The final publication is available at Springer via http://dx.doi.org/10.1007/s00477-014-0855-x

CITATION:
Shafii, M., B. Tolson, and L. Matott (2014), Uncertainty-based multi-criteria calibration of rainfall-runoff models: a comparative study, Stoch Environ Res Risk Assess, 28(6), 1493-1510.

Uncertainty-based multi-criteria calibration of rainfall-runoff models: A comparative study

Mahyar Shafii[1,*], Bryan Tolson[2], and L. Shawn Matott[3]

[1,*] Corresponding author, Dept. of Civil and Environ. Eng, University of Waterloo, Waterloo, ON, Canada; email: mshafiih@uwaterloo.ca, Tel: 1-519-888-4567, Ext. 38864, FAX: 1-519-888-4349

[2] Dept. of Civil and Environ. Eng, University of Waterloo, Waterloo, ON, Canada

[3] Center for Computational Research, University at Buffalo, Buffalo, NY, USA

Abstract

This study compares formal Bayesian inference to the informal Generalized Likelihood Uncertainty Estimation (GLUE) approach for uncertainty-based calibration of rainfall-runoff models in a multi-criteria context. Bayesian inference is accomplished through Markov Chain Monte Carlo (MCMC) sampling based on an auto-regressive multi-criteria likelihood formulation. Non-converged MCMC sampling is also considered as an alternative method. These methods are compared along multiple comparative measures calculated over the calibration and validation periods of two case studies. Results demonstrate that there can be considerable differences in hydrograph prediction intervals generated by formal and informal strategies for uncertainty-based multi-criteria calibration. Also, the formal approach generates definitely preferable validation period results compared to GLUE (i.e., tighter prediction intervals that show higher reliability) considering identical computational budgets. Moreover, non-converged MCMC (based on the standard Gelman-Rubin metric) performance is reasonably consistent with those given by a formal and fully-converged Bayesian approach even though fully-converged results requires significantly larger number of samples (model evaluations) for the two case studies. Therefore,

research to define an alternative and more practical convergence criteria for MCMC applications to computationally intensive hydrologic models may be warranted.

Keywords: Hydrologic modelling, Multi-criteria calibration, Uncertainty analysis, Bayesian inference, GLUE

1 Introduction

Hydrologic modelling has benefited from significant developments over the past two decades, and this has led to increasing complexity in hydrologic models and an advance from lumped conceptual models toward semi-distributed and distributed physics-based models. These models include many parameters which need to be estimated through an adjustment procedure using historical observation data. The automatic calibration conducted without sufficient hydrological expertise might yield improper parameter values which can result in unreasonable regimes of model responses that are not controlled by measurements (Refsgaard, 1997; Wagener et al., 2001). Moreover, even 'well calibrated' parameter values can yield poor performance with respect to an independent validation data set.

Problems with parameter adjustment in hydrologic models can be attributed to different factors. Conceptually, aggregation of all residuals into a single objective function during calibration does not provide sufficient detail about model inadequacy (Gupta et al., 1998). For example, single-objective metrics do not distinguish between high-flow and low-flow model behaviour. This realization has motivated multi-criteria calibration approaches in which multiple sets of observations and/or multiple evaluation criteria are employed (Gupta et al., 1998; Legates and McCabe, 1999; Madsen, 2000; Yapo et al., 1998). Multi-criteria calibration uses more than one index to describe the characteristics of the error vector (e.g., separate Nash-Sutcliffe values for high-flow and low-flow data), resulting in an objective-function tradeoff curve and corresponding set of “Pareto” optimal parameter values.

Another strategy for increasing the usefulness of predictive hydrologic models is to rigorously account for different sources of uncertainty (e.g., uncertainties associated with estimated parameter values as well as uncertainties in meteorological inputs and other non-calibrated forcing functions). In fact, it is very important to include an assessment of uncertainty in the calibration process. Razavi et al. (2010) named such approaches ‘uncertainty-based calibration’ which refers to the coupling of an environmental model with an uncertainty engine such that the uncertainty engine repeatedly samples model parameter configurations to develop a calibrated probability distribution for the parameters. Other research has emphasized comprehensive model assessment (or model evaluation) procedures whereby parameter estimation is done probabilistically to derive the probability density function (PDF) of the model outcome(s) of interest, through traditional ‘frequentist’ approaches (e.g. Bates and Watts, 1988; Reichert, 1997; Seber and Wild, 1989) and Bayesian inference approaches.

From a Bayesian perspective, uncertainty-based calibration seeks to elucidate posterior PDFs for various parameters and model outcomes given some prior information and available data. These posterior PDFs then form the basis of a complementary predictive uncertainty analysis (Bates and Campbell, 2001; Box and Tiao, 1973; Gelman et al., 2004; Kavetski et al., 2002; Kuczera, 1983; Kuczera and Parent, 1998; Thiemann et al., 2001). The Bayesian approach to model specification and uncertainty analysis is particularly appealing as it allows for formal specification and propagation of an error model (Marshall et al., 2007). Furthermore, in the Bayesian approach, any a priori knowledge about model parameters can be used in terms of prior distributions, which are then updated for any particular catchment using the data available. For complex hydrologic models, Bayesian inference is aided by the use of numerical procedures that implement Markov Chain Monte Carlo (MCMC) sampling. In this regard, a number of MCMC samplers have been proposed, including BaRE (Bayesian Recursive Estimation) (Thiemann et al., 2001); SCEM-UA (Shuffled Complex Evolution Metropolis – University of Arizona) (Vrugt et al., 2003b), BATEA (Bayesian Total Error Analysis) (Kavetski et al., 2002; Kavetski et al., 2006), and DREAM (Differential Evolution Adaptive Metropolis) (Vrugt et al., 2009).

At the heart of Bayesian inference is the use of formal likelihood functions to analyse parameter uncertainty. A given likelihood function must make explicit assumptions about the form of the model residuals (i.e., deviations between simulations and observations) (Stedinger et al., 2008). Thus, a major criticism of the Bayesian approach is that in hydrologic modelling the appropriate statistical form for a given set of model residuals is not always clear, and this makes it difficult to establish an appropriate likelihood function (e.g., Beven et al., 2008). To address this issue, some researchers have emphasized the development of more appropriate likelihood functions by using hierarchical Bayesian structures that disaggregate different sources of uncertainties (e.g., Huard, 2008; Kuczera et al., 2006; Moradkhani et al., 2005; Renard et al., 2010; 2011; Wei et al., 2010). However, development and application of such formulations to complex non-linear hydrological models is non-trivial and may be computationally intractable in some case studies using existing state-of-the-art MCMC samplers. The issue of defining an appropriate Bayesian likelihood formulation becomes even more challenging when one considers a multi-response or multi-criteria approach – an approach that some have argued is the most appropriate for hydrological modelling (e.g., Hamilton, 2007; Montanari, 2007).

Recently, the concept of epistemic and aleatory uncertainties in hydrological modelling has been discussed among researchers (Beven et al., 2012; Beven et al., 2011; Clark et al., 2012;

Montanari, 2011). Uncertainties are categorized as aleatory (also called natural uncertainty) if they are presumed to be the intrinsic randomness of a stochastic process which can be represented in terms of the probabilities of different outcomes. On the other hand, many of the errors that enter into the modelling process stem from a lack of knowledge about processes and boundary conditions. These errors are called epistemic or limited-knowledge uncertainty. In statistical models (including Bayesian inference structures), uncertainties are accounted for by providing a representation of all of the important sources of uncertainty as aleatory (Beven et al., 2011). As a consequence, the results of Bayesian methods might not be robust when many of the errors that affect modelling uncertainty in hydrology are epistemic (Beven et al., 2011; Beven et al., 2008). However, statistical methods are believed to be able to fit epistemic uncertainties provided that the inherent regularities are well represented by the statistical model itself (Montanari, 2011). Similar to almost all studies in the literature on uncertainty analysis of rainfall-runoff models, the Bayesian method of our paper also considers all uncertainties to be aleatory.

Despite the robust theoretical underpinnings of a formal Bayesian approach to parameter inference, a variety of alternative and informal approaches have been proposed for uncertainty-based multi-criteria calibration of complex hydrological models. Examples include a Pareto-based calibration approach (Gupta et al., 1998) and informal MCMC sampling (Blasone et al., 2008a; Vrugt et al., 2003a). Importance sampling techniques have also been used for informal uncertainty-based calibration, with GLUE (Generalized Likelihood Uncertainty Estimation) (Beven and Binley, 1992) being the most commonly used approach. GLUE is based on the concept of 'equifinality' and classifies Monte Carlo samples as having produced model output that is either 'behavioural' (i.e., plausible, given the data and one's knowledge of the system) or ‘non-behavioural’. The behavioural solutions are then used to derive the probability distribution function for parameters and model outputs. The GLUE methodology can be easily extended to multi-criteria calibration problems (e.g. Blazkova and Beven, 2009). A drawback of informal methods is that such approaches do not require formal specification of an error model and might not be reliable for uncertainty analysis (Kavetski et al., 2002).

Along with development of a variety of uncertainty-based calibration routines, some researchers have focused their efforts towards comparison between formal and informal methods. Overall, these efforts generally indicate relatively close agreement among alternative methods, in terms of predictive capability (Beven et al., 2008; Jeremiah et al., 2011; Jin et al., 2010; Li et al., 2010; Qian et al., 2003; Vrugt et al., 2008; Yang et al., 2008). Note that some studies have only considered informal methods in their comparisons (e.g., Blasone et al., 2008b).

From both a comparative and theoretical perspective, previous literature demonstrates that MCMC sampling and Bayesian inference can be considered a preferred approach to deal with uncertainty-based calibration, as long as the computational budget allows full convergence of the MCMC sampler. Achieving convergence is not problematic if one is dealing with rainfall-runoff models with manageable simulation runtimes. However, when computational budget limitations exist, MCMC sampling may not be an appropriate choice. Furthermore, the observed similarity between the predictive capabilities of formal and informal approaches suggests that one might be able to gain insight into predictive uncertainty by means of informal approaches without getting involved in likelihood definition and corresponding assumptions. Most of previous papers comparing formal and informal approaches have only considered single-criterion calibration scenarios. Balin-Talamba (2004) and Balin-Talamba et al. (2010) considered multi-criteria calibration of hydrologic models applying GLUE and MCMC sampling. These studies evaluated the impact of multi-response calibration on predictive uncertainty using GLUE and MCMC, in comparison with single-criterion calibration. However, the GLUE and MCMC techniques are only visually compared in Balin-Talamba (2004) and no comparative measures are reported. To the best of our knowledge, comparison among formal and informal techniques from a multi-criteria perspective using quantitative comparative measures has yet to be reported on in the literature.

