Stoch Environ Res Risk Assess
DOI 10.1007/s00477-008-0274-y

ORIGINAL PAPER

Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling?

Jasper A. Vrugt · Cajo J. F. ter Braak · Hoshin V. Gupta · Bruce A. Robinson

Abstract In recent years, a strong debate has emerged in the hydrologic literature regarding what constitutes an appropriate framework for uncertainty estimation. Particularly, there is strong disagreement whether an uncertainty framework should have its roots within a proper statistical (Bayesian) context, or whether such a framework should be based on a different philosophy and implement informal measures and weaker inference to summarize parameter and predictive distributions. In this paper, we compare a formal Bayesian approach using Markov Chain Monte Carlo (MCMC) with generalized likelihood uncertainty estimation (GLUE) for assessing uncertainty in conceptual watershed modeling. Our formal Bayesian approach is implemented using the recently developed differential evolution adaptive metropolis (DREAM) MCMC scheme with a likelihood function that explicitly considers model structural, input and parameter uncertainty. Our results demonstrate that DREAM and GLUE can generate very similar estimates of total streamflow uncertainty. This suggests that formal and informal Bayesian approaches have more common ground than the hydrologic literature and ongoing debate might suggest. The main advantage of formal approaches is, however, that they attempt to disentangle the effect of forcing, parameter and model structural error on total predictive uncertainty. This is key to improving hydrologic theory and to better understand and predict the flow of water through catchments.

J. A. Vrugt
Center for NonLinear Studies (CNLS), Mail Stop B258,
Los Alamos National Laboratory (LANL),
Los Alamos, NM 87545, USA

J. A. Vrugt (✉)
Institute for Biodiversity and Ecosystems Dynamics,
University of Amsterdam, Amsterdam, The Netherlands
e-mail: vrugt@lanl.gov

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

H. V. Gupta
Department of Hydrology and Water Resources,
The University of Arizona, Tucson, AZ 85737, USA

B. A. Robinson
Civilian Nuclear Program Office (SPO-CNP), LANL,
Los Alamos, NM 87545, USA

1 Introduction and scope

Uncertainty quantification is currently receiving a surge in attention in hydrology, as researchers are trying to better understand what is well and what is not well understood about the watersheds that are being studied and as decision makers push to better quantify accuracy and precision of model predictions. Various methodologies have been developed in the past decade to better treat uncertainty. These approaches include state-space filtering, model averaging, and Bayesian approaches, and they differ in the underlying assumptions, mathematical rigor, and how the various sources of error are being treated (Montanari 2007).

Despite these advances, the more recent approaches for uncertainty estimation require considerable understanding of mathematics and statistics, and significant experience with implementation of these methods on a digital computer. For example, sequential filtering methodologies not only require models to be written in a state-space formulation, but also need a mathematical procedure that defines

how and which states to update when new information becomes available. This programming task can become quite difficult and cumbersome, especially in the absence of general-purpose software that enables the use of these state-of-the-art methods in a user-friendly environment. Simpler methods, on the contrary, are easier to understand and use and require less modifications to existing source codes of hydrologic models. They, therefore, have an important advantage over more sophisticated filtering and Bayesian approaches. We, therefore, posit that many researchers and practitioners will, at least for the foreseeable future, prefer to keep using simple methods for uncertainty estimation.

A relatively simple approach for uncertainty estimation is the generalized likelihood uncertainty estimation (GLUE) method of Beven and Binley (1992). This method is inspired by the Hornberger and Spear (1981) method of sensitivity analysis and operates within the context of Monte Carlo analysis coupled with Bayesian or fuzzy estimation and propagation of uncertainty. Since its introduction in 1992, GLUE has found widespread application for uncertainty assessment in many fields of study, including modeling of the rainfall-runoff transformation (Beven and Binley 1992; Freer et al. 1996; Lamb et al. 1998), soil erosion (Brazier et al. 2001), tracer dispersion in a river reach (Hankin et al. 2001), groundwater and well capture zone delineation (Feyen et al. 2001; Jensen 2003), unsaturated zone (Mertens et al. 2004), flood inundation (Romanowicz et al. 1996; Aronica et al. 2002), land–surface–atmosphere interactions (Franks et al. 1997), soil freezing and thawing (Hansson and Lundin 2006), crop yields and soil organic carbon (Wang et al. 2005), and ground radar-rainfall estimation (Tadesse and Anagnostou 2005). Recent applications of GLUE are also found in distributed hydrologic modeling (McMichael et al. 2006; Muleta and Nicklow 2005). The popularity of GLUE is probably best explained by its conceptual simplicity and relative ease of implementation, requiring no modifications to existing source codes of simulation models. In addition, GLUE can take great advantage of the property of being "embarrassingly parallel" and thus result in nearly linear speed ups on distributed computer systems.

Recent contributions to the hydrologic literature have criticized GLUE for not being formally Bayesian, resulting in parameter and predictive distributions that are statistically incoherent, unreliable, and that should therefore not be used (Christensen 2004; Montanari 2005; Mantovan and Todini 2006; Vogel et al. 2008). The GLUE method is most often used with a statistically informal likelihood function, does not attempt to find the maximum likelihood estimate of the parameters to benchmark the performance of the best model, and does not explicitly consider model errors in the derivation and communication of predictive distributions. In recent years, a strong debate has emerged in the hydrologic community between those that adhere strongly to the underlying philosophy of GLUE and believe that the method is a useful working methodology for assessing uncertainty in non-ideal cases (see Beven 2006), and researchers and practitioners that strongly oppose incorrect usage of statistics, and prefer to use coherent probabilistic approaches. The goal of this paper is to establish common ground between these two different view points, and highlight that under a variety of different conditions both Bayesian and informal Bayesian methods can result in very similar estimates of predictive uncertainty. This paper builds further on our previous work (Blasone et al. 2008) and compares informal GLUE with a formal Bayesian approach using the recently developed differential evolution adaptive metropolis (DREAM) Markov Chain Monte Carlo (MCMC) scheme (Vrugt et al. 2008a, b). The DREAM algorithm has important advantages over the shuffled complex evolution metropolis (SCEM-UA) global optimization algorithm (Vrugt et al. 2003), and maintains detailed balance and ergodicity which enables it to provide an exact Bayesian estimate of uncertainty.

The remainder of this paper is organized as follows. Section 2 provides a general overview of the inference problem considered in this paper, and presents a formal (DREAM) and informal (GLUE) Bayesian approach to estimating uncertainty of model predictions. In Sect. 3, we consider the application of these methods to hydrologic modeling using the hydrologic model (HYMOD) conceptual watershed model. In this section we are especially concerned with inference of parameter and predictive uncertainty. Finally, a summary with conclusions is presented in Sect. 4.

2 The inverse problem

Let us consider a model, f that simulates the response $\mathbf{Y} = \{y_1,...,y_n\}$ with length n of a real-world system for measured boundary $\hat{\zeta}$ and initial conditions $\hat{\phi}$, using a vector of d model parameters, $\theta = \{\theta_1,...,\theta_d\}$:

$$\mathbf{Y} = f(\theta, \hat{\zeta}, \hat{\phi}) \tag{1}$$

The model hypothesis is typically represented by a deterministic or stochastic function $f : \hat{\zeta}, \hat{\phi} \rightarrow \mathbf{Y}$ closed by the parameter vector θ (Kavetski et al. 2006a). Note that in typical time series analysis the influence of $\hat{\phi}$ on the model output diminishes with increasing distance from the start of the simulation. In those situations, it is common to use a spin-up period to reduce sensitivity to state-value initialization.

To establish whether f provides an accurate description of the underlying system it is intended to represent, it is a standard practice to confront the model-simulated response with measurements of observed system behavior,

$\hat{\mathbf{Y}} = \{\hat{y}_1, \ldots, \hat{y}_n\}$. The difference between $\mathbf{Y}$ and $\hat{\mathbf{Y}}$ defines the vector of residuals:

$$\varepsilon_i(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) = y_i(\theta|\hat{\zeta}, \hat{\phi}) - \hat{y}_i \quad i = 1, \ldots, n \tag{2}$$

The closer the residuals are to zero, the better the model represents the observational data. However, because of errors in the observed initial and boundary (forcing) conditions, (and hence $\hat{\phi}$ and $\hat{\zeta}$), structural inadequacies in the model, errors in the output measurements, $\hat{\mathbf{Y}}$ and uncertainty associated with the correct choice of θ, the residual values are not expected to go to zero.

The common approach that has historically developed is to attempt to force the residual vector to be as close to zero as possible by tuning the values of the parameters, without considering forcing and structural model uncertainty as potential sources of error. A measure that is commonly minimized during parameter estimation is the sum of squared residuals (SSR):

$$\mathrm{SSR}(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) = \sum_{i=1}^{n} \varepsilon_i(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi})^2 \tag{3}$$

This is the standard least squares (SLS) formulation. Various numerical optimization methods have been developed during the past decades to efficiently minimize this measure for d-dimensional parameter spaces (see e.g., Duan et al. 1992). Unfortunately, such algorithms only provide an estimate of the best values of θ. It would also be desirable to have an estimate of the underlying posterior probability density function (pdf) of θ, $p(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi})$. This distribution will help assess the information content of the data, and help generate predictive distributions of $\mathbf{Y}$.

One approach to estimate uncertainty of parameters, state variables, and model output prediction is through Bayesian statistics coupled with Monte Carlo sampling. The Bayesian paradigm provides a simple way to combine multiple probability distributions using Bayes theorem. In a hydrologic context, this method is admirably suited for systematically addressing and quantifying the various error sources within a single cohesive, integrated, and hierarchical manner (Kuczera and Parent 1998; Bates and Campbell 2001; Engeland and Gottschalk 2002; Vrugt et al. 2003; Marshall et al. 2004; Liu and Gupta 2007).

