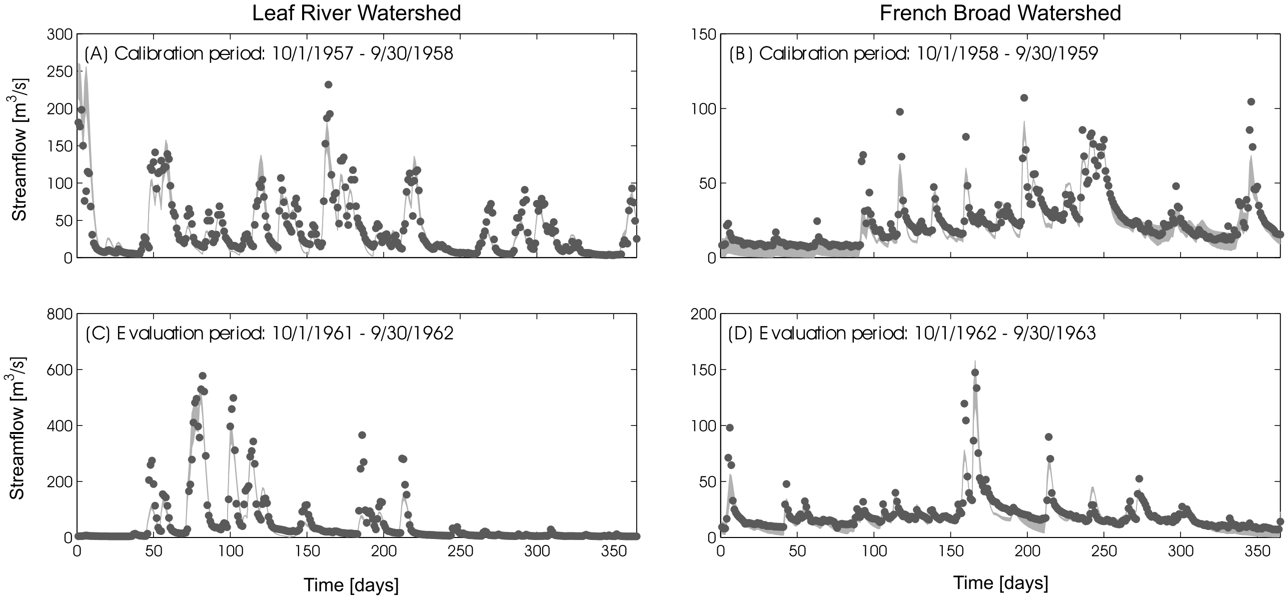

Figure 4. Classical hydrologic model calibration: 95% streamflow prediction uncertainty ranges for the (left) Leaf River and (right) French Broad watersheds. A distinction is made between the (top) calibration and (bottom) evaluation periods. The uncertainty bounds represent HYMOD parameter uncertainty only. Observed streamflows are indicated with solid circles.

using uniform priors for the rainfall multipliers, as this might result in ill posedness of the resulting parameter estimation problem. Yet, the results with DREAM presented below do not seem to support that conjecture. To reduce sensitivity to state value initialization, we used a 365-day warm up period prior to the calibration data time series, during which no updating of the posterior density was performed.

5.1. Case Study 1: Estimation of HYMOD Parameters

[47] This first case study focuses on estimation of the HYMOD parameters without explicit assessment of forcing data error. The results of this analysis serve as benchmark for the next studies that explicitly incorporate rainfall data error in the model calibration process. In this first study, we use the following classical density function:

$$\pi(\theta|\mathbf{S}) \propto c \cdot \pi(\theta) \cdot F_{SLS}(\theta|\mathbf{S})^{-\frac{1}{2}n} \qquad (8)$$

where c is a normalizing contact, and $\pi(\theta)$ signifies the prior distribution of θ. This distribution combines the data likelihood with a prior distribution using Bayes theorem. *Vrugt et al.* [2008b] extend the formulation of this density function to explicitly include structural error through the use of a first-order autoregressive scheme of the error residuals. The resulting inference problem is solved with DREAM.

[48] Figure 3 presents the posterior marginal probability density distributions for each of the HYMOD model parameters for the Leaf River and French Broad watersheds using the samples generated with the DREAM algorithm. For both data sets, we used a population size of $N = 2d$ with a maximum total of 25,000 model evaluations. The first 60% of the samples in each of the 10 chains were discarded and used as burn in. No outlier chains were reported with DREAM during burn in.

[49] The marginal posterior pdfs of most of the individual parameters are well defined and occupy only a relatively small region interior to the uniform prior distributions (e.g., Table 1) of the individual dimensions. This shows that the observed streamflow data contains sufficient information to estimate these parameters. This is further confirmed with relatively small (linear) correlation values between the 5 parameters. Note that most histograms appear approximately Gaussian with the exception of the marginal pdfs of *Alpha* and R_s for the Leaf River, which significantly depart from normality and tend to concentrate most of the probability mass at their upper and lower bounds, respectively. The ranges for these parameters cannot be further relaxed without resulting in physically unrealistic behavior of the model. This raises the question of whether these two parameters are actually representing invariant behavior of the underlying catchment, or whether they are compensating for structural deficiencies in the model, or systematic errors in the forcing data. Greater insight into this issue requires a more explicit treatment of these two error sources. The next case study contrasts the results of this classical model calibration approach against those obtained when using an explicit treatment of forcing data error through a comparison of parameter estimates and streamflow prediction uncertainty bounds.