The main objective of this research is to evaluate the applicability of different uncertainty analysis approaches to multi-criteria calibration and uncertainty analysis of hydrologic models considering identical computational budget. The methodologies addressed in this paper are statistically-based Bayesian inference using MCMC sampling (Bates and Campbell, 2001; Kuczera, 1983; Schaefli et al., 2007; Vrugt et al., 2009), and sampling-based uncertainty estimation using GLUE (Beven and Binley, 1992; Blazkova and Beven, 2009). Bayesian inference was implemented using the DREAM MCMC sampler (Vrugt et al., 2009) through a robust multi-criteria formulation. Also, we consider an alternative Bayesian method based on the results of MCMC sampling up to a limited computational budget (i.e., using the MCMC before convergence). Such a method cannot be viewed informal, as it uses formal likelihood function; however, it would not be formal either, as convergence has not occurred, meaning that the solutions in the chain could not be considered as samples from posterior distributions.

2 Methodology

A typical multi-criteria model calibration process can involve multiple likelihood functions used for different sets of measurements, e.g., discharge, sediment, snow, etc. However, even in the case of a model with only one output flux to be simulated, the model evaluation may still be considered to be inherently multi-criteria (Gupta et al., 1998). The multi-criteria numerical experiments in this study only deal with one response (discharge), splitting it into high- and low-flows. This strategy is expected to be adequate for an initial exploration of multiple uncertainty-based calibration techniques within a multi-criteria formulation.

The comparison framework of this study uses the posterior distribution of model parameters derived from MCMC sampling, as well as the behavioural or optimal parameter sets obtained from other methods. In order to be consistent in wording, the term "posterior" is applied to all of the considered techniques even though the results of non-converged MCMC sampling and GLUE are not a formal statistical posterior distribution. Results are then compared with respect to computational burden, complexity, and predictive capacity. Numerical experiments are aimed at exploring advantages and disadvantages of the uncertainty analysis techniques addressed in this study in multi-criteria calibration of rainfall-runoff models. The reliability of these methods is evaluated using two rainfall-runoff models, a 5-parameter lumped model, HYMOD (Hydrology model) (Boyle, 2000), and an 11-parameter semi-distributed model, WetSpa (Water and Energy Transfer between Soil, Plants and Atmosphere) (Liu et al., 2003; Wang et al., 1996).

The GLUE approach of this paper employs informal likelihood functions and results are compared with those obtained from formal Bayesian inference as well as non-converged MCMC sampling. The use of GLUE without a formal likelihood function has been the subject of much debate (e.g., Beven et al., 2008; Mantovan and Todini, 2006; Montanari, 2005; Thiemann et al., 2001). Nevertheless, we used GLUE with an informal generalized likelihood function in this study because the objective of the study was to assess the performance of informal methods. Much of the reason informal methods like GLUE are so well utilized in practice is because they can use informal likelihood functions based on long utilized deterministic calibration objective functions like sum of squared errors or the Nash Sutcliffe coefficient. It is also worth noting that GLUE could also be applied using formal likelihood functions (Freni and Mannina, 2009; Romanowicz et al., 1994), but this is not addressed in the present paper.

The comparison approach (informal to formal methods) of this study is exactly consistent with previous comparative studies of uncertainty-based calibration in hydrological modelling (e.g. Vrugt et al., 2008; Yang et al., 2008). Beven (2009) noticed that in Vrugt et al. (2008) the formal Bayes estimates are based on an autoregressive error model, while such information is not supplied to the GLUE simulations. Despite the difference between the formulations of the Bayesian approach and GLUE in Vrugt et al. (2008), it is shown in that paper that formal and informal uncertainty analysis methods have some common ground with respect to the total predictive uncertainty in single-criterion calibration cases. In this paper, multiple quantitative comparative measures are applied and we evaluate the similarity in behavior of MCMC and GLUE in the multi-criteria context. As such, we consider the same implementations of MCMC sampling and GLUE as used in Vrugt et al. (2008).

2.1 Formal multi-criteria Bayesian inference

Bayesian statistics have been shown to be a robust methodology for formal multi-criteria calibration and uncertainty analysis of hydrologic models, as long as all underlying assumptions are satisfied. Both analytical and numerical Bayesian approaches have been used to deal with multi-criteria calibration (Balin-Talamba et al., 2010; Hong et al., 2005; Kuczera, 1983; Kuczera and Mroczkowski, 1998; Mroczkowski et al., 1997; Schaefli et al., 2007). The notion of multi-criteria in Bayesian inference structures is mostly concerning cases in which multiple responses of observations are employed (e.g., measured streamflows and measured soil water content), and thus, it is also called multi-response calibration in the literature. There are also reports of multi-criteria Bayesian formulations using a single response. For instance, Schaefli et al. (2007) considered multiple likelihood functions which were associated with high- and low-streamflows. The research presented here used a previously published multi-criteria formulation (Balin-Talamba et al., 2010; Schaefli et al., 2007).

Moreover, the initial experiments of Bayesian inference in these case studies showed that errors were correlated. As a result, we had to consider development of a formal likelihood function which accounts for auto-correlation. As such, auto-regressive (AR) parameters were introduced to the high- and low-flow time series to address auto-correlation among residuals (e.g., Bates and Campbell, 2001; Kuczera, 1983). The resulting Bayesian inference formulation introduces a first-order AR scheme to represent the residuals (Balin-Talamba et al., 2010; Schaefli et al., 2007), details of which are provided in the Appendix of this paper. Note that the AR scheme was applied separately to the low- and high-flow regimes and this resulted in the addition of two AR parameters (ρ_L for low-flows and ρ_H for high-flows) to the set of calibrated parameters.

For this paper, the DREAM MCMC sampler was used for formal Bayesian inference (Vrugt et al., 2009). DREAM maintains ergodicity while showing excellent efficiency even if the target posterior distributions are complex, highly nonlinear, and/or multimodal. DREAM runs multiple Markov chains simultaneously to facilitate efficient global exploration of the parameter space. Like other adaptive samplers, DREAM speeds convergence by dynamically adjusting the scale and orientation of the proposal distribution.

2.2 Sampling-based Uncertainty Estimation using Non-converged MCMC

Even though applications of MCMC sampling with pseudo-likelihood functions have been previously reported in the literature (Blasone et al., 2008b; Vrugt et al., 2003a), there has been no report on evaluation of the results from non-converged MCMC samplers with formal likelihood functions. In this paper, non-converged DREAM results are used to approximate the converged MCMC sampling strategy. The number of solutions taken from a given DREAM chain was defined to be consistent with the informal methods considered in this paper (explained below). For example, if the informal methods use a budget of 10000 simulations, then we only consider 10000 solutions from the initial part of the long DREAM chain. Afterwards, the last 1000 solutions of this set would be treated as posterior solutions to derive prediction intervals. Clearly, such an approach is neither formal (as convergence has not occurred) nor informal (as it uses formal likelihood function). That is the reason why we separated this approach from formal Bayesian and informal GLUE approaches.

2.3 Sampling-based Uncertainty Estimation using GLUE

The GLUE technique (Beven and Binley, 1992) is the most commonly applied method in the family of informal sampling-based methods. In GLUE, parameter uncertainty accounts for all sources of uncertainty, because ‘‘the likelihood measure value is associated with a parameter set and reflects all these sources of error and any effects of the covariation of parameter values on model performance implicitly’’ (Beven and Freer, 2001). The GLUE analysis conducted here consisted of the following four steps:

1. Defining the generalized informal likelihood measure $l(\boldsymbol{\theta})$. Generally, the measure $l(\boldsymbol{\theta})$ is a pseudo-likelihood function which demonstrates the model performance for a particular

parameter set $\boldsymbol{\theta}$. In this study, we used the generalized likelihood function provided in previous multi-criteria GLUE studies (Balin-Talamba, 2004; Lamb et al., 1998) as follows:

$$l(\theta) = \prod_{i=1}^{M} \exp\left(-W_i \frac{\sigma_{\varepsilon,i}^2}{\sigma_{o,i}^2}\right) \tag{1}$$

where W_i represents the weighting factor for criterion i (explained later), M is the number of criteria, $\sigma_{\varepsilon,i}^2$ and $\sigma_{o,i}^2$ are the variance of simulation errors and the variance of observed data, respectively, over the time window in which criterion i is calculated. The likelihood function $l(\theta)$ equals 1 if the observed and simulated data are the same for all criteria, and reduces towards zero as the similarity decreases. Note that, in the multi-criteria calibration problem of this paper, we calculate this likelihood function based on the information in high- and low-flow time periods (M=2).

2. After defining $l(\boldsymbol{\theta})$, a large number of parameter sets are randomly sampled from the prior distribution and each parameter set is assessed as either ‘‘behavioural’’ or ‘‘non-behavioural’’ through a comparison of the likelihood measure with a selected threshold value which is explained in details later in this section of the paper.

3. Each behavioural parameter set is given a likelihood weight according to $\varpi_i = l(\boldsymbol{\theta}_i) \Big/ \sum_{k=1}^{N} l(\boldsymbol{\theta}_k)$, where N is the number of behavioural parameter sets.

4. Finally, prediction uncertainty of streamflow is described by quantiles of the cumulative distribution realized from the weighted behavioural parameter sets, i.e., at each time step, the model outcome associated to behavioural solutions are identified and prediction intervals (for example 95% intervals) are constructed based on quantiles (such as 2.5 and 97.5 percentiles).

The behavioural threshold for the GLUE pseudo-likelihood function defines the boundary between behavioural and non-behavioural solutions. In this study, based on the strategy in Balin-Talamba (2004) and Lamb et al. (1998), we followed the same strategy (also described below) to filter out behavioural samples. Once samples are taken from prior distributions, the generalized likelihood function Eq. (1) is calculated considering high- and low-flow time periods whereby the weights are equal for both periods, i.e., $W_L = 0.5$ and $W_H = 0.5$ (note that L and H stand for low- and high-flows, respectively):

$$l(\theta) = \exp\left(-W_L \frac{\sigma^2_{\varepsilon,L}}{\sigma^2_{o,L}}\right).\exp\left(-W_H \frac{\sigma^2_{\varepsilon,H}}{\sigma^2_{o,H}}\right) = \exp\left[-\left(W_L \frac{\sigma^2_{\varepsilon,L}}{\sigma^2_{o,L}} + W_H \frac{\sigma^2_{\varepsilon,H}}{\sigma^2_{o,H}}\right)\right] \qquad (2)$$

Parameter sets are now sorted based on the combined criterion, and the top N samples are considered behavioural solutions. Identifying N is in fact a subjective decision in GLUE, and would probably affect the uncertainty bounds computed using the GLUE method. Among the traditional choices reported in literature is N being equal to the number of top 10% of solutions (Binley and Beven, 1991; Lamb et al., 1998) sampled from the prior distributions. However, Lamb et al. (1998) showed that relaxation of the rejection threshold to define a larger proportion of the total number of samples as behavioural would cause only slight modifications of uncertainty bounds. The reason for this insensitivity to the rejection threshold is that even after selecting a larger number of behavioural samples, the majority of samples would achieve only small likelihood values. Therefore, the predictions associated with these poor samples would fall within the tails of the cumulative distributions of model outcome. Given the rescaling stage in GLUE, these predictions would have little effect on the location of uncertainty bounds (Lamb et al., 1998). In this paper, we also considered the top 10% strategy to define behavioural samples.