If we assume that the measurement errors in Eq. 2 are mutually independent (uncorrelated) and Gaussian-distributed with a constant variance, σ_e^2, the posterior pdf takes the following form:

$$p(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) = c \cdot p(\theta) \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_e^2}} \times \exp\left(-\frac{(y_i(\theta|\hat{\zeta}, \hat{\phi}) - \hat{y}_i)^2}{2\sigma_e^2}\right) \tag{4}$$

where c is a normalizing contact, and $p(\theta)$ signifies the prior distribution of θ. This distribution combines the data likelihood (multiplicative part of Eq. 4) with a prior distribution using Bayes theorem. It is convenient to maximize the logarithm of the likelihood function (or log-likelihood function) rather than the likelihood function itself, for reasons of both algebraic simplicity and numerical stability; the same parameter values that maximize one also maximize the other. The log-likelihood, ℓ of Eq. 4 is:

$$\ell(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma_e^2) - \frac{1}{2}\sigma_e^{-2} \times \sum_{i=1}^{n}(y_i(\theta|\hat{\zeta}, \hat{\phi}) - \hat{y}_i)^2 \tag{5}$$

The use of this formulation is convenient, but the assumption of uncorrelated errors is not very realistic in hydrologic modeling. The time series of residuals typically exhibit considerable non-stationarity and autocorrelation. These error characteristics need to be explicitly accounted for to result in parameter and predictive uncertainty estimates that can be considered coherent from a statistical viewpoint.

One approach to at least partially account for correlated errors is through use a first-order autoregressive (AR) scheme of the residuals:

$$\varepsilon_i = \rho\varepsilon_{i-1} + v_i \quad i = 1, \ldots, n \tag{6}$$

where ρ is the first-order correlation coefficient, and $v \sim N(0, \sigma_v^2)$ is the remaining (unexplained) error with zero mean and constant variance σ_v^2. The AR-1 corrected time series of residuals is then:

$$\delta_i(\theta, \rho|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) = \varepsilon_i(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) - \rho\varepsilon_{i-1}(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) \quad i = 1, \ldots, n. \tag{7}$$

with $\varepsilon_0 = 0$. Sorooshian and Dracup (1980) have shown how to incorporate this AR-1 model into the formulation of the log-likelihood function:

$$\ell(\theta, \rho|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\frac{\sigma_v^{2n}}{1-\rho^2} - \frac{1}{2}(1-\rho)^2 \times \sigma_v^{-2}\varepsilon_1(\theta|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi})^2 - \frac{1}{2}\sigma_v^{-2} \times \sum_{i=2}^{n}\delta_i(\theta, \rho|\hat{\mathbf{Y}}, \hat{\zeta}, \hat{\phi})^2 \tag{8}$$

Note that for $\rho = 0$, Eq. 8 automatically reduces to Eq. 5. In the Bayesian approach we will assume Jeffrey's prior for σ_v^2 and the uniform prior for ρ. The first-order AR formulation of Eq. 8 explicitly accounts for autocorrelation in the residuals, and thus the effect of model structural error. However, Eq. 8 ignores potential error in forcing conditions. Hence, $\hat{\zeta}$ is only an approximation of the true forcing conditions, ζ.

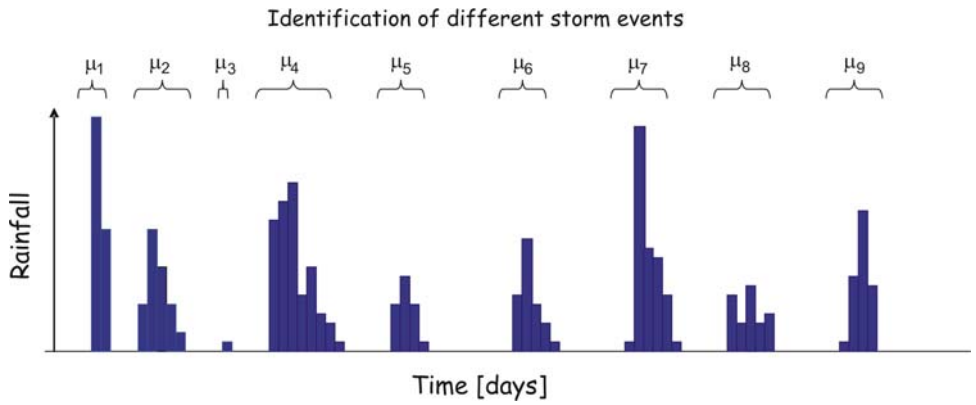

Fig. 1 Illustrative example of how rainfall multipliers are assigned to individual storm events. The values of these multipliers are estimated simultaneously with the hydrologic model parameters by minimizing the mismatch between observed and simulated catchment response. In the example considered here $\Upsilon = 9$ different rainfall evens are identified and hence nine different multipliers, u_j, $j = 1,...,9$ are used to characterize forcing uncertainty within the formal Bayesian approach

In a previous paper (Vrugt et al. 2008b), we have shown how we can treat forcing error in hydrologic modeling by assigning rainfall multipliers to each individual storm event in the forcing time series. This follows the approach introduced by (Kavetski et al. 2006a, b). Prior to calibration, individual storm events are identified from the measured hyetograph and hydrograph. A simple example of this approach is illustrated in Fig. 1. Each storm, $j = 1, \ldots, \Upsilon$ is assigned a different rainfall multiplier μ_j, and these scalar values are added to the vector of model parameters θ and ρ to be optimized:

$$\ell(\theta, \rho, \mu|\hat{\mathbf{Y}}, \hat{\phi}) = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln\frac{\sigma_v^{2n}}{1-\rho^2} - \frac{1}{2}(1-\rho)^2 \times \sigma_v^{-2}\varepsilon_1(\theta, \mu|\hat{\mathbf{Y}}, \hat{\phi})^2 - \frac{1}{2}\sigma_v^{-2} \times \sum_{i=2}^{n}\delta_i(\theta, \rho, \mu|\hat{\mathbf{Y}}, \hat{\phi})^2 \quad (9)$$

Note that the individual storms are clearly separated in time in the hypothetical example considered in Fig. 1. 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 streamflow data must be combined to identify different rainfall events. It can be quite difficult in practice to identify and quantify individual error sources, because input, parameter and structural error are likely to interact strongly through multiplication in Bayes law and nonlinear processing of input errors by the model necessarily leads to structured, non-stationary residuals (Beven et al. 2008). Nevertheless, to improve hydrologic theory through modeling, it is necessary that we attempt to separate and quantify individual error sources. This will help to find out what parts of the model can potentially be improved.

Unfortunately, in many hydrologic studies, the probability distribution defined in Eq. 9 cannot be derived through analytical means nor by analytical approximation. Iterative approximation methods such as Monte Carlo sampling are therefore needed to generate a sample from the posterior pdf. In the following two sections we discuss two methods that have found widespread use in the field of hydrology to estimate parameter and predictive distributions within a Bayesian context.

2.1 Generalized likelihood uncertainty estimation (GLUE)

Simple assumptions about the error characteristics of the residuals in Eq. 2 are convenient in applying statistical theory but are not often borne out in the actual calibration time series of residual errors which may show changing bias, variance (heteroscedasticity), skewness, and correlation structures under different hydrologic conditions (and for different parameter sets). For linear systems it is known that ignoring such characteristics, or wrongly specifying the structure of the error model, will lead to bias in the estimates of parameter values. There does not appear to be a way around this problem without making some very strong (and generally difficult to justify) assumptions about the nature of the errors (Beven 2006).

The origins of the GLUE method lie in trying to deal with uncertainty estimation problems for which simple theoretical likelihood assumptions do not seem appropriate. The GLUE methodology rejects the traditional statistical basis for the likelihood function in favor of finding a set of representations (model inputs, model structures, model parameter sets, model errors) that are behavioral in the sense of being acceptably consistent with the (non-error-free) observations. To this end, it uses an informal

likelihood measure to avoid over conditioning and exclude parts of the model (parameter) space that might provide acceptable fits to the data and be useful in prediction. Many different informal measures have been used within the context of GLUE. Of these, the inverse error variance, introduced by Beven (1989) and Beven and Binley (1992), is most commonly used to measure the closeness between model predictions and observations:

$$L(\theta|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi}) = (\sigma_e^2)^{-T} = \left(\frac{\mathrm{SSR}(\theta|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi})}{n-2}\right)^{-T} \quad (10)$$

where T is a parameter chosen by the user. Note that when $T = 0$, every simulation will have equal likelihood and when $T \to \infty$ the emphasis will be placed on a single best simulation, while the other solutions are assigned a negligible likelihood. To estimate parameter and model output uncertainty, the GLUE method works as follows:

1. Draw a sample of points Θ of size N using the specified prior distribution, $p(\theta)$.
2. Compute the likelihood $L(\theta^i|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi})$ of each point of Θ, $i = 1,...,N$.
3. Define a cutoff threshold to separate good solutions from non-behavioral parameter combinations of Θ. Collect the k behaviorial solutions in $\mathbf{D}$.
4. Normalize the likelihood values of the behavioral solutions, $i = 1,...,k$ of $\mathbf{D}$, $\bar{L}(\mathbf{D}^i|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi}) = L(\mathbf{D}^i|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi})/\sum_{i=1}^{k} L(\mathbf{D}^i|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi})$ so that $\sum_{i=1}^{k} \bar{L}(\mathbf{D}^i|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi}) = 1$.
5. Assign each output prediction $\mathbf{Y}^i$, $i = 1,...,k$ of $\mathbf{D}$, probability $\bar{L}(\mathbf{D}^i|\hat{\mathbf{Y}},\hat{\zeta},\hat{\phi})$.
6. Sort the $\mathbf{Y}^i$, $i = 1,...,k$ with their corresponding probabilities to create the pdf of the model output prediction, and use these to generate uncertainty intervals.