[50] To understand how the uncertainty in the model parameters translates into HYMOD predictive uncertainty, consider Figure 4, which presents the 95% streamflow uncertainty bounds for a selected period of the calibration (top plots) and evaluation (bottom plots) period for the Leaf River (left) and French Broad (right) watersheds. The observed discharge data are separately indicated with solid circles. The model seems to be unable to match large

Table 2. Synthetic Data[a]

		Euclidean Distance	
Parameter	Prior Range	Leaf River	French Broad
$C_{\max}$	1.00–500.00	0.32	0.037
b_{exp}	0.10–2.00	0.008	0.0002
$Alpha$	0.10–0.99	0.0026	0.0002
R_s	0.001–0.10	0.0001	0.0001
R_q	0.10–0.99	0.0001	0.0000
β_j, j = 1,. . .,ζ	0.25–2.50	0.007	0.009

[a] Shown are average normalized Euclidean distance between true values of HYMOD model parameters and storm multipliers and their estimates derived using the DREAM adaptive sampling scheme. Listed statistics denote averages over 25 different calibration cases.

portions of the hydrograph. This is indicated by large sections where the darkly shaded region does not bracket the observed streamflow data. These findings are consistent with other results presented in the literature, and stimulate the development of an inference framework that takes explicit consideration of the role of forcing and model error.

5.2. Case Study 2: Estimation of HYMOD Parameters and Storm Multipliers Using Streamflow Data

[51] The second case study involves simultaneous estimation of the HYMOD model parameters and rainfall multipliers using observed streamflow data. To verify whether this approach is computationally feasible, synthetically generated streamflow data are used first, followed by real-world observations of discharge.

5.2.1. Synthetic Streamflow Data

[52] To generate the synthetic discharge observations, a total of $\zeta = 57$ and $\zeta = 59$ different rainfall multipliers were first drawn using Latin hypercube sampling within $[0.25, 2.50]^{\zeta}$ (see Table 1). These two vectors of multipliers are then combined with the observed rainfall depths of both watersheds to generate two rainfall hyetographs. These rainfall records are subsequently used with randomly sampled values of the HYMOD parameters (within the bounds specified in Table 1) to create a 5-year time series of synthetic daily discharge data for the Leaf River and French Broad watersheds. Then, DREAM is executed with equation (8) to back out the posterior pdf of the HYMOD model parameters and storm multipliers. This is done for both catchments using a maximum total of 1,500,000 model evaluations. We used a uniform initial sampling distribution of the model parameters and rainfall multipliers with ranges specified in Table 1 to test the robustness of DREAM when confronted with relative poor prior information on the location of the posterior pdf in the parameter space. In lieu of sampling variability and model nonlinearity, we repeated this experiment 25 different times using different values of the multipliers and model parameters. The results of this analysis are reported in Table 2.

[53] Table 2 summarizes the average Euclidean distance between the true HYMOD parameter values and rainfall multipliers used to generate the synthetic streamflow data, and the maximum likelihood estimates of these parameters derived with DREAM. The listed statistics represent averages over the 25 different calibration time series, and were obtained using a population size of $N = 2d$. Two main conclusions can be drawn from this analysis. First, the estimates of the multipliers and model parameters derived with DREAM are very close to their values used to generate the synthetic streamflow data. This highlights the robustness of DREAM, being able to consistently solve $d = 62$ (Leaf River) and $d = 64$ (French Broad) dimensional parameter estimation problems. Second, streamflow data contain sufficient information to warrant the simultaneous identification of the HYMOD model parameters and rainfall