2.4 Comparison measures

The main goal of calibration and uncertainty analysis is to assess models' predictive capability. Therefore, in order to evaluate uncertainty-based calibration techniques, it seems necessary that we focus more on the validation time period rather than the calibration period. Nonetheless, a portion of our analysis examined differences between calibration and validation results. The comparative measures are calculated based on the results obtained using the posterior parameter sets. It should be noted that the parameter uncertainty is derived based on the envelope of model outputs using the posterior parameter sets. Moreover, in order to derive the predictive uncertainty, the entire set of posterior parameters is first used in simulation model to derive the parameter uncertainty. Afterwards, error parameters are sampled to generate a correlated residual time series which is then added to model outputs.

To evaluate the quality of resulting model outcomes, efficiency measures such as NS is used to assess model performance. In the multi-criteria context of this paper, we illustrate the scatterplot of posterior parameter sets in bi-criteria space (i.e., NS for high- and low-flows).

In addition, the generated model outcomes using the posterior solutions derived from different techniques are used to derive the predictive uncertainty which can be assessed using a variety of measures. Among the simplest measures for comparing alternative realizations of predictive uncertainty are the reliability and sharpness measures (Yadav et al., 2007). For a given prediction interval, the reliability measure is the percentage of discharge observations that are captured by the prediction interval. Reliability values are calculated by counting the number of times the observed streamflow falls within the prediction band, divided by the length of the time series. Sharpness is a measure of the prediction intervals' width relative to the hydrograph prediction bounds obtained from sampling prior feasible parameter ranges. If the posterior prediction bounds for the hydrograph form a single line, sharpness would be 100%. Whereas when the posterior prediction bounds are the same as those obtained using priori feasible parameter ranges, sharpness would be 0% (clearly undesirable). Ideally, and for a given prediction interval, the reliability should be equal to the desired interval percentage (i.e., 90% of observations should be captured by a 90% prediction interval) and larger values of the corresponding sharpness measure are better than smaller values.

The Bayesian posterior predictive p-value is another measure of the predictive capacity of uncertainty-based calibration techniques (Gelman et al., 2004, pp. 162–163). The Bayesian p-value is the probability that the model prediction at a particular time step could be more extreme than the observed data at that same time step. Such values may be estimated by the proportion of simulations for which the simulated value equals or exceeds the observed value. Probability distributions of p-values can be constructed from the complete series of p-value calculations. If the model output and measured data are consistent, the corresponding p-value distribution should be uniformly distributed over the interval [0,1]. This can be checked graphically using QQ-plots (Laio and Tamea, 2007; Thyer et al., 2009) and deviations from the bisector (the 1:1 line) denote interpretable deficiencies (see Figure 1).

[Figure 1 goes here.]

Our approach to compute comparative performance metrics with GLUE such as reliability, sharpness and Bayesian p-values is consistent with studies computing one or more of these metrics for GLUE results based on a pseudo-likelihood function such as Vrugt et al. (2008), Yang et al. (2008) and Jin et al. (2010)

2.5 Case Studies

Bayesian inference is expected to result in robust expression of predictive uncertainty, as long as all assumptions are satisfied and the posterior PDFs are taken from a converged MCMC sampler. Two case-studies involving real data from two catchments are used in this paper, for which the DREAM sampler is run to convergence to extract formal posterior distributions. The non-converged MCMC sampling and GLUE methods are also applied to the same problems. One case-study applies the HYMOD hydrologic model to the Leaf River catchment, and one applies the WetSpa hydrologic model to the Hornad River catchment, where details about these catchments are provided below.

The first study area addressed in this paper is the 1994 km^2 Leaf River watershed located north of Collins, Mississippi. This catchment has been studied intensively in the past (e.g., Boyle, 2000; Sorooshian et al., 1993; Thiemann et al., 2001; Vrugt et al., 2003b; Vrugt et al., 2008) and may be considered a standard benchmark for parameter estimation of hydrological models. In this regard, three years (i.e., 1953-1955) of hydrologic data (i.e., mean areal precipitation [mm/d], potential evapotranspiration [mm/d], and streamflow [m^3/s]) were used. The first two years of data were used for model calibration, while the third year served as a validation dataset for assessing predictive capability. We used the simulation model HYMOD in this catchment to predict streamflow at a single location in the Leaf River channel network. The HYMOD model is a relatively simple rainfall excess model (Moore, 1985) connected with a series of linear reservoirs. HYMOD requires estimation of five parameters and these are listed in Table 1 along with their prior range.

[Table 1 goes here.]

The second case study is the 1,131 km^2 Hornad River catchment located in Slovakia. The observations for this catchment were collected from 1991 to 2000, and the first five years (i.e., 1991 to 1995) were used for calibration and the remaining data (i.e., 1996 to 2000) was used for validation. We used the simulation model WetSpa in this catchment to predict streamflow at a single location in the Hornad River channel network. Unlike HYMOD, WetSpa is a grid-based hydrologic model that simulates water and energy transfer between soil, plants and the atmosphere. WetSpa can be configured to run in semi-distributed or fully distributed mode of which the former was chosen for this study. According to the previous applications of WetSpa model to Hornad catchment (Bahremand et al., 2007; Liu et al., 2003; Shafii and Smedt, 2009), and as shown in Table 2, 11 WetSpa parameters were targeted for calibration.

[Table 2 goes here.]

The multi-criteria formulation used in this paper was created by splitting a single time series of responses (i.e., discharges) into high- and low-flows. Following Schaefli et al. (2007), high-flows corresponded to time steps in which the hydrograph was rising, and low-flows were defined based on the recession part of hydrograph. Separate Nash-Sutcliffe values (or formal likelihood values, in the case of MCMC sampling) were then calculated for each flow regime, yielding a bi-criteria calibration problem.

The computational overhead required for GLUE and DREAM are both dominated by the simulation model run time and as such, for the same number of model simulations completed, GLUE and DREAM require approximately the same computation time. The simulation model run time for HYMOD and WetSpa are 0.65 and 2.25 seconds, respectively, on a PC with 3-GHz Intel processor.

3 Results

For each of the case studies, the DREAM sampler was first applied to establish a converged chain of samples, and the non-converged DREAM and GLUE were then applied. Note that, as mentioned earlier, we used an AR-based Bayesian formulation in this paper. Transformation and/or scaling of parameters is an important factor that can affect the difficulty of parameter estimation (Bates and Watts, 1981; Johnston and Pilgrim, 1976; Kuczera, 1983) and the convergence behaviour of MCMC samplers (Hills and Smith, 1992). For the HYMOD Leaf River and WetSpa Hornad River case studies, a series of preliminary numerical experiments were performed to explore alternative parameter transformations within the DREAM sampler. These experiments indicated that the most suitable transformation was to logarithmically transform HYMOD and WetSpa model parameters and use un-transformed auto-regressive parameters. It should also be noted that, in the formal Bayesian approach, discharges were also transformed logarithmically to stabilize the error variance.

3.1 HYMOD

When applied to the HYMOD Leaf River case study, the DREAM sampler converged after approximately 143000 simulations. The convergence of MCMC sampler was checked using the Gelman-Rubin convergence metric, which was also cross-checked to verify residuals normality (via inspection of a QQ-plot) and non-correlation (via inspection of the auto-correlation function). Furthermore, 1000 out of the last 10000 post-convergence samples were taken from the DREAM chain and used to derive baseline posterior parameter distributions. For the non-converged DREAM approach, a new trial of DREAM was considered up to 10000 simulations of which the last 1000 samples were used to derive corresponding posterior distributions. The

GLUE method was applied using the generalized likelihood function Eqs (1-2) considering two scenarios, (i) a budget of 10000 simulations called 'GLUE Low-budget', and (ii) identical computation budget to DREAM (i.e., 143000 simulations in HYMOD case study) and is called 'GLUE Full-budget'.

Figure 2 illustrates the posterior parameter information derived by the various calibration methods when applied to the HYMOD Leaf River case study. As observed in Figure 2 the posterior parameter ranges varied across methods, especially with respect to parameters `Rs` and `Rq`. Most of the ranges given by non-converged DREAM were wider than those given by converged DREAM. The difference between the location of posterior solutions derived from Bayesian inference and GLUE is not surprising, and can be explained by the fact that different likelihood functions have been used in these methods. However, comparison between these posterior ranges indicates that incorporating two additional error parameters (i.e., higher complexity in comparison to informal formulation) resulted in a higher level of identifiability, especially for parameters `Rs` and `Rq`.

[Figure 2 goes here.]

Figure 3 illustrates the Nash-Sutcliffe (NS) values of the HYMOD Leaf River case study for calibration (upper panel) and validation (lower panel) period, demonstrating the results of DREAM (light points) versus non-converged DREAM and GLUE (dark points) along low and full computational budget. Conversion of DREAM likelihood values into equivalent NS values was non-trivial because the fitted error series should also be accounted for. Proper conversion into equivalent NS values must consider additional elements of the revised Bayesian formulation, namely, the two extra auto-regressive parameters (i.e., ρ_L and ρ_H) and the AR-based residuals term (δ_t). Thus, for a given parameter vector $\boldsymbol{\varphi}_i$ containing a model parameter set $\boldsymbol{\theta}_i$ and corresponding $\rho_{L,i}$ and $\rho_{H,i}$ auto-regressive parameters, the corresponding error variances were sampled to generate 100 different time series of error realizations. These errors were then combined with simulated discharges and auto-regressive terms to yield 100 different NS values for parameter vector $\boldsymbol{\varphi}_i$. The average of these NS values was then used as the equivalent NS value converted from the original DREAM likelihood value. Repeating this process for all parameter vectors contained in the DREAM posterior samples yielded the equivalent NS values plotted in Figure 3 for calibration and validation period.

[Figure 3 goes here.]

As shown in the calibration part in Figure 3 (upper panel), the results obtained from DREAM were superior (based on NS values) to those given by other methods, and there was some overlap between the posterior sets of solutions given by converged and non-converged DREAM sampler. Note that we sometimes call these sets of solutions 'posterior clouds', as they look like a cloud in NS space. In the validation part of Figure 3 (lower panel), the non-converged DREAM posterior cloud very closely resembles the DREAM posterior cloud. This is a good indication that much of the high-density areas of the parameter space were explored prior to the DREAM sampler satisfying the Gelman-Rubin convergence criteria.

The results of GLUE in Figure 3 also indicate that regardless of the computation budget considered, the samples were located in fairly identical space in NS space (but with different densities) both in calibration and validation period. However, GLUE with full computational budget performed slightly better considering extreme NS values of GLUE in Figure 3. The GLUE results in calibration period showed that 8% of behavioural samples resulted in negative NS values for low-flows, but since their NS values for high-flows were high, they could rank in the top 10% of all GLUE samples. It should be pointed out that, similar to previous studies (Balin-Talamba, 2004; Lamb et al., 1998), the threshold for classifying solutions as behavioural utilized the formulations in Eqs. (1-2), and did not take into consideration the condition of positive NS values. This explains why there are some solutions with negative low-flows NS values among posterior samples.