To summarize, a large number of runs are performed for a particular model with different combinations of the parameter values, chosen randomly from prior parameter distributions. By comparing predicted and observed responses, each set of parameter values is assigned a likelihood value, i.e. a function that quantifies how well that particular parameter combination (or model) simulates the system. Higher values of the likelihood function typically indicate better correspondence between the model predictions and observations. Based on a cutoff threshold, the total sample of simulations is then split into behavioral and non-behavioral parameter combinations. This threshold is either defined in terms of a certain allowable deviation of the highest likelihood value in the sample, or more commonly as a fixed percentage of the total number of simulations. The likelihood values of the retained solutions are then rescaled to obtain the cumulative distribution function (cdf) of the output prediction. The deterministic model prediction is then typically given by the median of the output distribution, and the associated uncertainty is derived from the cdf, normally chosen at the 5 and 95% prediction quantiles in most of the published GLUE studies. The likelihood weights of the GLUE procedure attempt to approximate and reflect all sources of error in the modeling process and allow the uncertainties associated with those errors to be carried forward into the predictions. Note that the limits of acceptability approach developed in Beven (2006) can be applied at every single time step if required before combination into a single likelihood weight.

Because of its conceptual simplicity and ease of implementation, the GLUE method has found widespread use. If used with a formal Bayesian likelihood function such as Eq. 4, GLUE generally will result in very similar estimates of parameter and predictive uncertainty as Markov Chain Monte Carlo simulation through DREAM. However, DREAM will have a much better efficiency in finding "acceptable" models as it uses adaptive proposal updating to search for high quality solutions. Use of a simple uniform sampling distribution of model parameters over a relatively large region, as typically done in GLUE, can result in an algorithm that, even after billions of model evaluations, may only have generated a handful of good solutions (Iorgulescu et al. 2005), even if Latin Hypercube sampling has been used.

Most applications of GLUE, however presented in the hydrologic literature and beyond use an informal likelihood function to distinguish between behavioral and non-behavioral solutions (or models). An informal likelihood function such as Eq. 10 does not properly account for the number of measurements n used to condition the parameter estimates. A small number of measurements in Eq. 10 is considered as informative as a data set that contains many more observations and spans a much wider range of conditions. This is counter intuitive, but is done to avoid over-conditioning and thus ensure that parameter uncertainty reflects total uncertainty. Each model implicitly carries along an error series that is known exactly in calibration, and assumed to have similar characteristics in prediction (evaluation). Moreover, the cutoff threshold introduced in step (3) to separate behavioral from non-behavioral is entirely subjective, and not based on proper statistical arguments. But if it is accepted that equifinality, input and model structural errors are important issues, then GLUE is a useful working paradigm to avoid overconditioning and to summarize parameter and predictive distributions. Note that GLUE can be used with sequential updating which should further reduce chances of overfitting (Beven et al. 2008).

2.2 Markov Chain Monte Carlo Sampling with DREAM

A more sophisticated and elegant approach to estimate the posterior pdf of the parameters and model output prediction is MCMC simulation. Not only has this methodology a proper statistical foundation, but it is also more efficient than GLUE in finding behavioral models. Unlike GLUE, MCMC simulation uses a formal likelihood function, appropriately samples the high-probability-density region of the parameter space, and separates behavioral from non-behaviorial solutions using a cutoff threshold that is based on the sampled probability mass, and thus underlying probability distribution. Vrugt et al. (2008a, b) have recently presented a novel adaptive MCMC algorithm to efficiently estimate the posterior pdf of parameters in complex, high-dimensional sampling problems. This method, entitled 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. This scheme is an adaptation of the SCEM-UA global optimization algorithm (Vrugt et al. 2003) and has the advantage of maintaining detailed balance and ergodicity while showing excellent efficiency on complex, highly nonlinear, and multimodal target distributions (Vrugt et al. 2008a). The code of DREAM is given below. For convenience, we assemble the parameters θ, ρ and μ into a single vector $\mathbf{x}$.

1. Draw an initial population $\mathbf{X}$ of size N, typically $N = d$ or $2d$, using the specified prior distribution. The symbol d signifies the number of parameters to be estimated.
2. Compute the density $p(\mathbf{x}^i|\hat{\mathbf{Y}}, \hat{\phi})$ of each point of $\mathbf{X}$, $i = 1,...,N$ using the antilog of Eq. 9.

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

3. Generate a candidate point, $\mathbf{z}^i$ in chain i,

$$\mathbf{z}^i = \mathbf{x}^i + \gamma(\delta) \cdot \left(\sum_{j=1}^{\delta} \mathbf{x}^{r(j)} - \sum_{n=1}^{\delta} \mathbf{x}^{r(n)} \right) \tag{11}$$

where δ signifies the number of pairs used to generate the proposal (candidate point), and $r(j)$, $r(n) \in \{1,...,N\}$; $r(j) \neq r(n) \neq i$. The value of γ depends on the number of pairs used to create the proposal. By comparison with random walk metropolis, a good choice for $\gamma = 2.38/\sqrt{2\delta d_{\text{eff}}}$, with $d_{\text{eff}} = d$, but potentially decreased in the next step.

4. Replace each element, $j = 1,...,d$ of the proposal $\mathbf{z}_j^i$ with $\mathbf{x}_j^i$ using a binomial scheme with crossover probability CR,

$$\mathbf{z}_j^i = \begin{cases} \mathbf{x}_j^i & \text{if } U \leq 1 - \text{CR}, \quad d_{\text{eff}} = d_{\text{eff}} - 1 \\ \mathbf{z}_j^i & \text{otherwise} \end{cases} \qquad j = 1, \ldots, d \tag{12}$$

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

5. Compute $p(\mathbf{z}^i|\hat{\mathbf{Y}}, \hat{\phi})$ and accept the candidate point with Metropolis acceptance probability, $\alpha(\mathbf{x}^i,\mathbf{z}^i)$,

$$\alpha(\mathbf{x}^i, \mathbf{z}^i) = \begin{cases} \min\left(\frac{p(\mathbf{z}^i|\hat{\mathbf{Y}},\hat{\phi})}{p(\mathbf{x}^i|\hat{\mathbf{Y}},\hat{\phi})}, 1\right) & \text{if } p(\mathbf{x}^i|\hat{\mathbf{Y}}, \hat{\phi}) > 0 \\ 1 & \text{if } p(\mathbf{x}^i|\hat{\mathbf{Y}}, \hat{\phi}) = 0 \end{cases} \tag{13}$$

6. If the candidate point is accepted, move the chain, $\mathbf{x}^i = \mathbf{z}^i$; otherwise remain at the old location, $\mathbf{x}^i$.

END FOR (CHAIN EVOLUTION)

7. Remove potential outlier chains using the inter-quartile-range (IQR) statistic.
8. Compute the Gelman–Rubin, R_{stat} convergence diagnostic.
9. If $R_{\text{stat}} \leq 1.2$, stop, otherwise go to CHAIN EVOLUTION.

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. The members of $\mathbf{X}$ are used to globally share information about the progress of the search of the individual chains. Hence, at every individual step, the points in $\mathbf{X}$ contain the most relevant information about the search. 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. Convergence of the individual chains is monitored using the R-statistic of Gelman and Rubin (1992). Detailed balance and ergodicity of DREAM have been proved in Vrugt et al. (2008a).

We did not add the error variance σ_v^2 to the parameter vector $\mathbf{x}$. The reason is that the posterior distribution of σ_v^2 given the other parameters is known to be inverse chi-square with n degrees of freedom and scale s with

$$s^2 = \frac{1}{n}\left(\varepsilon_1^2(1-\rho^2) + \sum_{i=2}^{n} \delta_i^2\right). \tag{14}$$

We can therefore update σ_v^2 after step 6 in the DREAM algorithm by Gibbs sampling, as follows. We draw a value z from a chi-squared distribution with n degrees of freedom and calculate $(\sigma_e^2)^j = \frac{n}{z}s^2$.

2.3 Predictive inference using MCMC simulation with DREAM

The posterior pdf of the model parameters derived with DREAM contains all required information to summarize predictive uncertainty. An estimate of the predictive distribution for $f(\mathbf{x}, \hat{\phi})$ is obtained by evaluating the model output, $\mathbf{Y}$ for each $\mathbf{x}^j$ of J draws derived with DREAM after convergence has been achieved to a stationary distribution. The so-obtained values $\{\mathbf{Y}^j, j = 1,...,J\}$ are summarized in the desired way, e.g. by calculating the 2.5 and 97.5% percentiles of each individual model prediction, y^i, $i = 1,...,n$. This predictive distribution only includes the effect of parameter uncertainty. The remaining (unexplained) error is assumed to be additive and can be summarized as follows.

For each model outcome, $\{\mathbf{Y}^j, j = 1,...,J\}$ the residual error $\varepsilon_j \sim N(0,(\sigma_v^2)^j/(1-(\rho^2)^j))$ is added to the prediction. The desired output percentiles can be summarized in a similar way as described in the previous paragraph. A slightly more efficient approach is to draw the outcome variable y^j for each $\mathbf{x}^j$ directly from a Student distribution with n degrees of freedom, mean $f(\mathbf{x}^j, \hat{\phi})$ and variance $s^2/(1-(\rho^2)^j)$, where s^2 is calculated for the current draw using Eq. 14. An even more precise approach for obtaining the 95% prediction uncertainty intervals including parameter, model and measurement error is presented in the Appendix.

3 Case study

We compare formal (DREAM) and informal (GLUE) Bayesian inference to parameter and model output uncertainty estimation by application to streamflow forecasting using the HYMOD conceptual watershed model. This study is used to demonstrate that formal and informal Bayesian approaches can yield very similar estimates of total predictive uncertainty.