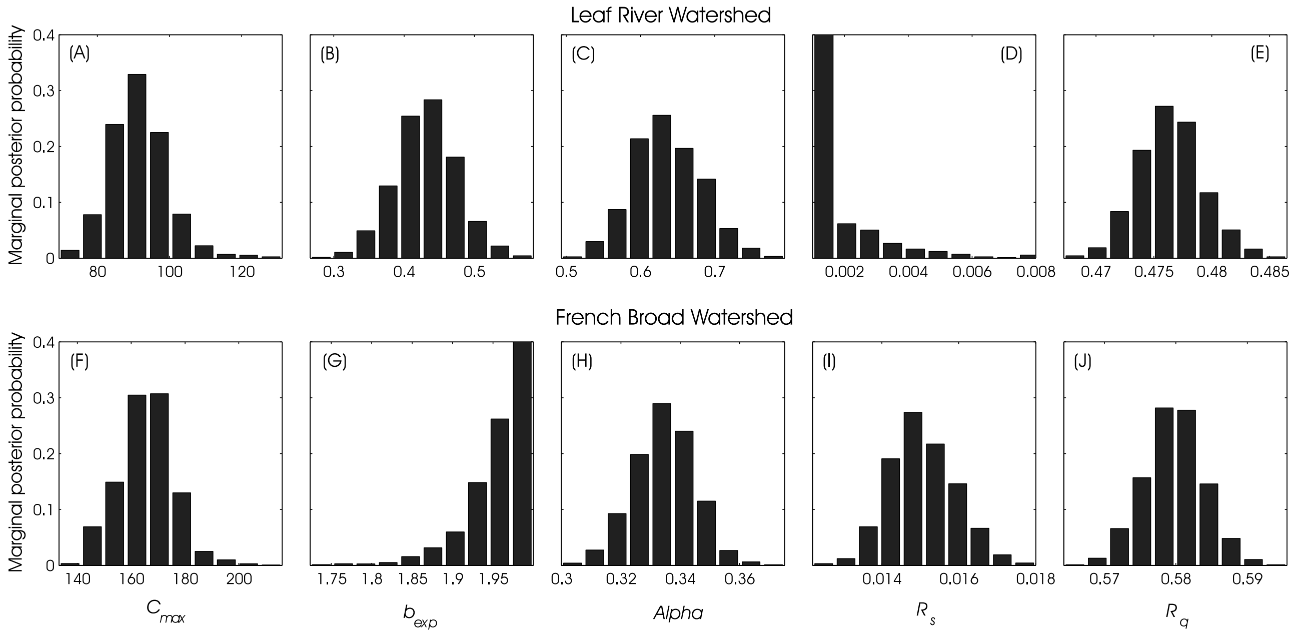

Figure 5. Simultaneous estimation of HYMOD model parameters and rainfall multipliers: marginal posterior probability distributions of the HYMOD model parameters $C_{\max}$, b_{exp}, $Alpha$, R_s, and R_q for the (top) Leaf Rive (bottom) French Broad watersheds. The histograms were constructed using the last 150,000 sample erated with DREAM after convergence to the posterior distribution.

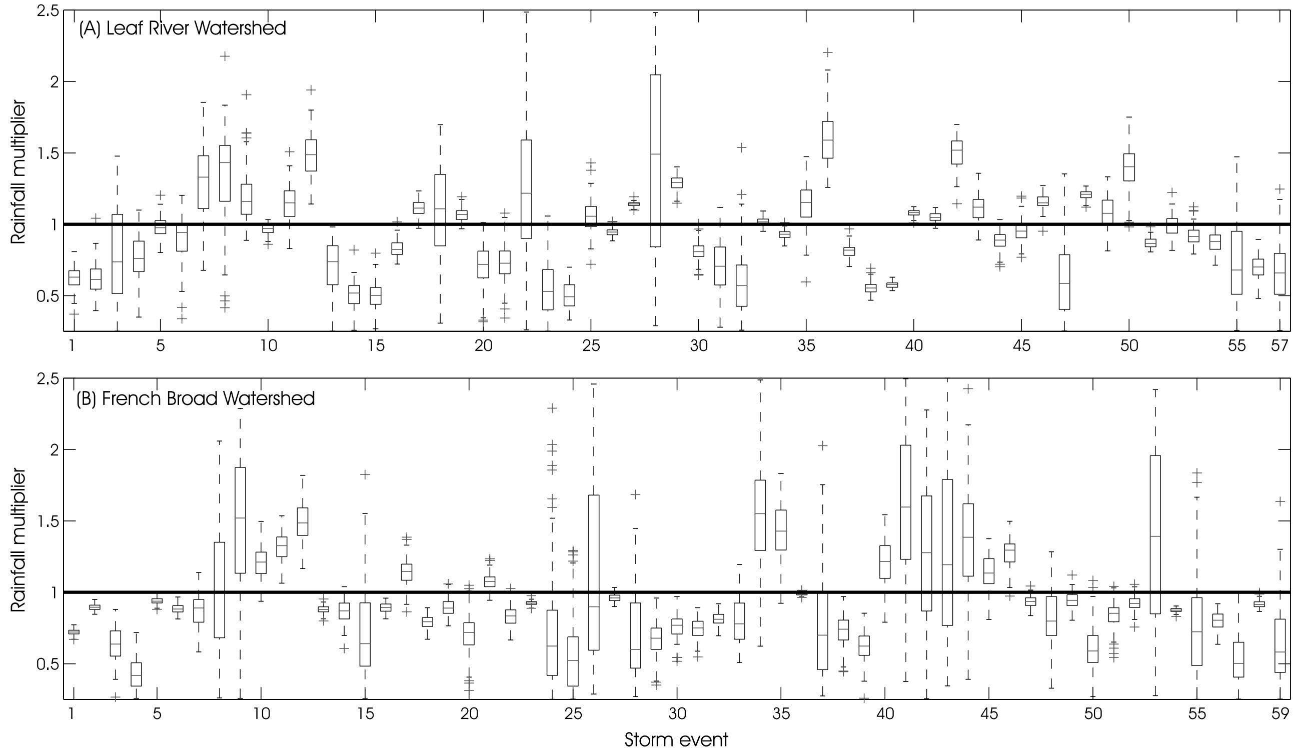

Figure 6. Simultaneous estimation of HYMOD model parameters and rainfall multipliers: box plots of the marginal posterior distributions of the rainfall multipliers for the (top) Leaf River and (bottom) French Broad watersheds.

multipliers. Hence, DREAM has converged to the appropriate values of the parameters. These findings inspire confidence that this inference methodology can be successfully applied to real-world streamflow data.

5.2.2. Observed Streamflow Data

[54] Using measured discharge data, a total of 28 (Leaf River) and 53 (French Broad) different outlier chains were detected with DREAM during burn in using the *IQR* statistic. Figure 5 presents histograms of the HYMOD model parameters using observed streamflow data of the Leaf River and French Broad catchments in the US. These marginal distributions were created using the last 150,000 samples generated with DREAM after convergence to a limiting distribution.

[55] The histograms of the HYMOD model parameters are quite different than those obtained previously in case study 1. Simultaneous estimation of watershed model parameters and rainfall multipliers not only increases the uncertainty for most of the HYMOD parameters, but also results in significantly different values for the mode of the distribution. The only exception is the residence time of the linear quick flow reservoir, R_q, which maintains a similar distribution. It is interesting to observe that the distribution of *Alpha* for the Leaf River data set (Figure 5c) has now become approximately normal, with a value of the mode that appears physically more reasonable. In contrast, for the French Broad River system, the spatial variability of soil moisture storage, b_{exp} changed from a normal distribution in Figure 3g to a truncated distribution with highest probability mass at the upper bound. To closely match the observed streamflow with overall re rainfall amounts (as will be shown later) HYMOD needs to increase the spatial variability in soil moisture storage.