Figure 3 also shows that GLUE yielded good performance in terms of matching the simulations with observation in validation low-flows, but not as good in high-flows compared to DREAM sampler. In contrast, the posterior cloud generated by DREAM in validation period (Figure 3 lower panel) emphasized matching high-flows (i.e., points clustered in the 0.8 to 1.0 range for NS_{high}) at the expense of matching low-flows (i.e., points clustered around $NS_{low} = 0.5$).

Ideally, all posterior samples would generate positive NS values in validation period for low- and high-flows. The vertical dashed lines in Figure 3 (lower panel) separates the region with positive NS values for low-flows, and thus, the ideal region would be the right half of the scatter plots. It is observed that all posterior samples from DREAM and all but one of the non-converged DREAM posterior samples were located in this ideal region. However, almost 40% of posterior GLUE (full-budget) samples generated negative validation period NS values for 'low-flows'. It should be pointed out that almost 92% of these samples had resulted in positive NS values both for low- and high-flows in calibration period.

Figure 4 (left panels) illustrates the tradeoff between reliability and sharpness measures for the HYMOD Leaf River case study (only in validation period) for the various methods that were considered (i.e., DREAM, non-converged DREAM, and GLUE with low and full computational budget). The reliability and sharpness values were calculated based on 95% prediction intervals on the corresponding posterior PDFs of simulated discharges. The reliability was calculated based the percentage of coverage of observations by prediction bounds, whereas sharpness was based on the amount of reduction in discharge ranges through comparison with the range of model simulations using prior parameter ranges. In order to define such prior intervals, 100000 Latin hypercube samples were taken from prior parameter ranges which were used in HYMOD to generate 100000 discharge hydrographs. The minimum and maximum of discharges at each time steps were then identified to serve as prior discharge ranges.

[Figure 4 goes here.]

The HYMOD results in Figure 4 (left panel) show that the converged DREAM sampler and 'GLUE Full-budget' cannot dominate each other with respect to both reliability and sharpness. Compared to 'GLUE Full-budget', the converged DREAM resulted in improved sharpness both for low- and high-flows. In terms of reliability, as the goal was to generate 95% prediction intervals, both methods came fairly close to this goal given that reliabilities in validation period ranged from 93% to 97%. Comparison between non-converged DREAM and 'GLUE Low-budget' shows that neither of these two methods is superior to the other one with respect to both reliability and sharpness. The reliabilities of these two methods were close to 95%. The sharpness of non-converged DREAM was larger than 'GLUE Low-budget' in low-flows, and approximately the same in high-flows.

Figure 5 contains Bayesian p-values for both the calibration and validation periods of the HYMOD Leaf River case study for non-converged and converged DREAM approaches. Note that the p-values were derived using the entire set of posterior solutions. Figure 5 shows that even though the p-value results for the converged and non-converged DREAM sampler were different during the calibration period, the results in validation period, however, were fairly similar. Also, both methods yielded underestimation of predictive uncertainty with respect to low-flows in validation period. This might be due to the fact that we used standard Bayesian formulation without disaggregation of different sources of uncertainty, which will be discussed later in the discussion section.

[Figure 5 goes here.]

Figure 6 illustrates the prediction bounds given by the posterior simulations of the considered calibration techniques for the validation period in HYMOD case study. The bounds shown in Figure 6 are derived in a manner similar to those given for posterior parameters of Figure 2 and are assumed to represent 95% prediction intervals. As shown in Figure 6, the converged DREAM sampler reliably covers the validation dataset. Prediction bounds of the non-converged DREAM sampler resemble those generated from the converged DREAM sampler but at the cost of larger width and larger peak flow values. Figure 6 also shows that the prediction bounds associated with 'GLUE Full-budget' are larger than those derived with 'GLUE Low-budget', but covered the observations better.

[Figure 6 goes here.]

Across the various comparative measures that were evaluated in the context of the HYMOD Leaf River case study, we observed that the formal Bayesian method (both converged and non-converged MCMC sampling) turned out to be more appropriate than informal GLUE strategy in calibration period. Once the validation period was used to evaluate the methods, the formal Bayesian inference (given the formulation of this paper) resulted in a level of underestimation of predictive uncertainty, which would be probably solved through more complex HBS systems, as elaborated in discussions section. On the other hand, the GLUE methodology was only successful in partially meeting the predictive criteria in validation period. The WetSpa Hornad River real case study (Section 3.2) investigates whether these findings would hold for a more complex hydrological model (involving more uncertain parameters) applied to a different catchment.

3.2 WetSpa

For the WetSpa case study (i.e., application to Hornad River catchment), the DREAM sampler was again configured to use a formal auto-regressive Bayesian inference formulation and the method converged (based on the Gelman-Rubin statistic) after 470,000 simulations. As with the HYMOD studies, 10000 post-convergence DREAM samples were taken to construct the Bayesian posterior distributions. Similar to the previous case, the results of non-converged DREAM were derived based on running DREAM only up to 10000 simulations (independent trial than converged DREAM). GLUE was also applied to the WetSpa case study using low and full computational budget as described in HYMOD Leaf River case study.

Figure 7 contains normalized posterior ranges of the WetSpa model parameters generated by the various calibration methods. The first result noted in Figure 7 is that some parameters were deemed non-identifiable (i.e., K_S, K_{GI}, and K_{RD}) by the converged DREAM sampler, as indicated by 95% posterior intervals covering almost the entire prior range. When informal likelihood functions were used (i.e., GLUE), most of parameters appeared to be poorly-identifiable.

However, it should be noted that the difference between the location of posterior parameter ranges and identifiability levels obtained by formal and informal methods would be explained by the difference in the likelihood functions used in these methods. It is also observed in Figure 7 that the posterior parameter ranges derived from non-converged DREAM covered those obtained from converged DREAM, and this shows how the sampler located a smaller posterior region after it converged.

[Figure 7 goes here.]

Figure 8 illustrates the Nash-Sutcliffe values for calibration (upper panel) and validation (lower panel) period of the WetSpa Hornad River case study as evaluated by non-converged DREAM and GLUE (dark points), in comparison to those calculated based on the posterior solutions of the converged DREAM sampler (light points). Note that two cases were reported for GLUE, one with low and one with full computational budget. Also note that the axes in lower panel of Figure 8 were centred between ±1, the dashed lines showing the origin where both NS values were zero. A number of GLUE solutions were not within this range and were not depicted in Figure 8. The ideal region for a given calibration method to sample from would be the upper right quadrant of validation panel where both low- and high-flow NS values were positive. It is observed in the calibration panel that DREAM yielded the best NS values both for low- and high-flows. Given that non-converged DREAM and DREAM achieve these high NS values, it seems the inclusion of an error term is important to achieve such high performance. The posterior cloud from non-converged DREAM overlaps substantially the converged DREAM posterior cloud, which indicates that the posterior distribution has likely been sampled from well before the Gelman-Rubin statistic indicated convergence. The results of GLUE (low and full computational budgets) also indicate that increasing the number of simulations in GLUE did not result in comparable model performance as DREAM (see distance between the location of posterior clouds). It is also observed in GLUE results (both low and full budgets) that there were a considerable number of points not located in the ideal region, that is, positive NS values for low and high-flows or the upper right quadrant identified by dashed lines, even though they were all behavioural in the calibration period.

[Figure 8 goes here.]

The sharpness and reliability measures for the validation period of the WetSpa Hornad River case study are given in Figure 4 (right panel). These measures were computed in the same manner as those for the HYMOD Leaf River case study. In terms of reliability, as the goal was to generate 95% prediction intervals, all methods came fairly close to this goal for high flows given that reliabilities in validation period ranged from 94% to 98%. The same is true for validation period low flows except that 'GLUE Low-budget' results have a slightly lower reliability of 88%. Comparing converged DREAM with 'GLUE Full-budget', it is observed that DREAM results dominate GLUE in both low-flows and high flows (i.e., larger reliability and larger sharpness). In other words, DREAM generates tighter 95% prediction intervals and simultaneously improves reliability. Similarly, non-converged DREAM dominates 'GLUE Low-budget' results in high flows and practically dominates 'GLUE Low-budget' results in low flows (very similar reliabilities but significantly improved sharpness for DREAM).

Figure 9 compares the Bayesian p-value QQ plots for non-converged and converged DREAM sampling for the calibration (upper panel) and validation periods (lower panel) of the WetSpa Hornad River case study. As implied by the sigmoid shapes of their respective p-value curves, both DREAM samplers (i.e., converged and non-converged) exhibited systematic under-estimation of uncertainty for low-flows in validation period, even though the results of converged DREAM in calibration period were promising both for low-flows and high-flows. This finding is similar to results in Thyer et al. (2009) and the previous HYMOD case study. The under-estimation of only low-flow uncertainty by the converged DREAM procedure can be considered as indication of model structural error. This suggests that improving the low-flow modules in WetSpa may be a worthwhile enterprise. Such insight highlights the usefulness of multi-criteria Bayesian p-value separation as a post-diagnostic measure for detecting model structural deficiencies. However, it is also possible that the above-mentioned issue may be due to mis-specification of likelihood function.

[Figure 9 goes here.]

Figure 10 illustrates the prediction bounds given by the posterior simulations of the considered calibration techniques for one year (i.e., 1999) of the 5-year validation period (whereas Figure 4 reliability and sharpness values summarize prediction bounds over the entire 5-year period). The bounds shown in Figure 10 were derived in a manner similar to those given for posterior parameters of Figure 7 and are assumed to represent 95% prediction intervals. As shown in Figure 10, the converged DREAM sampler reliably covered the validation dataset even though the Bayesian p-value analysis indicated that the results were not perfect with respect to low-flows. Prediction bounds of the non-converged DREAM sampler resemble those generated from the converged DREAM sampler but at the cost of larger width and larger peak flow values. The prediction bounds associated with ‘GLUE Full-budget’ are larger than those derived with ‘GLUE Low-budget’, but covered the observations better.

[Figure 10 goes here.]

Across all comparative measures, the results of the WetSpa case study suggest the following conclusions: (1) the formal Bayesian inference through the standard formulation of this paper using converged DREAM yielded good results with respect to almost all predictive measures, except for p-values of low-flows in validation period; (2) the non-converged DREAM sampler yielded results that were nearly universally consistent with the converged DREAM sampler while requiring a fraction (i.e., 2%) of the computational budget; and (3) considering the predictive measures addressed in this study, GLUE did not meet all measures as satisfactorily as formal DREAM methodology, even when the full computational budget was considered.