3.1 Rainfall-runoff modeling

In this study, we use the HYMOD conceptual watershed model which is schematically presented in Fig. 2. HYMOD is a hierarchical and parsimonious rainfall-runoff model whose parameters are thought to vary between watersheds. This model has been used in a number of studies in the past and has five parameters that need to be specified by the user (Table 1). Inputs to the model include mean areal precipitation (MAP), and potential evapotranspiration (PET), while the outputs are estimated channel inflow. The HYMOD model has been discussed extensively in many previous papers that study streamflow forecasting and automatic model calibration (Boyle 2000; Wagener et al. 2001; Vrugt et al. 2003). Details of the model can be found therein.

To compare GLUE and DREAM we use historical data from the Leaf River (1,950 km^2) and French Broad (767 km^2) watersheds in the USA. The data consists of

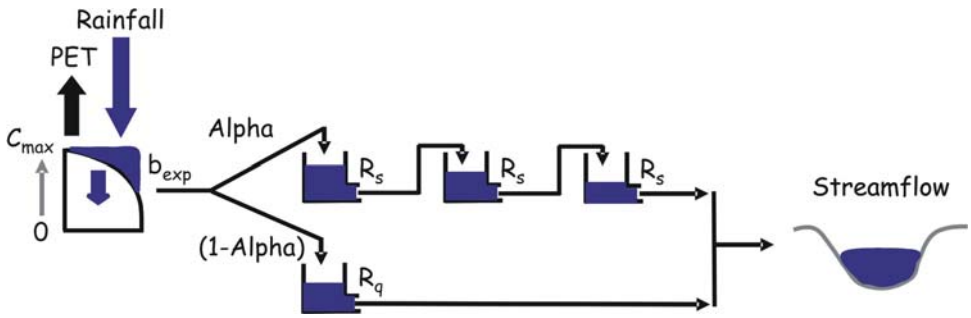

Fig. 2 Schematic representation of the HYMOD conceptual watershed model

Table 1 Prior ranges and description of the hydrologic model (HYMOD) parameters and rainfall multipliers

Parameter	Description	Minimum	Maximum
C_{max} (mm)	Maximum storage in watershed	1.00	500.00
b_{exp}	Spatial variability of soil moisture storage	0.10	2.00
Alpha	Distribution factor between two reservoirs	0.10	0.99
R_s (days)	Residence time slow flow reservoir	0.001	0.10
R_q (days)	Residence time quick flow reservoir	0.10	0.99
ρ	First-order correlation coefficient	−1.00	1.00
$\mu_j, j = 1,\ldots,\Upsilon$	Rainfall multipliers	0.25	2.50

MAP (mm/day), PET (mm/day), and streamflow (m^3/s). For both catchments 5 years of data is used for model calibration, whereas the remainder of the data is used for evaluation purposes. The calibration data set consists of October 1, 1953 to September 30, 1958 for the Leaf River and spans the period of October 1, 1954 to September 30, 1959 for the French Broad river. In this 5-year calibration time series, a total of $\Upsilon = 57$ and $\Upsilon = 59$ storm events are identified for the Leaf River (October 1, 1954–September 30, 1959) and French Broad (October 1, 1953–September 30, 1958) watersheds, respectively. This results in a total of $d = 63$ (Leaf River) and $d = 65$ (French Broad) parameters to be estimated within the formal Bayesian inference procedure using DREAM. The upper and lower bounds that define the prior uncertainty ranges of the HYMOD model parameters, first-order correlation coefficient and rainfall multipliers are given in Table 1. These ranges are based on previous work (HYMOD parameters), mathematics (correlation coefficients) or analysis of rain-gauge data (multipliers) to make sure that the parameter values remain hydrologically realistic.

To approximate the posterior pdf of the HYMOD model parameters, storm multipliers and first-order correlation coefficient in the likelihood function of Eq. 9, a total of 2,000,000 HYMOD model evaluations are performed with DREAM using uniform prior ranges over the hypercube specified in Table 1. We use $N = 100$ different Markov chains. In GLUE, a sample size of $N = 100{,}000$ is used with a value of $T = 1$ in the informal likelihood function of Eq. 10 and cutoff threshold in step (3) as the best 1% of the sample. These are rather standard settings with GLUE and in the present context will result in a total of $k = 1{,}000$ different behavioral solutions present in $\mathbf{D}$. Similar results with GLUE and DREAM are obtained for larger sample sizes.

To stabilize the total error variance, σ_v^2 and reduce heteroscedasticity we use a Box–Cox transformation (Box and Cox 1964) of the simulated and measured streamflow data:

$$\tau(\mathbf{Y}, \lambda) = \begin{cases} (\mathbf{Y}^{\lambda} - 1)/\lambda & \text{if } \lambda \neq 0 \\ \ln(\mathbf{Y}) & \text{if } \lambda = 0 \end{cases} \tag{15}$$

using $\lambda = 0.3$, which is consistent with previous studies (Misirli et al. 2003; Vrugt et al. 2003, 2006).

Figure 3 presents histograms of the HYMOD model parameters using the formal (top panels) and informal (bottom panels) Bayesian inference considered here for the Leaf River streamflow time series. The x-axis in each graph is fixed to the prior range of each individual parameter, to facilitate pairwise comparison of the results of the formal and informal Bayesian approaches. For DREAM, the last 20% of the samples in each individual chain are used to compute and summarize the marginal densities, whereas for GLUE the marginal frequencies of the $k = 1{,}000$ different behavioral solutions are plotted.

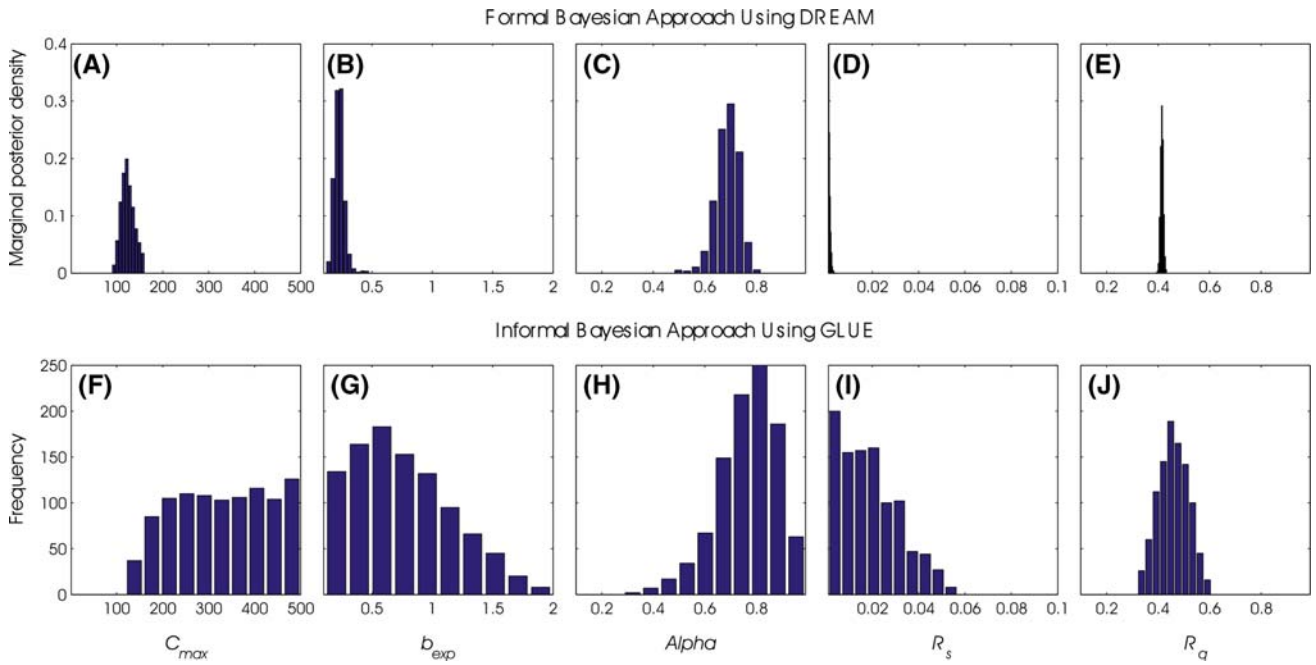

Fig. 3 Histograms of the HYMOD model parameters inferred using a formal likelihood function (*top panels* **a**–**e**) which explicitly considers input, parameter, and model structural error, and informal likelihood function (*bottom panels* **f**–**j**) that maps all uncertainty onto the parameter space. The model parameters are much better identifiable when using a formal Bayesian approach for statistical inference and analysis. Is equifinality the outcome of a weak inference procedure that lumps all uncertainty onto the model parameters?

The histograms in the top panels show that formal Bayesian inference results in parameter distributions that are well identified and encompass only a relatively small region interior to the prior uncertainty bounds. Note that the recession parameters of the quick and slow flow tanks are particularly well defined with very small dispersion around the mode of their respective distributions. Hence, this is relative to the prior uncertainty ranges. Nevertheless, these results demonstrate that the explicit treatment of forcing data error and model structural inadequacies through the use of $d = 58$ additional parameters ($\Upsilon = 57$ storm multipliers and one first-order correlation coefficient) in the definition of the likelihood function in Eq. 9 does not negatively affect the identifiability of the HYMOD model parameters. They remain well calibrated with relatively tight uncertainty bounds, and small correlation among the individual parameters (not shown). On the contrary, using an informal Bayesian approach with GLUE results in parameter distributions that are much wider and almost cover the entire prior defined hypercube of the individual parameters. Implicit projection of forcing and structural uncertainty onto the HYMOD model parameters gives rise to what Beven et al. in a series of papers since 1992 have called equifinality (Beven 1993). Qualitatively similar findings, as presented here, are also found for the French Broad watershed.