[56] To provide more insights into the values of the rainfall multipliers, consider Figure 6, which presents box plots of the sampled rainfall multipliers for the Leaf River (top plot) and French Broad (bottom plot) catchments. These box plots were created using the last 150,000 samples generated with DREAM in the $N = 2d$ parallel chains. The marginal pdfs of the multipliers vary widely between individual storm events. Some events are very well defined, while others show considerable uncertainty. For instance, compare the box plots of β_{27} and β_{28} for the Leaf River, and β_{26} and β_{27} for the French Broad watershed. These adjacent storms differ substantially in their posterior width, but exhibit approximately similar mean values. The overall mean posterior values of the storm multipliers is $\overline{\beta} = 0.95$ for the Leaf River and $\overline{\beta} = 0.94$ for the French Broad watershed. This shows that, on average our inferred rainfall from the streamflow data is in close correspondence with the observed rainfall depths from the rain gauge data. Detailed analysis further demonstrates that the rainfall multipliers exhibit small temporal autocorrelation, and show no obvious time or seasonality pattern. Furthermore, the d-dimensional correlation matrix of the posterior demonstrates that correlation among the multipliers is small. This confirms our earlier finding that observed daily streamflow data contain sufficient information to warrant the identification of an additional $\zeta = 57$ and $\zeta = 59$ storm multipliers, simultaneous with the five HYMOD model parameters.

[57] It is interesting to observe that most of the storm multipliers are clustered in the vicinity of 1 for both catchments. This illustrates that the measured rainfall is not

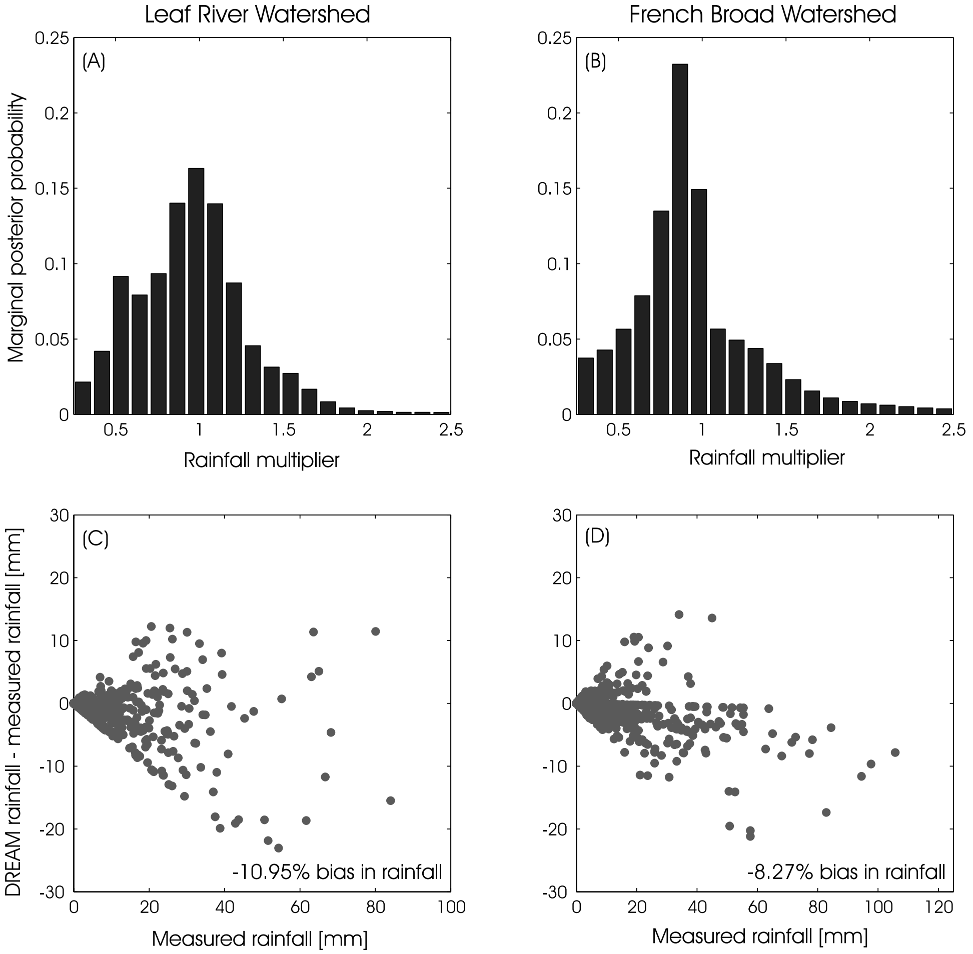

Figure 7. Simultaneous estimation of HYMOD model parameters and rainfall multipliers. (top) Histograms of all storm multipliers combined for the (a) Leaf River and (b) French Broad watersheds. These marginal posterior pdfs were derived by pooling the individual multipliers together using the information depicted in Figure 6. (bottom) Two-dimensional scatterplots of observed precipitation against the deviation between DREAM-optimized and measured rainfall for the (c) Leaf River and (d) French Broad watersheds.

significantly over or underestimating the actual precipitation, but is generally consistent in pattern and depth with the estimated rainfall record derived from the discharge data. This is an important diagnostic and provides support for the claim that the rain gauge data, albeit having a very small spatial support provide, on average, a good proxy of whole-catchment precipitation for both watersheds. This is further demonstrated in Figure 7, which presents a histogram of all precipitation multipliers combined for the Leaf River (Figure 7a) and French Broad (Figure 7b) watersheds. These histograms exhibit an approximate Gaussian distribution, with mean values centered around 1.0 and truncated lower and upper bounds. These bounds force the DREAM estimated multipliers to remain hydrologically realistic.