4 Discussion

The DREAM results suggest that the Gelman-Rubin convergence criterion is too stringent since non-converged DREAM results closely approximates converged DREAM results and yet requires a fraction of the computational budget. It may also be possible to further improve the results of the non-converged DREAM sampler (i.e., make it more closely approximate the converged DREAM results) by filtering out obviously low quality solutions for the calibration period (e.g., those with NS values smaller than 0.5 in upper left panels of Figures 3 and 8). Also, one might think of applying alternative convergence measures. A potential hydrology-based convergence metric can be the reproduction of hydrological signatures that represent the overall hydrologic behaviour of the catchment (Gupta et al., 2008; Yilmaz et al., 2008). Future research should explore these and other alternative convergence measures in a multi-criteria context.

Comparison between formal and informal methods could also be viewed from the standpoint of aleatory and epistemic uncertainties, which was also elaborated in the introduction section of this paper. The errors in the case studies of this paper are assumed to be aleatory (especially in Bayesian inference methodology), even though in reality they could be a mixture of both aleatory and epistemic uncertainties. The results reveal that validation period performance measures are generally poorer compared to calibration period which is expected to be caused by epistemic errors (Beven et al., 2011). Thus, in the presence of epistemic errors, neither the standard Bayesian formulation nor the informal methods (such as GLUE) would be perfectly reliable in prediction mode. There are improved informal and formal approaches for case studies where epistemic errors are thought to be significant, e.g., the use of hierarchical Bayesian structures (e.g., Huard, 2008; Kuczera et al., 2006; Moradkhani et al., 2005; Renard et al., 2010; Wei et al., 2010), or the concept of 'limits of acceptability' used for identifying behavioural models in GLUE (Blazkova and Beven, 2009; Liu et al., 2009). Comparison between these two more advanced formal and informal uncertainty analysis methods is an interesting future research avenue.

5 Concluding Remarks

This paper evaluates the applicability of formal (Bayesian inference) and informal (GLUE) multi-criteria methods to uncertainty-based calibration in hydrological modelling. Bayesian inference is implemented through DREAM sampling based on a multi-criteria formulation. The results of non-converged DREAM are also evaluated. The results are compared with those obtained from two scenarios for GLUE, using a restricted computational budget and the full computational budget equivalent to the budget required for DREAM sampler to converge. The various methods are applied to two cases involving the 5-parameter HYMOD model and the 11-parameter WetSpa model. Results demonstrate that there can be considerable differences in prediction intervals generated by formal and informal strategies for uncertainty-based multi-criteria calibration. Future uncertainty-based calibration studies for simulation models with a large number of parameters should be aware of the potential considerable difference between the results of formal and informal strategies.

Results also demonstrate that it is advisable to consider multiple comparative measures, including traditional metrics like the Nash-Sutcliffe efficiency, when comparing alternative calibration strategies. Furthermore, it is observed that the choice of using the validation period or the calibration period for selected comparative measures would influence the analysis and as such it is recommended that future uncertainty-based calibration method comparison studies should include and largely focus on comparative performance assessment for the validation period.

In general, the Bayesian inference methodology performs well (in comparison with other methods) along all comparative measures except for low-flows in validation period considering the same computational budget, e.g., DREAM validation period prediction intervals are simultaneously tighter and more reliable than corresponding GLUE intervals. In case of limited computational budget (i.e., only 10000 simulations in this paper), non-converged MCMC sampling using DREAM proves to be fairly consistent with formal Bayesian inference. This indicates the potential value of utilizing formal MCMC sampling results before convergence as a promising alternative to informal methods such as GLUE.

The results obtained through application of Bayesian inference to the two cases of this paper indicated under-estimation of predictive uncertainty for low-flows in the validation period. We applied a standard Bayesian formulation which lumps all uncertainties into a single additive error term. More recently, Renard et al. (2010; 2011) showed that consideration of rainfall and model structural uncertainties outside of the error term used in Bayesian formulation yielded more reliable estimation of the predictive uncertainty for all runoff ranges, as opposed to the typical Bayesian formulation in our paper. Application of hierarchical Bayesian structures to the case studies of this paper is currently being investigated.

There are many ways to formulate and conduct GLUE analyses, and to some extent DREAM calibration experiments. Our experiments require a number of subjective decisions and as such our results are conditional on these decisions. However, we believe that the subjective decisions we make are consistent with the decisions others have made in the literature. For example, although it is possible to apply GLUE using a formal likelihood function, the literature suggests that is relatively uncommon and thus we do not examine this. We used GLUE with an informal generalized likelihood function in this study because the objective of the study was to assess its performance as an informal method. It may be possible that applying informal methods such as GLUE using formal likelihood functions would improve their performance, but this is not the focus of the present study. Future comparative studies systematically varying such subjective decisions would be valuable.

6 Acknowledgement

The authors would like to acknowledge Dr. Jasper A. Vrugt for providing the code of his DREAM algorithm used in this study and the four anonymous reviewers for their helpful comments that have improved the paper.

References

Bahremand, A. et al., 2007. WetSpa Model Application for Assessing Reforestation Impacts on Floods in Margecany–Hornad Watershed, Slovakia. Water Resources Management, 21(8): 1373-1391.

Balin-Talamba, D., 2004. Hydrological behaviour through experimental and modeling approaches; Application to the Haute-Mentue catchment, PhD Thesis, Swiss Federal School of Technology of Lausanne.

Balin-Talamba, D., Parent, E., Musy, A., 2010. Bayesian multiresponse calibration of TOPMODEL: Application to the Haute-Mentue catchment, Switzerland. Water Resources Research, 46: W08524.

Bates, B.C., Campbell, E.P., 2001. A Markov Chain Monte Carlo Scheme for parameter estimation and inference in conceptual rainfall-runoff modeling. Water Resources Research, 37(4): 937-947.

Bates, D.M., Watts, D.G., 1981. Parameter transformations for improved approximate confidence regions in nonlinear least squares. The Annuals of Statistics, 9(6): 1152-1167.

Bates, D.M., Watts, D.G., 1988. Nonlinear regression analysis and its applications. Wiley, NY.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrological Processes, 6(3): 279-298.

Beven, K., Freer, J., 2001. Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. J. Hydrology, 249((1-4)): 11-29.

Beven, K., Smith, P.J., Westerberg, I., Freer, J., 2012. Comment on "Pursuing the method of multiple working hypotheses for hydrological modeling" by P. Clark et al. Water Resources Research, 48: W11801.

Beven, K., Smith, P.J., Wood, A., 2011. On the colour and spin of epistemic error (and what we might do about it). Hydrol. Earth Syst. Sci., 15: 3123-3133.

Beven, K.J., 2009. Comment on ''Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling; by Jasper A. Vrugt, Cajo J. F. ter Braak, Hoshin V. Gupta and Bruce A. Robinson. Stoch Environ Res Risk Assess, 23: 1059-1060.

Beven, K.J., Smith, P.J., Freer, J., 2008. So just why would a modeller choose to be incoherent? . J. Hydrology, 354: 15-32.

Binley, A., Beven, K., 1991. Physically-based modelling of catchment hydrology: a likelihood approach to reducing predictive uncertainty. In: Rycroft, D.G.F.a.M.J. (Ed.), Computer Modelling in the Environ-mental Sciences, The Institute of Mathematics and its Applications Conference Series. Clarendon Press, Oxford, pp. 75–88.

Blasone, R.-S., Madsen, H., Rosbjerg, D., 2008a. Uncertainty assessment of integrated distributed hydrological models using GLUE with Markov chain Monte Carlo sampling. J. Hydrology, 353: 18– 32.

Blasone, R.S. et al., 2008b. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. Adv. Water Res., 31: 630-648.

Blazkova, S., Beven, K., 2009. A limits of acceptability approach to model evaluation and uncertainty estimation in flood frequency estimation by continuous simulation: Skalka catchment, Czech Republic. Water Resources Research, 45: W00B16.

Box, G.E.P., Tiao, G.C., 1973. Bayesian Inference in Statistical Analysis. Addison-Wesley, Boston, MA.

Boyle, D.P., 2000. Multicriteria calibration of hydrological models. Ph.D. Dissertation Thesis, Univ. of Ariz.

Clark, M.P., Kavetski, D., Fenicia, F., 2012. Reply to comment by K. Beven et al. on “Pursuing the method of multiple working hypotheses for hydrological modeling”. Water Resources Research, 48: W11802.

Freni, G., Mannina, G., 2009. Bayesian approach for uncertainty quantification in water quality modelling:The influence of prior distribution. J. Hydrology, 392: 31–39.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B., 2004. Bayesian Data Analysis. Chapman & Hall/CRC, Boca Raton, Florida.

Gupta, H.V., Sorooshian, S., Yapo, P.O., 1998. Toward improved calibration of hydrologic models: Multiple and noncommensurable measures of information. Water Resources Research, 34(4): 751-764.

Gupta, H.V., Wagener, T., Liu, Y., 2008. Reconciling theory with observations: Elements of a diagnostic approach to model evaluation. Hydrological Processes.

Hamilton, S., 2007. Just say NO to equifinality. Hydrological Processes, 21(14): 1979-1980.

Hills, S.E., Smith, A.F.M., 1992. Parameterization issues in Bayesian inference. In: J. M. Bernardo, J.B., A. P. Dawid and A. F. M. Smith (Ed.), Bayesian Statistics 4. Oxford University Press, pp. 227-46.

Hong, B., Strawderman, R.L., Swaney, D.P., Weinstein, D.A., 2005. Bayesian estimation of input parameters of a nitrogen cycle model applied to a forested reference watershed, Hubbard Brook Watershed Six. Water Resources Research, 41.

Huard, D., and A. Mailhot, 2008. Calibration of hydrological model GR2M using Bayesian uncertainty analysis. Water Resources Research, 44: W02424.

Jeremiah, E., Sisson, S., Marshall, L., Mehrotra, R., Sharma, A., 2011. Bayesian calibration and uncertainty analysis of hydrological models: A comparison of adaptive Metropolis and sequential Monte Carlo samplers. Water Resources Management, 47(W07547).

Jin, X., Xu, C.-Y., Zhang, Q., Singh, V.P., 2010. Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model. Journal of Hydrology, 383: 147-155.

Johnston, P.R., Pilgrim, D.H., 1976. Parameter optimization for Watershed models. Water Resources Research, 12(3): 477-486.

Kavetski, D., Franks, S.W., Kuczera, G., 2002. Confronting Input Uncertainty in Environmental Modeling. In: Duan, Q., Gupta, H.V., Sorooshian, S., Rousseau, A.N., Turcotte, R. (Eds.), Calibration of Watershed Models. AGU, Washington, DC, pp. 49-68.

Kavetski, D., Kuczera, G., Franks, S.W., 2006. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resources Research, 42, W03407, doi: 10.1029/2005WR004368.

Kuczera, G., 1983. Improved parameter inference in catchment models: 1. Evaluating parameter uncertainty. Water Resources Research, 19(5): 1151-1162.

Kuczera, G., Kavetski, D., Franks, S., Thyer, M., 2006. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: Characterising model error using storm-dependent parameters. Journal of Hydrology, 331(3-4): 161–177.