Although not further demonstrate herein, the optimized distributions of the first-order correlation coefficient are approximately Gaussian with 95% uncertainty bounds ranging between 0.76 and 0.85 for the Leaf River and 0.39 and 0.48 for the French Broad watershed. These values of ρ confirm the presence of significant autocorrelation between the error residuals, and establish a clear need for explicit modeling of the (non-random) input and model structural errors. The finding that ρ is relatively well defined is encouraging as it provides support for the claim that within the context of our assumptions model structural and input error are identifiable from the observed streamflow time series. Separating these two error sources is necessary to be able to understand if, and what parts of, the model can be improved. This is the key to improving hydrologic theory.

To provide more insights into the values of the rainfall multipliers, consider Fig. 4, which presents boxplots of the sampled rainfall multipliers for the Leaf River (top panel) and French Broad (bottom panel) catchments. These boxplots are created using the last 200,000 samples generated with DREAM in the $N = 100$ parallel chains. The marginal pdfs of the multipliers vary widely between individual

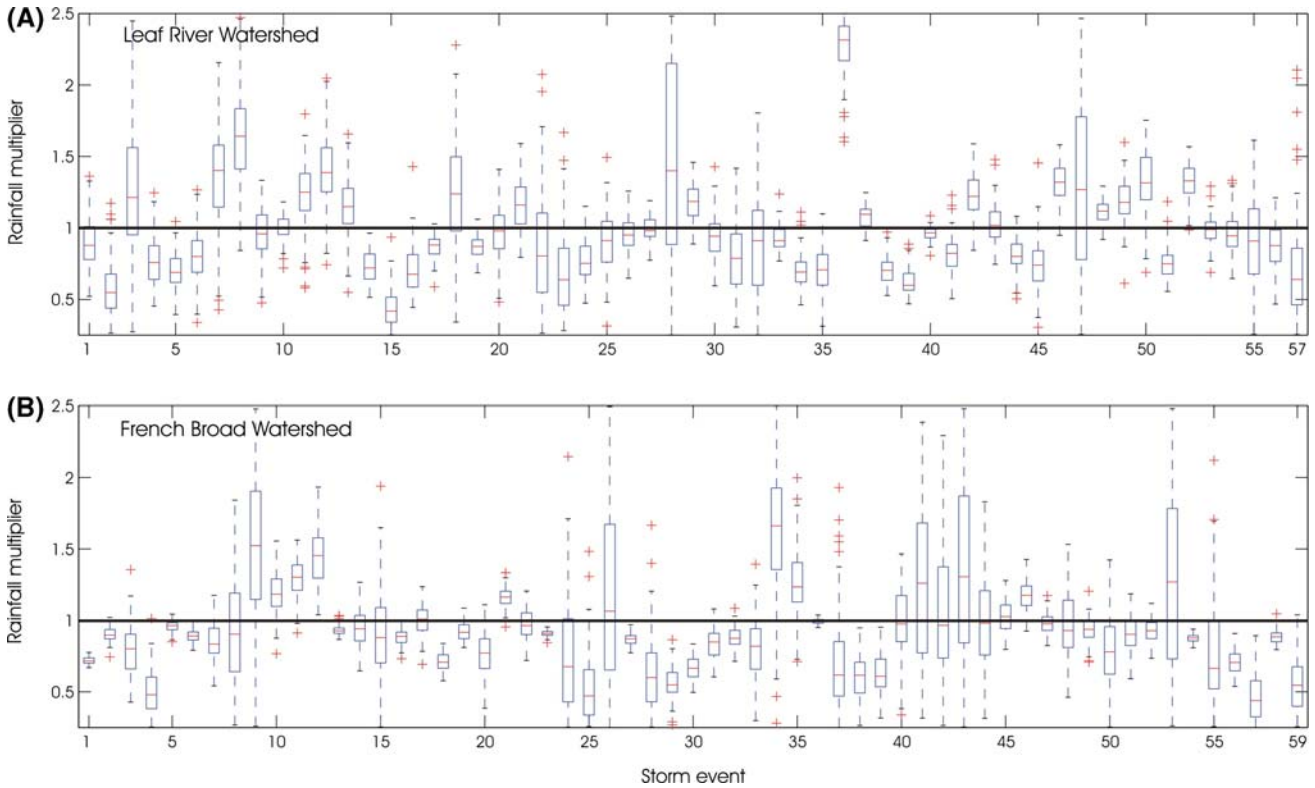

Fig. 4 Marginal posterior distribution of the rainfall multipliers for the **a** Leaf River, and **b** French Broad watersheds. These results are derived with DREAM using a total of 2,000,000 function evaluations. The *solid black lines* indicate no adjustment to the observed rainfall depths with multiplier values of 1 across both plots

storm events. Some events are very well defined, while others show considerable uncertainty. For instance, compare the boxplots of μ_{46} and μ_{47} for the Leaf River, and μ_{14} and μ_{15} for the French Broad watershed. These adjacent storms differ substantially in their posterior width, but exhibit approximately similar mean values. The overall mean posterior value of the storm multipliers is $\overline{\mu} = 0.99$) for the Leaf River and $\overline{\mu} = 0.93$ 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 amounts 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 $\Upsilon = 57$ and $\Upsilon = 59$ storm multipliers, simultaneous with the five HYMOD model parameters and first-order correlation coefficient.

Most of the storm multipliers are clustered in the vicinity of 1 for both catchments. This illustrates that the measured rainfall is on average unbiased and generally consistent in pattern and depth with the estimated rainfall record derived from the streamflow 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 a good proxy of whole-catchment precipitation for both watersheds.

Up to now, we have only discussed the parameter distributions as a main interest of the Bayesian inference, without recourse to examining the predictive uncertainty of the HYMOD model. Figure 5 illustrates how the marginal posterior pdf of the parameters ($p(\mathbf{x}|\hat{\mathbf{Y}}, \hat{\phi})$: DREAM) and behavioral solutions (**D**: GLUE) translates into 95% streamflow predictive uncertainty for a representative portion of the calibration (left column) and evaluation (right column) period for the Leaf River watershed. In the case of DREAM (top panels), the 95% prediction uncertainty of the HYMOD model predictions due to parameter uncertainty is indicated with the dark gray region, whereas the remaining prediction error is represented with the light gray region. For GLUE (bottom panels) only total error (due to parameter uncertainty) is assessed, and the streamflow uncertainty ranges denote 95% prediction quantiles.

The HYMOD model forecasts generally track the streamflow observations very well, especially when using the formal Bayesian inference. This is to be expected because individual rainfall events can be perturbed in their precipitation amounts to better match the hydrograph. Qualitatively, there is a strong agreement between the estimates of streamflow prediction uncertainty derived with

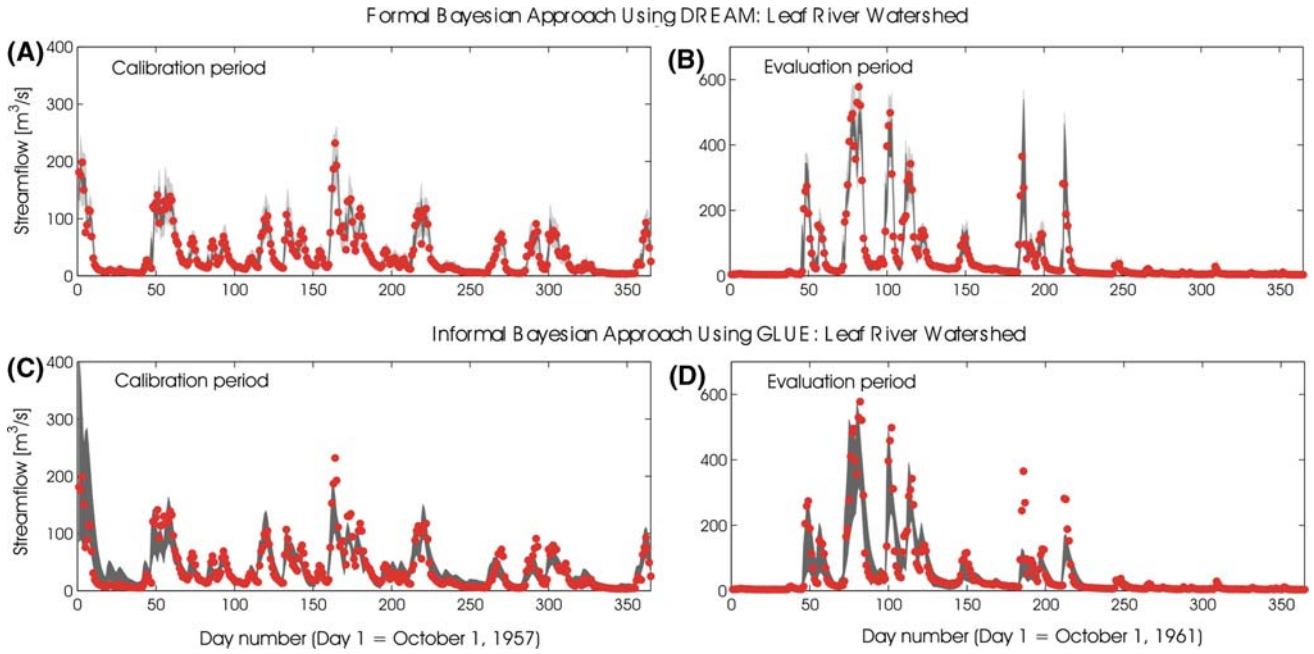

Fig. 5 Streamflow prediction uncertainty ranges derived with DREAM (*top panels*) and GLUE (*bottom panels*) for a representative portion of the calibration (*left column*) and evaluation period (*right column*) for the Leaf River watershed. In each DREAM graph, the *dark gray* region represents the 95% confidence intervals of the output prediction due to parameter uncertainty, whereas the *light gray* region represents the additional 95% ranges of the prediction uncertainty. For GLUE the 95% prediction quantiles are presented. The *solid circles* denote the streamflow observations

DREAM and GLUE, although the streamflow ranges of DREAM are slightly smaller and provide a better coverage at various rainfall events. For instance, consider the three storm events in the evaluation period between days 185 and 220. The informal Bayesian approach severely underestimates the actual streamflow data because error in rainfall is not explicitly considered within GLUE. This is an interesting result, because GLUE has often been criticized in the literature for grossly overestimating the actual uncertainty observed in the calibration data. Thus, GLUE can significantly underestimate total predictive uncertainty when input errors are large.