[58] The marginal posterior pdf of the multipliers presented here can be used to explicitly consider rainfall uncertainty during streamflow prediction. An easy way to do this is to sample a single multiplier for each individual storm event from the histograms presented in Figures 7a and 7b. By combining this vector of multipliers with the observed rainfall record, possible to generate different realizations of rainfall graphs for both watersheds during the evaluation period. This ensemble of rainfall records can be combined with posterior values of the HYMOD model parameters to generate streamflow hydrographs outside the calibration period that include explicit representation of model parameter and forcing data error. The results of this analysis will be presented later.

[59] Figures 7c and 7d quantify the difference between measured and inferred precipitation amounts for the Leaf River and French Broad watersheds, respectively. The proposed inference method suggests that the actual rainfall is, on average, about 10% lower than the measured rainfall for both watersheds. This difference is small, but nevertheless important as this bias explains the observed differences in optimized distributions of the HYMOD model parameters between case studies 1 and 2 (compare Figures 3 and 5). These results establish the need for appropriate characterization and inference of forcing data error during watershed model calibration. Not only to appropriately capture and quantify uncertainty, but also for better testing of hydrologic theory, diagnosis of structural error, and to maximize chances of finding useful regionalization relationships between rainfall-runoff model parameter values and catch-

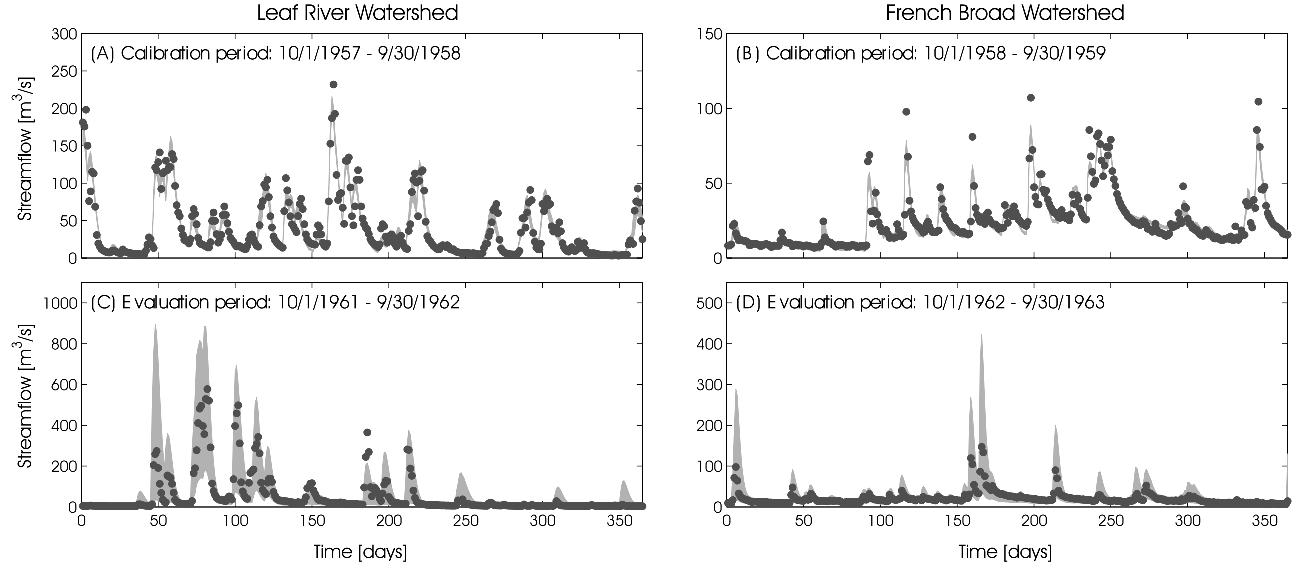

Figure 8. Simultaneous estimation of HYMOD model parameters and rainfall multipliers: 95% streamflow prediction uncertainty ranges for the (left) Leaf River and (right) French Broad watersheds. A distinction is made between the (top) calibration and (bottom) evaluation periods. Observed streamflows are indicated with solid circles.

ment properties. The inference method developed herein is especially designed to minimize the impact of rainfall error on hydrologic parameter estimates, and thus to enable getting the right answers for the right reasons. This latter is important, especially within the context of the PUB initiative.

[60] The validity of the inferred rainfall record can be checked by comparison against the observed spatial variation in rain gauge measurements, and estimates of precipitation from other methods such as rainfall radar. This analysis would help establish how reasonable the inferred rainfall records are, but is beyond the scope of the current paper. Note also, that the results presented here are contingent on HYMOD being a reasonable approximation of the underlying heterogeneous catchment it is trying to represent. This assumption is inappropriate at best, and will at least partially cause the DREAM-estimated rainfall to diverge from the measured precipitation depths. To further reduce ambiguity about the inferred record of whole-catchment rainfall, future analysis should include multiple conceptual watershed models using emerging (Bayesian) model averaging approaches in surface water hydrology [*Ajami et al.*, 2007; *Vrugt and Robinson*, 2007b]. Combining the proxy records of multiple different watershed models provides an explicit way to handle structural uncertainty when doing hydrology backward.