Kuczera, G., Mroczkowski, M., 1998. Assessment of hydrologic parameter uncertainty and the worth of multiresponse data. Water Resources Research, 34(6): 1481–1489.

Kuczera, G., Parent, E., 1998. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. Journal of Hydrology, 211(1-4): 69-85.

Laio, F., Tamea, S., 2007. Verification tools for probabilistic forecasts of continuous hydrological variables. Hydrol. Earth Syst. Sci., 11(4): 1267- 1277.

Lamb, R., Beven, K.J., 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.

Legates, D.R., McCabe, G.J., 1999. Evaluating the use of ‘‘goodness-of-fit’’ measures in hydrologic and hydro-climatic model evaluation. Water Resources Research, 35: 233-241.

Li, L., Xia, J., Xu, C.-Y., Singh, V.P., 2010. Evaluation of the subjective factors of the GLUE method and comparison with the formal Bayesian method in uncertainty assessment of hydrological models. Journal of hydrology, 390: 210-221.

Liu, Y., Freer, J., Beven, K., Matgen, P., 2009. Towards a limits of acceptability approach to the calibration of hydrological models: Extending observation error. J. Hydrology, 367: 93-103.

Liu, Y.B., Gebremeskel, S., De Smedt, F., Hoffmann, L., Pfister, L., 2003. A diffusive transport approach for flow routing in GIS-based flood modeling. Journal of Hydrology, 283(1-4): 91-106.

Madsen, H., 2000. Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. J. Hydrology(235): 276– 288.

Mantovan, P., Todini, E., 2006. Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology. J. Hydrology, 130((1 -2)): 368- 381.

Marshall, L., Nott, D., Sharma, A., 2007. Towards dynamic catchment modelling: a Bayesian hierarchical mixtures of experts framework. Hydrological Processes, 21: 847-861.

Montanari, A., 2005. Large sample behaviors of the generalized likelihood uncertainty estimation (GLUE) in assessing the uncertainty of rainfall-runoff simulations. Water Resources Research, 41: W08406.

Montanari, A., 2007. What do we mean by "uncertainty"? The need for a consistent wording about uncertainty in hydrology. HPToday, 21(6): 841-845.

Montanari, A., 2011. Interactive comment on “On the colour and spin of epistemic error (and what we might do about it)” by K. Beven et al. Hydrol. Earth Syst. Sci. Discussion, 8: C2885–C2891.

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

Moradkhani, H., Sorooshian, S., Gupta, H.V., Houser, P.R., 2005. Dual state-parameter estimation of hydrological models using ensemble Kalman filter. Adv. Water Resour., 28: 135-147.

Mroczkowski, M., Raper, G.P., Kuczera, G., 1997. The quest for more powerful validation of conceptual catchment models. Water Resources Research, 33: 2325- 2335.

Qian, S.S., Stow, C.A., Borsuk, M.E., 2003. On Monte Carlo methods for Bayesian inference. Ecological Modelling, 159: 269-277.

Razavi, S. et al., 2010. Reducing the Computational Cost of Automatic Calibration through Model Pre-Emption. Water Resources Research, 46(11): W11523.

Refsgaard, J.C., 1997. Parameterisation, calibration and validation of distributed hydrological models. J. Hydrology, 198: 69-97.

Reichert, P., 1997. On the necessity of using imprecise probabilities for modelling environmental systems. Water Science and Technology, 36(5): 149-156.

Renard, B., Kavetski, D., Kuczera, G., Thyer, M., 2010. Understanding predictive uncertainty in hydrologic modeling:The challenge of identifying input and structural errors. 2010, 46: W05521.

Renard, B. et al., 2011. Toward a reliable decomposition of predictive uncertainty in hydrological modeling : Characterizing rainfall errors using conditional simulation. Water Resour. Res., 47: W11516.

Romanowicz, R.J., K.J. Beven, Tawn, J., 1994. Evaluation of predictive uncertainty in nonlinear hydrological models using a Bayesian approach. In: Turkman, V.B.a.K.F. (Ed.), Statistics for the Environment 2, Water Related Issues, pp. 297-315.

Schaefli, B., Talamba, D.B., Musy, A., 2007. Quantifying hydrological modeling errors through a mixture of normal distributions. J. Hydrology(332): 303– 315.

Seber, G.A., Wild, C.J., 1989. Nonlinear Regression. John Wiley and Sons, New York (NY).

Shafii, M., Smedt, F.D., 2009. Multi-objective calibration of a distributed hydrological model (WetSpa) using a genetic algorithm. Hydrology and Earth System Sciences, 13: 2137-2149.

Sorooshian, S., Duan, Q., Gupta, V.K., 1993. Calibration of rainfall-runoff models: Application of global optimization to the Sacramento Soil Moisture accounting model. Water Resources Research, 29: 1185– 1194.

Stedinger, J.R., Vogel, R.M., Lee, S.U., Batchelder, R., 2008. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resources Research, 44: W00B06.

Thiemann, M., Trosset, M., Gupta, H.V., Sorooshian, S., 2001. Bayesian recursive parameter estimation for hydrologic models. Water Resources Research, 37(10): 2521-2535.

Thyer, M. et al., 2009. Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: A case study using Bayesian total error analysis. Water Resources Research, 45: W00B14.

Vrugt, J.A., Gupta, H.V., Bastidas, L.A., Bouten, W., Sorooshian, S., 2003a. Effective and efficient algorithm for multiobjective optimization of hydrologic models. Water Resources Research, 39(8): 1214, doi:10.1029/2002WR001746.

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

Vrugt, J.A. et al., 2009. Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling. International Journal of Nonlinear Sciences and Numerical Simulation, 10(3): 273-290.

Vrugt, J.A., ter Braak, C.J.F., Gupta, H.V., Robinson, B.A., 2008. Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stoch Environ. Res. Risk Assess., 44: 1-16.

Wagener, T. et al., 2001. A framework for development and application of hydrological models. Hydrol. Earth Syst. Sci., 5(1): 13– 26.

Wang, Z.-M., Batelaan, O., De Smedt, F., 1996. A distributed model for water and energy transfer between soil, plants and atmosphere (WetSpa). Physics and Chemistry of The Earth, 21(3): 189-193.

Wei, W., Clark, J.S., Vose, J.M., 2010. Assimilating multi-source uncertainties of a parsimonious conceptual hydrological model using hierarchical Bayesian modeling. J. Hydrology, 394: 436–446.

Yadav, M., Wagener, T., Gupta, H., 2007. Regionalization of constraints on expected watershed response behavior for improved predictions in ungauged basins. Advances in Water Resources, 30 (8): 1756–1774.

Yang, J., Reichert, P., Abbaspour, K.C., Yang, H., Xia, J., 2008. Comparing Uncertainty Analysis Techniques for a SWAT Application to the Chaohe Basin in China. J. Hydrology, 358: 1-23.

Yapo, P.O., Gupta, H.V., Sorooshian, S., 1998. Multi-objective global optimization for hydrologic models. Journal of Hydrology, 204(1-4): 83-97.

Yilmaz, K.K., Gupta, H.V., Wagener, T., 2008. A process-based diagnostic approach to model evaluation: Application to the NWS distributed hydrologic model. Water Resources Research, 44: W09417.

APPENDIX – Review of Bayesian Inference Procedure

This appendix provides a summary of the Bayesian formulation used in this paper, and the details can be found in previous studies (Balin-Talamba et al., 2010; Schaefli et al., 2007). We assume the AR-based formulation as follows:

$$Y_i = (Y_i^{sim} | \boldsymbol{\theta}, \mathbf{X}) + \rho \varepsilon_{i-1} + \delta_i \tag{A1}$$

where Y_t and Y_i^{sim} are the observed and simulated values for the model response at time step i, $\boldsymbol{\theta}$ is the model parameters vector, $\mathbf{X}$ is the model inputs vector, ρ is the lag-one AR parameter, $\varepsilon_i = \left(Y_i - Y_i^{sim}(\boldsymbol{\theta}, \mathbf{X})\right)$ is the residual between observation and model prediction at time step i (and $\varepsilon_0 = 0$), and δ_i is random error term:

$$\delta_i \sim N(0, \sigma_j^2) \tag{A2}$$

with σ_j^2 being the residual variance for response j, here considered unknown and should be estimated. If we consider J responses, then J parameters (representing error variance for J responses) need to be estimated in the Bayesian inference methodology. Under the assumption of multiple and statistically independent responses, the combined statistical likelihood function for multiple responses is simply the product of the individual likelihood functions:

$$\begin{aligned} l_{multiple} &= \prod_{j=1}^{J} l_j(\boldsymbol{\theta}, \rho, \sigma_j^2, \mathbf{X}) \\ &= \prod_{j=1}^{J} \frac{1}{\left(\sqrt{2\pi}\right)^{t_j} . \sigma_j^{t_j}} . \exp\left(-\frac{\sum_{i=1}^{t_j} \delta_{j,i}{}^2}{2\sigma_j^2} \right) \end{aligned} \tag{A3}$$

where $\delta_{j,i} = \varepsilon_{j,i} - \rho\varepsilon_{j,i-1}$ for observation set j and time step i (note that $\varepsilon_{j,0} = 0$), respectively; J is the number of observation sets, and t_j is the number of time steps for each observation set j. In order to derive the posterior distribution of parameters, a bounded uniform prior distribution is considered for $\boldsymbol{\theta}$ over prior feasible range, and the prior distribution of error variance is also considered to be Jeffrey non-informative distribution as follows:

$$p(\sigma_j^2) \propto 1/\sigma_j^2 \text{ for } 0 < \sigma_j^2 < \infty \tag{A4}$$

Using such prior distributions enables us to integrate out the error variances, and the Bayesian formulation results in the joint posterior distributions from which the marginal distribution of model parameters and error variances can be estimated conditioned on the observed data **Y**. Alternatively, we can use MCMC sampling to directly take samples from the posterior distributions, all of which are contained in the chain. In MCMC implementations, the acceptance/rejection criterion ratio (between posterior densities of the new candidate and old current samples) is used to accept/reject the candidate to be added to the chain. In the multi-criteria Bayesian formulation, let $\sigma^2_{j,current}$ and $\sigma^2_{j,candidate}$ be the error variance of the current and candidate solutions, respectively, which are estimated based on the residuals after running the simulation model. Also assume the quantity $S_j = 0.5\sum_{i=1}^{t_j} \delta_{j,i}{}^2$, such that $S_{j,current}$ and $S_{j,candidate}$ be the values for the current and the candidate solutions, respectively. The final form of the acceptance/rejection criterion can then be shown as follows:

$$\alpha = \prod_{i=1}^{J} \exp\left[\left(\frac{1}{\sigma^2_{j,current}} + \frac{1}{\sigma^2_{j,candidate}}\right).\left(S_{j,current} - S_{j,candidate}\right)\right] \cdot \left(\frac{S_{j,candidate}}{S_{j,current}}\right)^{\frac{t_j}{2}} \qquad \text{(A5)}$$