The formal Bayesian approach is less prone to errors in the measured forcing data, because these errors are explicitly considered through MCMC. When using DREAM, parameter uncertainty appears to be a rather small contribution to total uncertainty, with the exception of certain rainfall events during the evaluation period. This is because of incomplete knowledge of the rainfall multipliers outside the calibration period. These multipliers are assigned prior to each individual storm event by drawing from a specified probability distribution. The properties of this distribution are inferred using the calibration streamflow time series. How this is done is discussed below.

Figure 6 presents a scatter plot of the standard deviation of the rainfall multipliers as a function of the observed rainfall for the 5-year calibration period of the (a) Leaf River, and (b) French Broad watersheds. The standard deviation of the multipliers for each individual rainfall event is computed using the last 25 samples generated in each individual chain. This results in a total of $J = 2{,}500$ draws of multipliers from the posterior distribution. Both scatter plots depict a strongly nonlinear hyperbolic relationship between the actual measured precipitation and the standard deviation of the multipliers. 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. This finding is consistent with the 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. Note that we arrive at this conclusion based on the observed streamflow data only. This highlights the strength of a (formal) Bayesian approach that disentangles various error sources. To further benchmark the reasonableness of the rainfall error characteristics in Fig. 6, future work should include analysis of the spatial variability of rain-gauge measurements in both watersheds, as well as a comparison of the optimized rainfall depths against radar data. This is beyond the scope of the current paper.

The dotted black lines in Fig. 6a and b present the average standard deviation, σ_μ of all values of the multipliers. This information, albeit a bit crude is used to generate an ensemble of rainfall records during the evaluation period. To this end, we first draw 2,500 different rainfall multipliers for each individual storm event in the evaluation period of both watersheds using a Gaussian distribution with mean $\overline{\mu}$, and standard deviation σ_μ, $N(\overline{\mu}, \sigma_\mu)$. Using the information from Figs. 4 and 6, we use $\overline{\mu} = 0.99$ and $\sigma_\mu = 0.18$ for the Leaf River, and $\overline{\mu} = 0.93$ and $\sigma_\mu = 0.13$ for the French Broad watershed. We then combine each of these 2,500 multiplier vectors for both watersheds with the observed rainfall record, which results in an ensemble of 2,500 different rainfall hyetographs for the evaluation period for the Leaf River and French Broad. Finally, each rainfall hyetograph is assigned a posterior combination of the HYMOD model parameters and first-order correlation coefficient derived from calibration to create an ensemble of 2,500 different streamflow hydrographs for both data sets. Note that by setting $\overline{\mu}$ equal

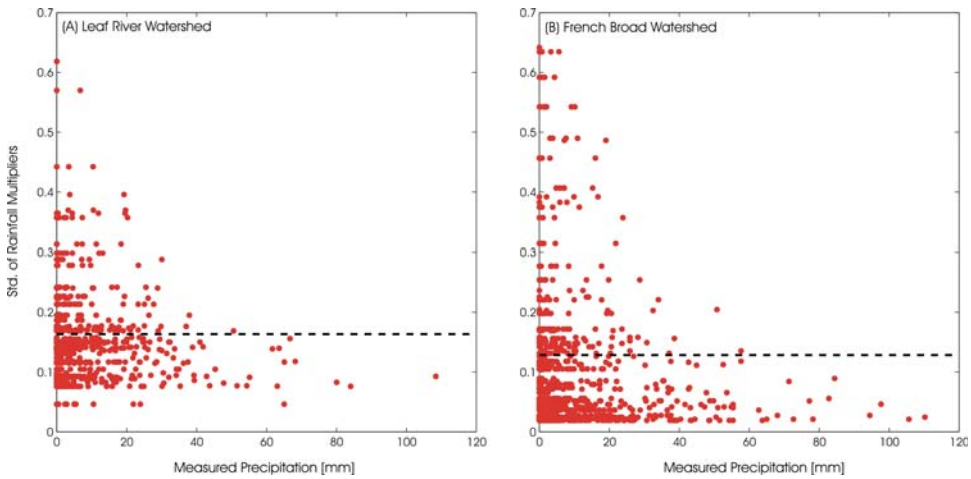

Fig. 6 The DREAM inferred standard deviation of the rainfall multipliers as a function of the observed rainfall for the **a** Leaf River, and **b** French Broad watersheds. The *dotted black line* denotes the average standard deviation that is used to generate ensembles of precipitation records during the evaluation period

to the overall posterior mean of the multipliers found during the calibration period, any potential bias in the measured rain gauge data is removed.

Figure 7 presents streamflow prediction uncertainty bounds derived with the formal (top row) and informal (bottom row) Bayesian approaches for the French Broad watershed. The left column depicts the results for the calibration period, whereas the right two plots correspond to the evaluation period. The results presented here are qualitatively very similar to those previously presented in Fig. 5 for the Leaf River watershed. The HYMOD predictions generally provide a good fit to the observed streamflow time series, and the total uncertainty ranges derived with DREAM and GLUE show a relatively close correspondence. Notice, however that GLUE has a tendency to overestimate the actual streamflow uncertainty during rainfall events. This is clearly visible in the evaluation period between days 160 and 180. Although, the formal and informal Bayesian approaches used here differ fundamentally in their underlying philosophy and representation of error, both methods receive quite similar performance in terms of ensemble spread and forecast.

This is further demonstrated in Table 2 that summarizes the probabilistic properties of the streamflow ensemble derived with the formal and informal Bayesian analyses considered herein. The coverage (%) measures the percentage of streamflow observations contained in the 95% uncertainty bounds (DREAM) or 95% prediction quantiles (GLUE), whereas the spread (m^3/s) quantifies the width of the prediction uncertainty intervals. A significant departure from a 95% coverage would indicate that the predictive uncertainty is either under- or overestimated, and would call into question the validity of the modeling approach for performing accurate probabilistic streamflow forecasting.

Table 2 Coverage (%) and spread (m^3/s) of the 95% streamflow prediction ranges associated with the total uncertainty estimated with DREAM (consisting of parameter and remaining residual error) and GLUE (parameter error only) for the Leaf River and French Broad watersheds. A distinction is made between the calibration and evaluation periods

Method	Leaf River watershed		French Broad watershed	
	Coverage	Spread	Coverage	Spread
Calibration period				
DREAM	94.2	18.2	94.8	15.9
GLUE	76.9	20.6	88.4	17.1
Evaluation period				
DREAM	92.2	30.3	93.2	18.4
GLUE	72.1	30.9	88.8	22.8

The results presented in this Table highlight a number of interesting results. The ensemble spread derived with the formal Bayesian approach is statistically coherent with a coverage of the streamflow observations that ranges

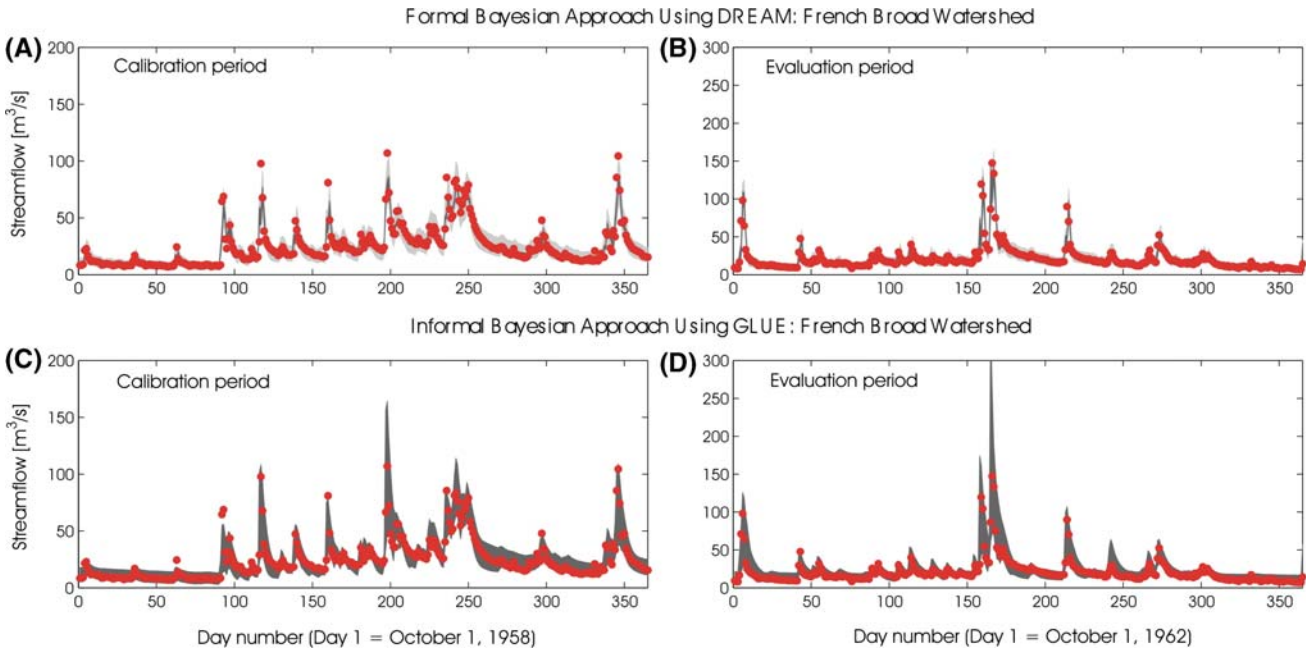

Fig. 7 Streamflow prediction uncertainty ranges derived with DREAM (*top panels*) and GLUE (*bottom panels*) for a representative portion of the calibration (*left column*) and evaluation period (*right column*) for the French Broad watershed. In each DREAM graph, the *dark gray* region represents the 95% confidence intervals of the output prediction due to parameter uncertainty, whereas the *light gray* region represents the additional 95% ranges of the prediction uncertainty. For GLUE the 95% prediction quantiles are presented. The *solid circles* denote the streamflow observations