[61] To understand how the uncertainty in the HYMOD model parameters and storm multipliers translates into predictive uncertainty, Figure 8 presents 95% streamflow uncertainty ranges for the Leaf River (left), and French Broad river (right) data sets for a selected portion of the calibration and evaluation period. In each plot, the observed streamflow observations are indicated with dots. The calibration results presented here for both catchments are very similar to those presented previously in Figure 4 for case study 1. Even though forcing error is explicitly considered, the multipliers are conditi or each individual storm to maximize the posterior density and minimize HYMOD prediction uncertainty. However, for the evaluation period, the width of the prediction uncertainty intervals have significantly increased, with streamflow bounds that show a much better coverage of the discharge observations. This is clearly visible in both plots, and particularly evident for two storm events around days 180–220 for the Leaf River watershed. While the classical model calibration with DREAM (Figure 4c) significantly underestimates the actual streamflow observations during these two rainfall events, explicit treatment of rainfall error provides an improved coverage of the data. Further improvements can be made by explicitly considering error in potential evapotranspiration during drying conditions of the watershed, and by including a more formal treatment of model error. *Vrugt et al.* [2008b] performed a similar analysis as done here, but explicitly treat model structural error through the use of a first-order autoregressive scheme of the error residuals.

Table 3. Posterior Mean and Standard Deviation of the HYMOD Model Parameters and Rainfall Multipliers Derived With DREAM for the Two Different Calibration Studies Considered in This Manuscript[a]

	Leaf River Watershed				French Broad Watershed			
	Case Study 1[a]		Case Study 2[b]		Case Study 1[a]		Case Study 2[b]	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
$C_{\max}$	225.02	2.142	91.54	7.92	305.80	7.36	165.91	10.92
b_{exp}	0.262	0.011	0.431	0.044	1.421	0.029	1.963	0.038
Alpha	0.986	0.004	0.636	0.046	0.362	0.006	0.346	0.011
R_s	0.001	0.002	0.002	0.001	0.009	0.000	0.015	0.001
R_q	0.488	0.003	0.476	0.003	0.573	0.004	0.585	0.004
β	N/A	N/A	0.951	0.134	N/A	N/A	0.942	0.247

[a] HYMOD model parameter estimation with observed streamflow data as calibration target.

[b] Simultaneous model parameter and forcing error estimation using observed streamflow data as calibration target. We report average values for the rainfall multipliers.

Table 4. Summary Statistics of the One-Day-Ahead Streamflow Forecasts For the Leaf River and French Broad Watersheds Using Two Different Model Calibration Approaches[a]

	Leaf River Watershed						French Broad Watershed					
	Calibration WY (1954–1958)			Evaluation WY (1959–1963)			Calibration WY (1953–1957)			Evaluation WY (1958–1962)		
	RMSE	CORR	BIAS	RMSE	CORR	BIAS	RMSE	CORR	BIAS	RMSE	CORR	BIAS
Case study 1	20.06	0.88	−0.08	33.46	0.92	−2.32	7.06	0.93	1.86	8.00	0.93	0.93
Case study 2	13.95	0.95	6.17	36.66	0.88	8.26	6.18	0.95	−0.41	7.74	0.94	−0.17

[a]Case study 1, parameter uncertainty only with observed streamflow data as calibration target. Case study 2, parameter and forcing uncertainty using observed streamflow data as calibration target. Distinction is made in performance between the calibration and evaluation period. Units of root-mean-square error (RMSE) and bias (BIAS) are m^3/s and %, respectively. Correlation coefficient (CORR) is dimensionless.

[62] Table 3 compares DREAM estimates of the posterior mean and standard deviation of the HYMOD model parameters and rainfall multipliers for case studies 1 and 2 for the Leaf River and French Broad watersheds. The results presented in Table 3 highlight that (1) explicit consideration of forcing error changes the mode of the posterior pdf of the HYMOD model parameters. This is most evident for the parameters $C_{\max}$, b_{exp} and *Alpha* and has significant implications for regionalization studies; (2) the uncertainty of the HYMOD parameters increases when rainfall estimates are directly inferred from the observed discharge data; and (3) the rainfall multipliers are relatively well defined by calibration against streamflow data, with an average standard deviation of about 0.20 for both watersheds.

[63] Finally, Table 4 presents summary statistics of the one-day-ahead streamflow forecasts of the HYMOD model for the Leaf River and French Broad River watersheds using the two different calibration studies considered in this paper. The listed numbers correspond to the mean ensemble discharge simulation of the posterior pdf derived with DREAM using the 5-year calibration period. As discussed previously, to simulate streamflow during the evaluation period, a precipitation ensemble was generated for each individual storm using different values of the rainfall multipliers randomly drawn from the respective marginal distributions presented in Figures 7a and 7b. These precipitation records were then combined with the posterior pdf of the HYMOD model parameters derived from the calibration period, and used for prediction.

[64] The results presented in Table 4 illustrate that the best performance (RMSE, CORR and BIAS) during the calibration period is obtained in case study 2, when rainfall depths are simultaneously inferred with the HYMOD model parameters. This result is not surprising, because modifications to the observed rainfall data allow the HYMOD model to more closely track the streamflow observations. The improvement in fit is most significant for the Leaf River (RMSE: 20.06 ↦ 13.95), whereas only a 10% reduction in RMSE is observed in case study 2 for the French Broad watershed (RMSE: 7.06 ↦ 6.18). The observed rainfall record for the French Broad catchment is quite consistent in depth with the observed streamflow data, and cannot be improved much through consideration of forcing error. Indeed, the marginal posterior pdf of many of the storm multipliers reside in the vicinity of 1 for the French Broad river system, indicating generally small modifications to the measured precipitation depths with the rain gauges.