Table 1. HYMOD parameters and their prior range

Parameter	Descritption	Unit	Prior Range
CMAX	Maximum storage capacity	mm	[1 , 500]
BEXP	Degree of the soil spatial variability moisture capacity	-	[0.1 , 2]
ALPHA	Distributing factor on flow between the two series of reservoirs	-	[0 , 0.1]
RQ	Residence time of the quick reservoirs	d	[0 , 0.1]
RS	Residence time of the slow reservoirs	d	[0.1 , 0.99]

Table 2. Parameters of WetSpa simulation model

Parameter	Description	Unit	Prior Range
Ki	Interflow scaling factor	-	[0 – 10]
Kg	Groundwater recession coefficient	d^{-1}	[0 - 0.05]
Ks	Initial soil moisture factor	-	[0 – 2]
Ke	Correction factor for PET	-	[0 – 2]
Kgi	Initial groundwater storage	mm	[0 – 500]
Kgm	Groundwater storage scaling factor	mm	[0 – 2000]
Kt	Base temperature for snowmelt	^{o}C	[-1 – 1]
Ktd	Temperature degree-day coefficient	mm $^{o}C^{-1}d^{-1}$	[0 – 10]
Krd	Rainfall degree-day coefficient	$^{o}C^{-1}$ d^{-1}	[0 - 0.05]
Km	Surface runoff coefficient	-	[0 – 5]
Kp	Rainfall scaling factor	mm	[0 – 500]

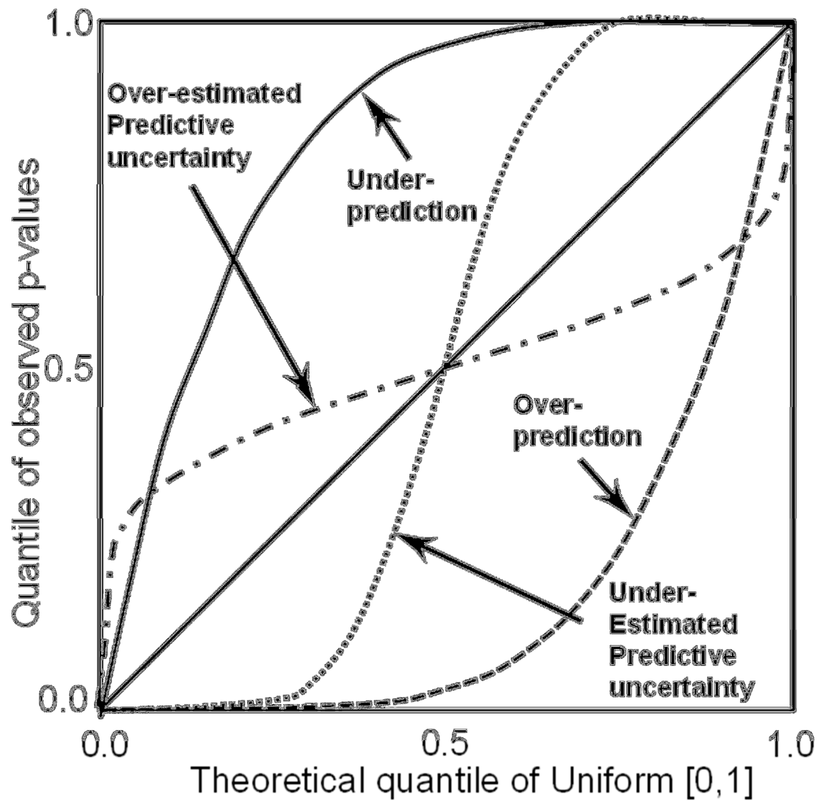

Figure 1. Schematic of the predictive QQ plot based on Thyer et al. (2009)

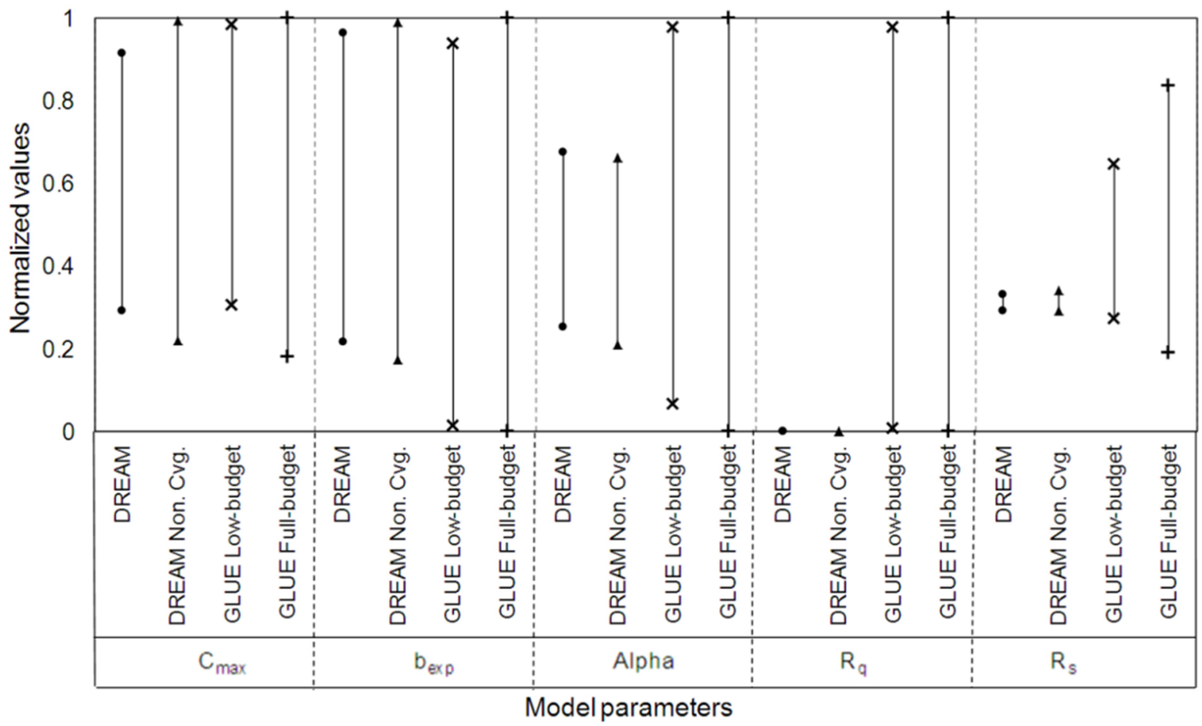

Figure 2. Posterior ranges of HYMOD parameters for the Leaf River case study; The parameter ranges correspond to 95% posterior intervals for different uncertainty analysis methods.

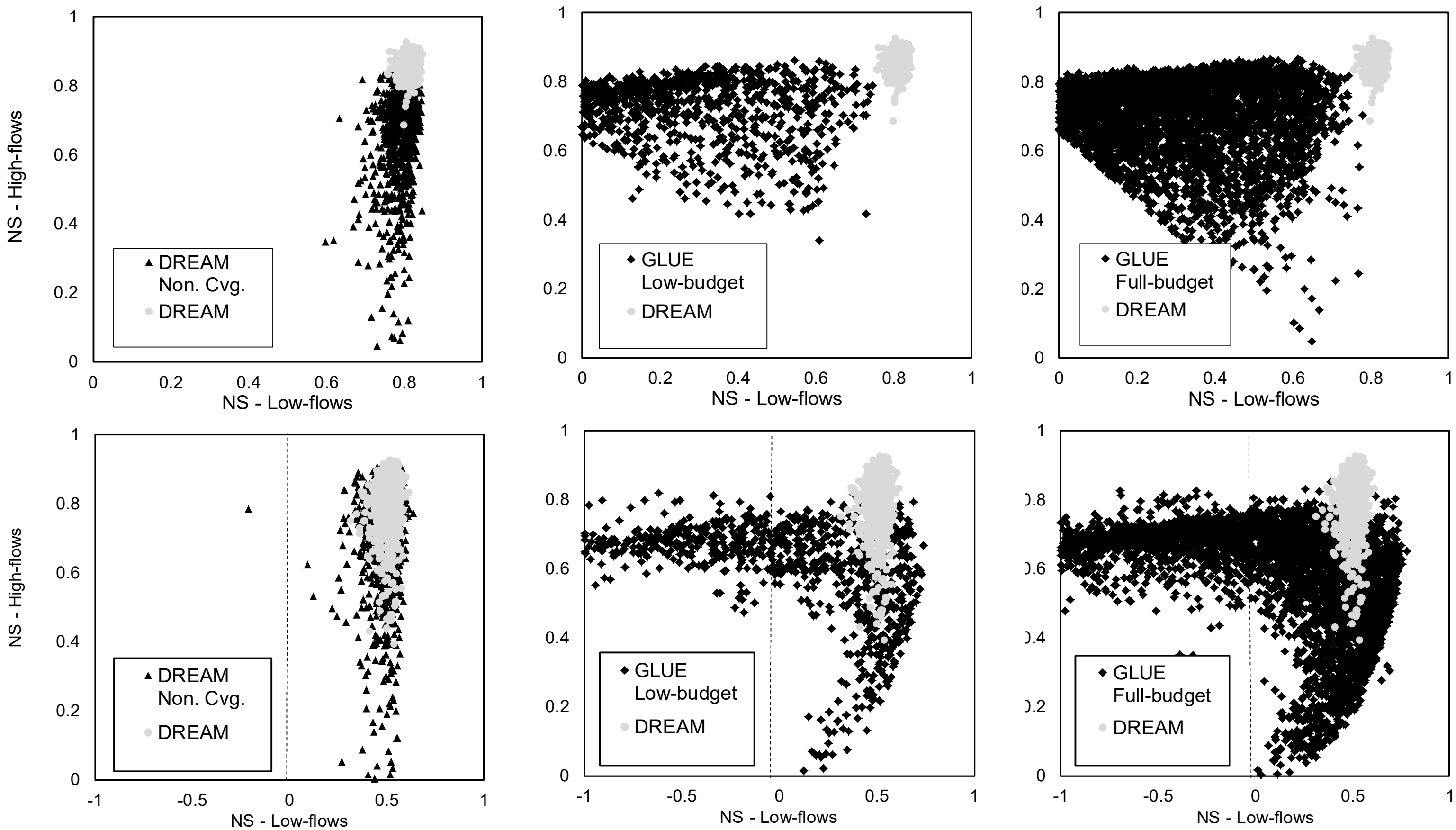

Figure 3. NS values of low-flows (horizontal axis) and high-flows (vertical axis) in calibration (upper panels) and validation (lower panels) period for HYMOD case study, derived from DREAM (light points) versus non-converged DREAM and GLUE methods (dark points).