between 92 and 95% at the 95% prediction level. This is an encouraging result, because it illustrates that our characterization of rainfall error during the evaluation period is consistent with the statistical properties of the streamflow observations. On the contrary, the prediction quantiles derived with GLUE underestimate the actual uncertainty in the streamflow measurements with a coverage that ranges between 72 and 89%. This seems counter-intuitive because the widths of the streamflow uncertainty bounds are, on average, about 10% larger with GLUE than with DREAM. The predictive pdf generated with DREAM is simply sharper and encompasses a larger percentage of the streamflow observations. This is a desirable characteristic for streamflow forecasting.

The results presented here warrant the conclusion that formal and informal Bayesian methods can receive very similar estimates of total predictive uncertainty. This is a rather unexpected result, considering that both methods rely on completely different philosophies and mathematical rigor. The formal Bayesian approach has its roots within classical statistical theory and applies formal mathematics and MCMC simulation to infer parameter and predictive distributions. The informal Bayesian approach (GLUE) makes use of subjective likelihood measures or probabilities and uses simple Monte Carlo sampling to estimate parameter and predictive uncertainty.

If the interest is in estimating total predictive uncertainty, there are several advantages in using GLUE over more formal Bayesian approaches. The method is very easy to implement and use, and is computationally efficient. For instance, in the examples considered here, GLUE is about 20 times more efficient than MCMC simulation with DREAM. The main disadvantage of GLUE, however is that it does not attempt to separate the effects of forcing, parameter and structural error on total predictive uncertainty. This makes it impossible to pinpoint what elements of the model are most uncertain and require improvement. Rather, the user of GLUE is left with a total estimate of uncertainty.

Figure 8 presents sample autocorrelation functions of the residuals for the evaluation period using the formal (top panels) and informal (bottom panels) Bayesian approaches for statistical inference of model parameter and output predictions. The mean posterior residuals are used. Similar to our previous graphs, the left two panels illustrate the results for the Leaf River, whereas the right two panels plot the results for the French Broad watershed. Note that the results are quite similar for both watersheds. Significant autocorrelation between the error residuals at the first lag (between 0.50 and 0.75) is found for GLUE. In the case of DREAM, the AR-1 model reduces the temporal correlation between the residuals.

Finally, Table 3 presents summary statistics of the one-day-ahead streamflow forecasts of the HYMOD model using the formal (DREAM) and informal (GLUE) Bayesian analyses for the Leaf River and French Broad

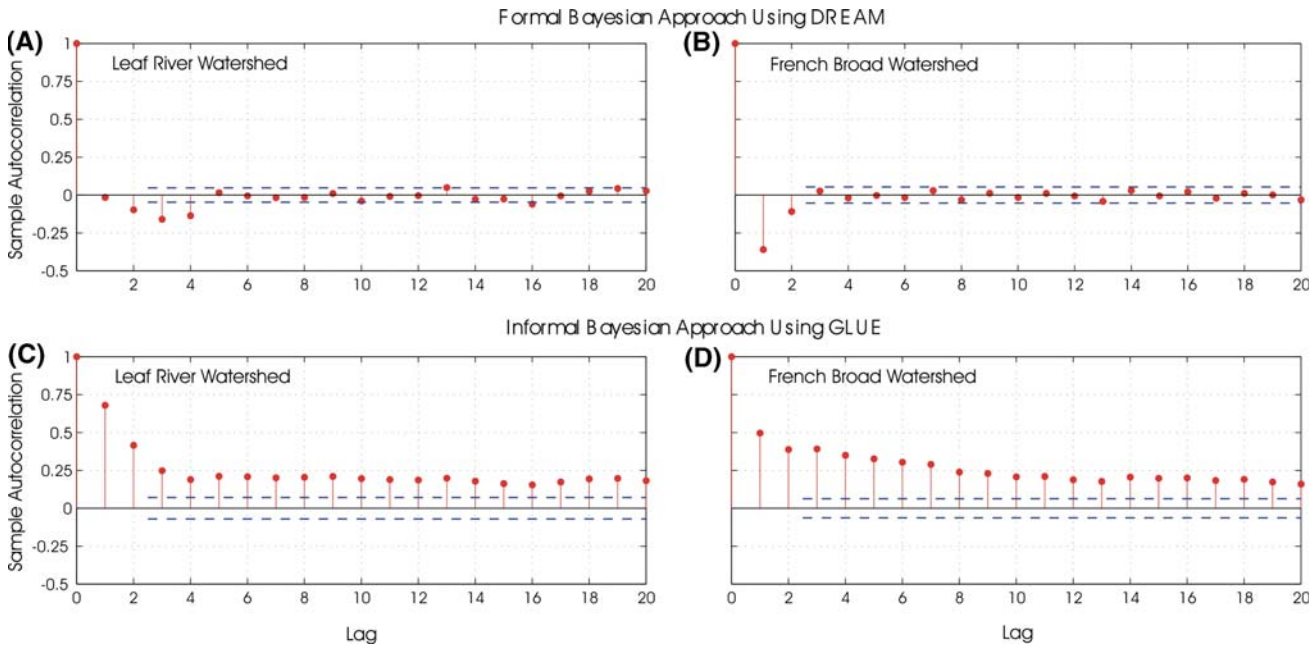

Fig. 8 Autocorrelation functions of the residuals of the mean ensemble streamflow forecasts and the verifying streamflow observations for the Leaf River (*left column*) and French Broad (*right column*) watersheds during the evaluation period. The *top panels* show the results using a formal Bayesian inference with DREAM, whereas the bottom results correspond to GLUE. The *dotted lines* in each of the individual panels denote the 95% confidence intervals for a series of uncorrelated and normally distributed residuals

Table 3 Summary statistics of the streamflow forecasts for the Leaf River and French Broad watersheds using formal (DREAM with MCMC simulation) and informal (GLUE) Bayesian inference

	Leaf River watershed						French Broad watershed					
	Calibration			Evaluation			Calibration			Evaluation		
	WY (1954–1958)			WY (1959–1963)			WY (1953–1957)			WY (1958–1962)		
	RMSE	CORR	BIAS	RMSE	CORR	BIAS	RMSE	CORR	BIAS	RMSE	CORR	BIAS
DREAM	13.46	0.95	−1.72	28.38	0.93	1.36	6.72	0.94	−1.33	7.83	0.93	0.09
GLUE	22.05	0.88	−4.21	34.82	0.91	−4.88	7.37	0.92	0.52	7.86	0.93	5.69

Units of RMSE, CORR, and BIAS are m^3/s, −, and %, respectively

RMSE root mean square error, *CORR* correlation coefficient, *BIAS* bias

watersheds. The statistics correspond to the mean ensemble forecast and distinguish between the calibration and evaluation period. The results in this Table show that the formal Bayesian approach consistently receives the best performance. This is not very surprising for the calibration period because rainfall and model structural inadequacies are explicitly inferred with the storm multipliers and first-order autoregressive (AR-1) scheme. This allows the HYMOD model to more closely track the streamflow observations. Yet, the difference in performance between the formal and informal Bayesian approach is generally smaller for the evaluation period. This is because our knowledge of precipitation multipliers for future events is at best incomplete. The difference in performance between DREAM and GLUE is most significant for the Leaf River, whereas a minor difference in RMSE, CORR and BIAS is found for the French Broad watershed. The rainfall record for the French Broad is structurally more consistent with the observed streamflow data, and cannot be improved much with storm multipliers. Indeed, their values reside in the vicinity of 1.

We like to emphasize that the findings presented in this paper are insensitive to the choice and length of the calibration time series. In all our calculations presented in this paper, we use $T = 1$ in Eq. 10. This is a standard setting that is most often used in GLUE applications. Larger values of T will increase the peakiness of the informal likelihood function in Eq. 10 and therefore reduce the parameter and output prediction (streamflow) uncertainty.

4 Summary and conclusions

In recent years, a strong debate has emerged in the hydrologic literature whether an uncertainty framework should have its roots within a proper statistical (Bayesian) context, or whether such a framework should implement informal measures and procedures to extract the information from the calibration data and summarize parameter and predictive distributions. The goal of this paper was to establish some common ground between these two different approaches, and compare GLUE with the more formal DREAM algorithm. This method implements Bayesian statistics, and uses state-of-the-art MCMC simulation to approximate the posterior probability distribution of the model parameter and output predictions. Our results demonstrate that:

- Formal Bayesian approaches that make very strong assumptions about the nature of the statistical properties of the residuals can generate very similar estimates of total predictive uncertainty as informal Bayesian approaches (such as GLUE used herein) that are based on a completely different philosophy of error representation. The debate that currently exits in the hydrologic literature between supporters of statistically coherent approaches for uncertainty estimation and champions of less formal approaches therefore might need serious reconsideration.
- The Bayesian method considered in this paper, has a somewhat smaller spread of the streamflow prediction uncertainty bounds than GLUE and better coverage of the streamflow observations.
- The GLUE procedure can reveal when no model can reproduce the observations given the available input data without compensation by a statistical error model or input adjustments. This is an important part of the learning process in hydrological modeling since it requires that model structure, input data or observations be questioned.
- Parameter uncertainty is made especially large in most GLUE applications because it includes implicit representation of model error. One should therefore be particularly careful in drawing conclusions about equifinality.
- The inability of GLUE to separate between individual error sources impairs our ability to identify structural deficiencies in models.