[65] A quite similar performance of HYMOD is observed during the evaluation for case studies 1 and 2. Whereas, a 10% deterioration in RMSE is visible when rainfall uncertainty is explicitly considered during streamflow simulation for the Leaf River (RMSE: 33.46 ↦ 36.66), a slight improvement in performance (RMSE: 8.00 ↦ 7.74) is seen for the French Broad watershed. This is a very interesting result, and provides support for the claim that the treatment of rainfall error presented herein, is useful and meaningful and structurally consistent with the observed discharge data outside the calibration period. Furthermore, a much better coverage of the streamflow observations is obtained when the rainfall depths are allowed to vary on the basis of the statistical distribution of the multipliers derived after calibration. We therefore conclude that the presented inference method provides important insights into the issue of forcing data error and inspires new thinking into how to disentangle input, parameter and model structural error. Further support for this is given by *Vrugt et al.* [2008b].

6. Summary and Conclusions

[66] Efficient and robust MCMC algorithms are indispensable for estimating and summarizing the posterior probability density function of input, parameter and model structural error in hydrologic modeling. In this paper, an adaptive MCMC algorithm was developed that can efficiently estimate the posterior pdf of model parameters in the presence of high-dimensional and complex response surfaces with multiple local optima. The method, entitled differential evolution adaptive Metropolis (DREAM), runs multiple chains in parallel and adaptively updates the scale and orientation of the proposal distribution during sampling. Candidate points are generated by using a fixed multiple of the difference of randomly chosen members of the population. The DREAM scheme is an extension to the SCEM-UA global optimization algorithm [*Vrugt et al.*, 2003], but has the advantage of maintaining detailed balance and ergodicity while showing good efficiency on complex, highly nonlinear, and multimodal target distributions [*Vrugt et al.*, 2008a].

[67] The usefulness and applicability of DREAM was demonstrated in the second part of this paper by application to streamflow forecasting using a five-parameter conceptual watershed model and daily data from the Leaf River and French Broad catchments in the USA. In particular, this study demonstrated how DREAM can be used to analyze forcing data error during watershed model calibration. The most important conclusions are as follows:

[68] 1. Explicit treatment of forcing error during hydrologic model calibration significantly alters the posterior

distribution of watershed model parameters. This finding has significant implications for regionalization studies that attempt to relate optimized rainfall-runoff model parameters to invariant properties of the underlying catchment.

[69] 2. The DREAM algorithm provides an accurate estimate of the posterior probability density function of hydrologic model parameters, and was demonstrated to successfully solve $d = 62$ and $d = 64$ dimensional parameter estimation problems. This facilitates estimating proxy records of whole-catchment rainfall from observed discharge data, including the underlying uncertainty in inferred rainfall depths.

[70] 3. The rainfall multipliers are grouped around 1 for both the Leaf River and French Broad watersheds. The estimated rainfall from the observed discharge data is, on average, about 10% lower than the measured rainfall for both watersheds. These findings are contingent on HYMOD being an accurate representation of the hydrologic functioning of both catchments.

[71] 4. Rainfall multipliers provide important diagnostic information to quantify rainfall error, better test hydrologic theory, and diagnose model structural errors.

[72] It would be desirable to use multiple different watershed models for posterior inference to explicitly consider structural uncertainty, and reduce ambiguity about the inferred proxy records of whole-catchment rainfall. Moreover, there is a urgent need to compare our estimates of precipitation against other available rainfall information. This might require selecting another catchment for which multiple types of precipitation data are available and for which spatially distributed models can be run. Future work should also focus on extending the method presented in this paper to consider potential evapotranspiration during drying conditions of the watershed as well. An initial step in that direction is presented by *Vrugt et al.* [2008b]. Work presented in that paper shows that low precipitation amounts are generally associated with relatively high uncertainty, whereas higher rainfall amounts appear to be better defined with smaller variation among the multipliers. That finding is consistent with recent work by *Villarini and Krajewski* [2008] who, for the Brue catchment in Southwest England have shown that the standard deviation of the spatial sampling error decreases with increasing rainfall intensity.

[73] The source code of DREAM is written in MATLAB and can be obtained from the first author (vrugt@lanl.gov) upon request.

[74] **Acknowledgments.** The first author is supported by a J. Robert Oppenheimer Fellowship from the LANL postdoctoral program. We gratefully acknowledge the ideas and comments of three reviewers and students and faculty at New Mexico Tech, Socorro, that have improved the current quality of this paper. Computer support, provided by the SARA center for parallel computing at the University of Amsterdam, Netherlands, is highly appreciated.

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.

Bates, B. C., and E. P. Campbell (2001), A Markov chain Monte Carlo scheme for parameter estimation and inference in conceptual rainfall-runoff modeling, *Water Resour. Res*, *37*(4), 937–948.

Beven, K. J., and A. M. Binley (1992), The future of distributed models: Model calibration and uncert rediction, *Hydrol. Processes*, *6*, 279–298.

Boyle, D. P. (2000), Multicriteria calibration of hydrological models, Ph.D. dissertation, Dep. of Hydrol. and Water Resour., Univ. of Ariz., Tucson.

Butts, M. B., J. T. Payne, M. Kristensen, and H. Madsen (2004), An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation, *J. Hydrol.*, *298*, 242–266, doi:10.1016/j.jhydrol.2004.03.042.

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

Clark, M. P., A. G. Slater, D. E. Rupp, R. A. Woods, J. A. Vrugt, H. V. Gupta, T. Wagener, and L. E. Hay (2008), Framework for Understanding Structural Errors (FUSE): A modular framework to diagnose differences between hydrological models, *Water Resour. Res.*, *44*, W00B02, doi:10.1029/2007WR006735.