- Formal Bayesian approaches attempt to disentangle the effect of input, output, parameter and model structural error, which is key to improving our hydrologic theory of how water flows through watersheds. Note, however that formal Bayes law suffers from interaction between these individual error sources, which makes statistical inference difficult, and therefore results should be carefully interpreted.
- Low precipitation amounts are generally associated with relatively high uncertainty, whereas higher rainfall events are well defined with relatively small variation among the multipliers. This finding, made possible through analysis of streamflow data with a formal Bayesian approach, is consistent with papers in the literature that have analyzed the spatial variability and measurement error of rain-gauge data.

The source codes of GLUE and DREAM are written in MATLAB and can be obtained from the first author (vrugt@lanl.gov) upon request.

Acknowledgments The first author is supported by a J. Robert Oppenheimer Fellowship from the LANL postdoctoral program. We would like to thank Sander Huisman, Jan Mertens, Benedikt Scharnagl and Jan Vanderborght for stimulating discussions. The authors gratefully acknowledge the many comments and suggestions of Alberto Montanari, Keith Beven and an anonymous reviewer that have greatly enhanced the quality of this manuscript.

Appendix

Calculation of predictive uncertainty from MCMC simulation

Assume that for each MCMC draw $\mathbf{x}^j$ the distribution of each model outcome y_i, $i = 1,...,n$ is F and that $F(c) = Pr(y < c|\mathbf{x}^j)$ can be calculated exactly for any value of c. For example, from the MCMC runs using Eq. 9 and AR-1 normally distributed model and measurement error as in Eq. 6, F is a Student distribution $t_v(\mu, \sigma^2)$ with $v = n$, $\mu = f(\mathbf{x}^j, \hat{\phi})$ and $\sigma^2 = s^2/(1-\rho^2)$ with s^2 in Eq. 14. Now $Pr(y < c)$ can be estimated from the J MCMC draws using the average of $Pr(y < c|\mathbf{x}^j)$. To estimate a $100\alpha\%$ percentile we thus need to find c such that:

$$\frac{1}{J}\sum_{j=1}^{J} Pr(y<c|\mathbf{x}^j) = \alpha \tag{16}$$

This can be done numerically by a root-finding algorithm. A 95% confidence interval is constructed by calculating the 2.5 and 97.5% percentile, respectively.

References

Aronica G, Bates PD, Horritt MS (2002) Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. Hydrol Proc 16:2001–2016

Bates BC, Campbell EP (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 KJ (1989) Changing ideas in hydrology. The case of physically based models. J Hydrol 105:157–172

Beven KJ (1993) Prophecy, reality and uncertainty in distributed hydrological modeling. Adv Water Res 16(1):41–51

Beven K (2006) A manifesto for the equifinality thesis. J Hydrol 320:18–36. doi:10.1016/j.jhydrol.2005.07.007

Beven KJ, Binley AM (1992) The future of distributed models: model calibration and uncertainty prediction. Hydrol Proc 6:279–298

Beven K, Smith PJ, Freer JE (2008) So why would a modeller choose to be incoherent? J Hydrol 354:15–32. doi:10.1016/j.jhydrol.2008.02.007

Blasone RS, Vrugt JA, Madsen H, Rosbjerg D, Zyvoloski GA, Robinson BA (2008) Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. Adv Water Res 31:630–648. doi:10.1016/j.advwatres.2007.12.003

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc Ser B 26:211–246

Boyle DP (2000) Multicriteria calibration of hydrologic models. Ph.D. dissertation, Department of Hydrology and Water Resources, University of Arizona, Tucson

Brazier RE, Beven KJ, Anthony SG, Rowan JS (2001) Implications of model uncertainty for the mapping of hillslope-scale soil erosion predictions. Earth Surf Proc Land 26:1333–1352

Christensen S (2004) A synthetic groundwater modeling study of the accuracy of GLUE uncertainty intervals. Nordic Hydrol 35:45–59

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

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

Feyen L, Beven KJ, De Smedt F, Freer JE (2001) Stochastic capture zone delineation within the generalized likelihood uncertainty estimation methodology: conditioning on head observations. Water Resour Res 37(3):625–638

Franks SW, Beven KJ, Quinn PF, Wright IR (1997) On the sensitivity of soil-vegetation-atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration. Agric For Met 86:63–75

Freer JE, Beven K, Ambroise B (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, Rubin DB (1992) Inference from iterative simulation using multiple sequences. Stat Sci 7:457–472

Hankin BG, Hardy R, Kettle H, Beven KJ (2001) Using CFD in a GLUE framework to model the flow and dispersion characteristics of a natural fluvial dead zone. Earth Surf Proc Land 26:667–687

Hansson K, Lundin C (2006) Equifinality and sensitivity in freezing and thawing simulations of laboratory and in situ data. Cold Reg Sci Tech 44:20–37

Hornberger GM, Spear RC (1981) An approach to the preliminary analysis of environmental systems. J Env Manag 12:7–18

Iorgulescu I, Beven K, Musy A (2005) Data-based modelling of runoff and chemical tracer concentrations in the Haute-Mentue research catchment (Switzerland). Hydrol Proc 19:2557–2573. doi:10.1002/hyp.5731

Jensen JB (2003) Parameter and uncertainty estimation in groundwater modelling. PhD thesis, Department of Civil Engineering, Aalborg University, Series Paper No. 23

Kavetski D, Kuczera G, Franks SW (2006a) Bayesian analysis of input uncertainty in hydrological modeling: theory. Water Resour Res 42:W03407. doi:10.1029/2005WR004368

Kavetski D, Kuczera G, Franks SW (2006b) Bayesian analysis of input uncertainty in hydrological modeling: application. Water Resour Res 42:W03408. doi:10.1029/2005WR004376

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

Lamb R, Beven K, Myrabø S (1998) Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. Adv Water Res 22(4):305–317

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

Mantovan P, Todini E (2006) Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. J Hydrol 330:368–381. doi:10.1016/j.hydrol.2006.04.046

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

McMichael CE, Hope AS, Loaiciga HA (2006) Distributed hydrological modeling in California semi-arid shrublands: MIKE SHE model calibration and uncertainty estimation. J Hydrol 317:307–324

Mertens J, Madsen H, Feyen L, Jacques D, Feyen J (2004) Including prior information in the estimation of effective soil parameters in unsaturated zone modelling. J Hydrol 294(4):251–269

Misirli F, Gupta HV, Sorooshian S, Thiemann M (2003) Bayesian recursive estimation of parameter and output uncertainty for watershed models. In: Duan et al (eds) Calibration of watershed models, Water Sci. Appl. Ser., vol 6. AGU, Washington, pp 113–124

Montanari A (2005) Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. Water Resour Res 41:W08406. doi:10.1029/2004WR003826

Montanari A (2007) What do we mean by uncertainty? The need for a consistent wording about uncertainty assessment in hydrology. Hydrol Proc 21(6):841–845. doi:10.1002/hyp.6623

Muleta MK, Nicklow JW (2005) Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. J Hydrol 306:127–145

Romanowicz RJ, Beven KJ, Tawn J (1996) Bayesian calibration of flood inundation models. In: Anderson MG, Walling DE (eds) Floodplain processes. Wiley, Chichester, pp 333–360

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

Tadesse A, Anagnostou EN (2005) A statistical approach to ground radar-rainfall estimation. J Atm Ocean Tech 22(11):1055–1071

Villarini G, Krajewski WF (2008) Empirically-based modeling of spatial sampling uncetainties associated with rainfall measurements by rain gauges. Adv Water Resour 31(7):1015–1023. doi:10.1016/j.advwatres.2008.04.007

Vogel RM, Stedinger JR, Batchelder R, Lee SU (2008) Appraisal of the Generalized Likelihood Uncertainty Estimation (GLUE) method. Water Resour Res (in review)

Vrugt JA, Gupta HV, Bouten W, Sorooshian S (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 JA, Gupta HV, Sorooshian S, Wagener T, Bouten W (2006) Application of stochastic parameter optimization to the Sacramento Soil Moisture Accounting model. J Hydrol 325(1–4):288–307. doi:10.1016/j.hydrol.2005.10.041

Vrugt JA, ter Braak CJF, Diks CGH, Robinson BA, Hyman JM, Higdon D (2008a) Accelerating Markov chain Monte Carlo simulation by self-adaptive differential evolution with randomized subspace sampling. Water Resour Res (in review)

Vrugt JA, ter Braak CJF, Clark MP, Hyman JM, Robinson BA (2008b) Treatment of input uncertainty in hydrologic modeling: doing hydrology nackwards with Markov chain Monte Carlo simulation. Water Resour Res (in press)

Wagener T, Boyle DP, Lees MJ, Wheater HS, Gupta HV, Sorooshian S (2001) A framework for development and application of hydrologic models. Hydrol Earth Syst Sci 5(1):13–26

Wang X, He X, Williams JR, Izaurralde RC, Atwood JD (2005) Sensitivity and uncertainty analyses of crop yields and soil organic carbon simulated with EPIC. Trans Am Soc Agr Eng 48(3):1041–1054