Duan, Q., S. Sorooshian, and V. Gupta (1992), Effective and efficient global optimization for conceptual rainfall-runoff models, *Water Resour. Res.*, *28*(4), 1015–1031.

Engeland, K., and L. Gottschalk (2002), Bayesian estimation of parameters in a regional hydrological model, *Hydrol. Earth Syst. Sci.*, *6*(5), 883–898.

Freer, J., K. Beven, and B. Ambroise (1996), Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach, *Water Resour. Res.*, *32*(7), 2161–2173.

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

Georgakakos, K. P., D. J. Seo, H. Gupta, J. Schaake, and M. B. Butts (2004), Towards the characterization of streamflow simulation uncertainty through multimodel ensembles, *J. Hydrol.*, *298*, 222–241.

Gupta, H. V., S. Sorooshian, and P. O. Yapo (1998), Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information, *Water Resour. Res.*, *34*(4), 751–763.

Kavetski, D., S. W. Franks, and G. Kuczera (2002), Confronting input uncertainty in environmental modeling, in *Calibration of Watershed Models*, *Water Sci. Appl.*, vol. 6, edited by Q. Duan et al., pp. 49–68, AGU, Washington, D. C.

Kavetski, D., G. Kuczera, and S. W. Franks (2006a), Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory, *Water Resour. Res.*, *42*, W03407, doi:10.1029/2005WR004368.

Kavetski, D., G. Kuczera, and S. W. Franks (2006b), Bayesian analysis of input uncertainty in hydrological modeling: 2. Application, *Water Resour. Res.*, *42*, W03408, doi:10.1029/2005WR004376.

Kirchner, J. W. (2008), Catchments as simple dynamical systems: Catchment characterization, rainfall-runoff modeling, and doing hydrology backward, *Water Resour. Res.*, doi:10.1029/2008WR006912, in press.

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

Kuczera, G., D. Kavetski, S. Franks, and M. Thyer (2006), Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters, *J. Hydrol.*, *331*, 161–177, doi:10.1016/j.jhydrol.2006.05.010.

Liu, Y., and H. V. Gupta (2007), Uncertainty in hydrologic modeling: Toward an integrated data assimilation framework, *Water Resour. Res.*, *43*, W07401, doi:10.1029/2006WR005756.

Marshall, L., D. Nott, and A. Sharma (2004), A comparative study of Markov chain Monte Carlo methods for conceptual rainfall-runoff modeling, *Water Resour. Res.*, *40*, W02501, doi:10.1029/2003WR002378.

Marshall, L. A., D. J. Nott, and A. Sharma (2006), Towards dynamic catchment modelling: A Bayesian hierarchical mixtures of experts framework, *Hydrol. Processes*, *21*, 847–861.

Mengersen, K., and C. P. Robert (2003), Population Markov chain Monte Carlo: The pinball sampler, in *Bayesian Statistics 7*, edited by J. O. Berger, A. P. Dawid, and A. F. M. Smith, pp. 277–292, Oxford Univ. Press, Oxford, U. K.

Moore, R. J. (1985), The probability-distributed principle and runoff production at point and basin scales, *Hydrol. Sci. J.*, *30*(2), 273–297.

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

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

Robert, C. P., and G. Casella (2004), *Monte Carlo Statistical Methods*, 645 pp., Springer, New York.

Roberts, G. O., and J. S. Rosenthal (2001), Optimal scaling for various Metropolis-Hastings algorithms, *Stat. Sci.*, *16*, 351–367.

Sivapalan, M. (2003), Prediction in ungauged basins: A grand challenge for theoretical hydrology, *Hydrol. Processes*, *17*, 3163–3170.

Slater, A. G., and M. P. Clark (2006), Snow data assimilation via an ensemble Kalman filter, *J. Hydrometeorol.*, *7*(3), 478–493.

ter Braak, C. J. F. (2006), A Markov chain Monte Carlo version of the genetic algorithm differential evolution: Easy Bayesian computing for real parameter spaces, *Stat. Comput.*, *16*, 239–249.

Villarini, G., and W. F. Krajewski (2008), Empirically-based modeling of spatial sampling uncetainties associated with rainfall measurements by rain gauges, *Adv. Water Resour.*, *31*, 1015–1023, doi:10.1016/j.advwatres.2008.04.007.

Vrugt, J. A., and B. A. Robinson (2007a), Improved evolutionary optimization from genetically adaptive multimethod search, *Proc. Natl. Acad. Sci. U. S. A.*, *104*, 708–711, doi:10.1073/pnas.0610471104.

Vrugt, J. A., and B. A. Robinson (2007b), 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., H. V. Gupta, W. Bouten, and S. Sorooshian (2003), A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters, *Water Resour. Res.*, *39*(8), 1201, doi:10.1029/2002WR001642.

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

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

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

Vrugt, J. A., C. J. F. ter Braak, C. G. H. Diks, B. A. Robinson, and J. M. Hyman (2008a), Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling, *Int. J. Nonlinear Sci. Numer. Simul.*, in press.

Vrugt, J. A., C. F. F. ter Braak, H. V. Gupta, and B. A. Robinson (2008b), Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?, *Stochastic Environ. Res. Risk Assess.*, in press.

M. P. Clark, NIWA, P.O. Box 8602, Riccarton, Christchurch, New Zealand.

J. M. Hyman, Mathematical Modeling and Analysis, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

B. A. Robinson, Civilian Nuclear Program Office, Los Alamos National Laboratory, Los Alamos, NM 87545, USA.

C. J. F. ter Braak, Biometris, Wageningen University and Research Centre, NL-6700 AC Wageningen, Netherlands.

J. A. Vrugt, Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM 87545, USA. (vrugt@lanl.gov)