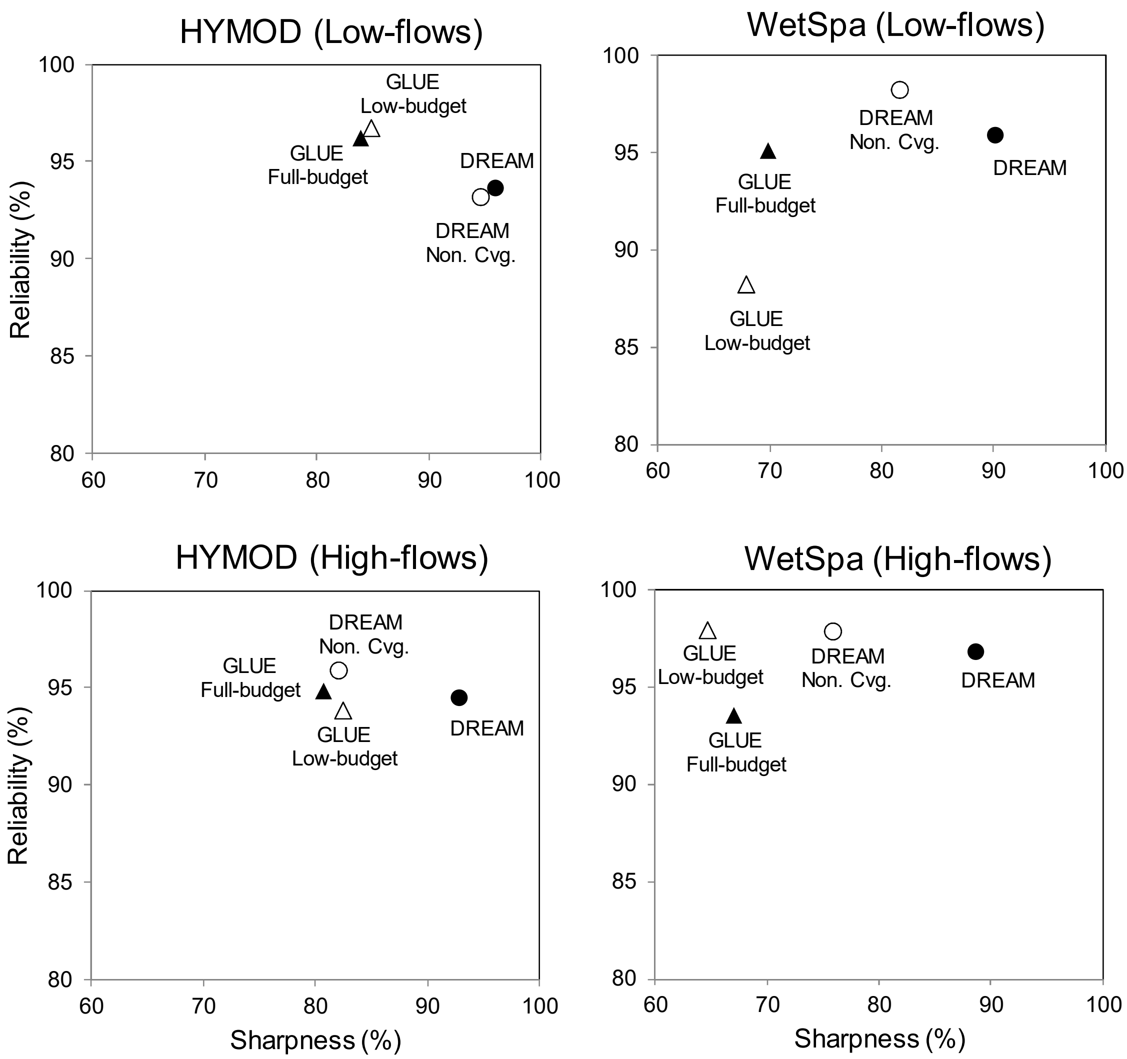

Figure 4. Validation period reliability and sharpness for low-flows (upper panels) and high-flows (lower panels) in application of different techniques (shown in different shapes) to the HYMOD and WetSpa simulation models.

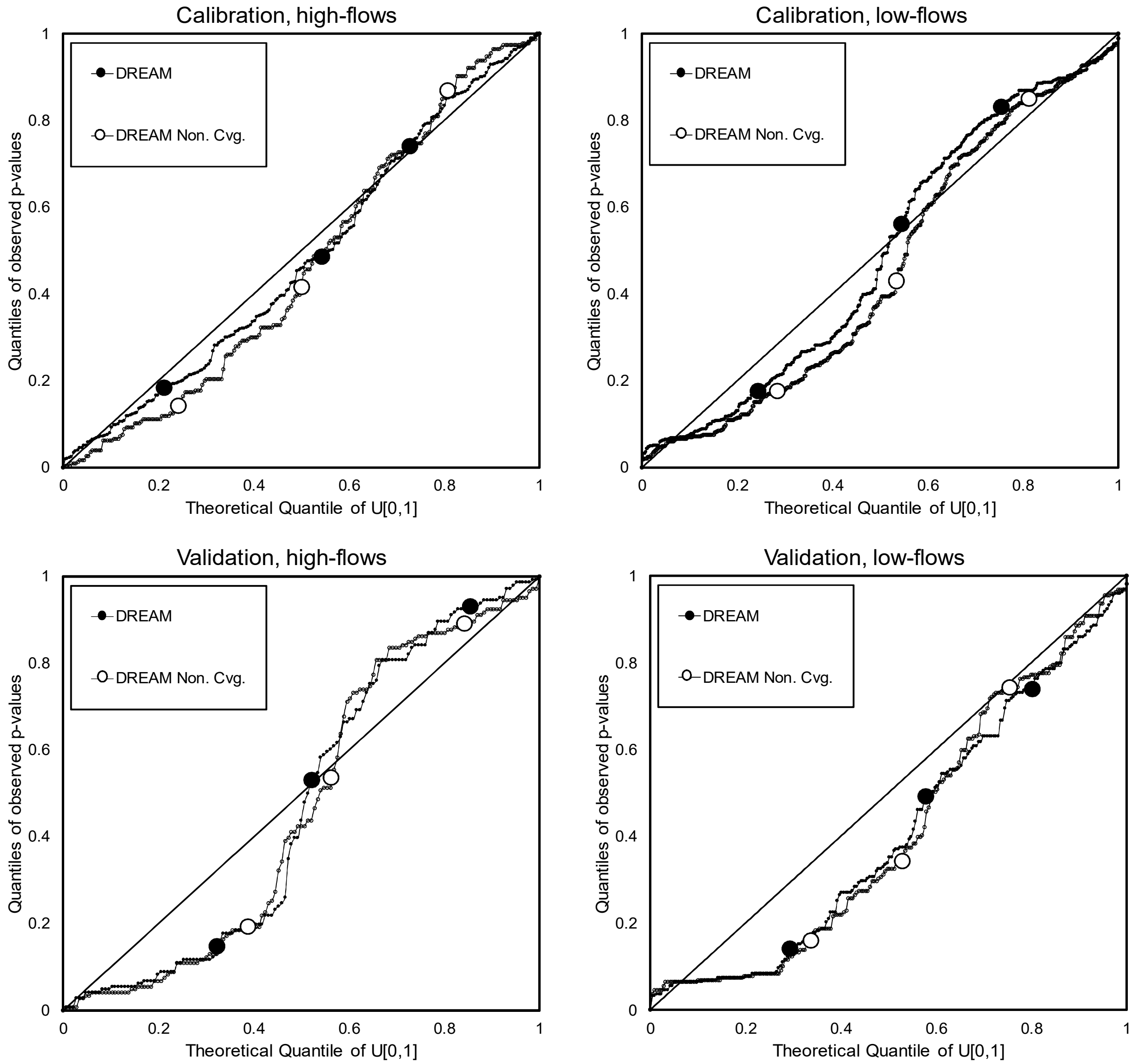

Figure 5. QQ plot of Bayesian p-values for high- and low-flows derived from converged and non-converged DREAM, for calibration (upper) and validation (bottom) periods of the HYMOD Leaf River case study.

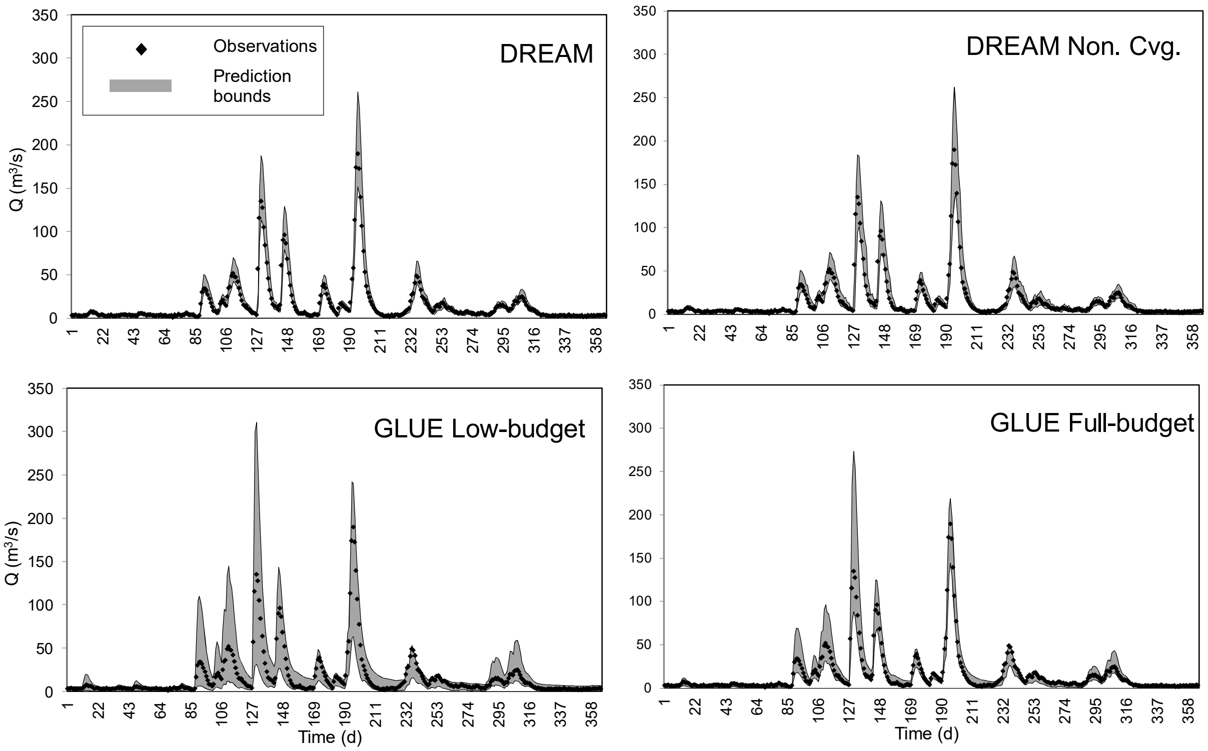

Figure 6. Prediction bounds and observations for the validation period in the HYMOD case study.

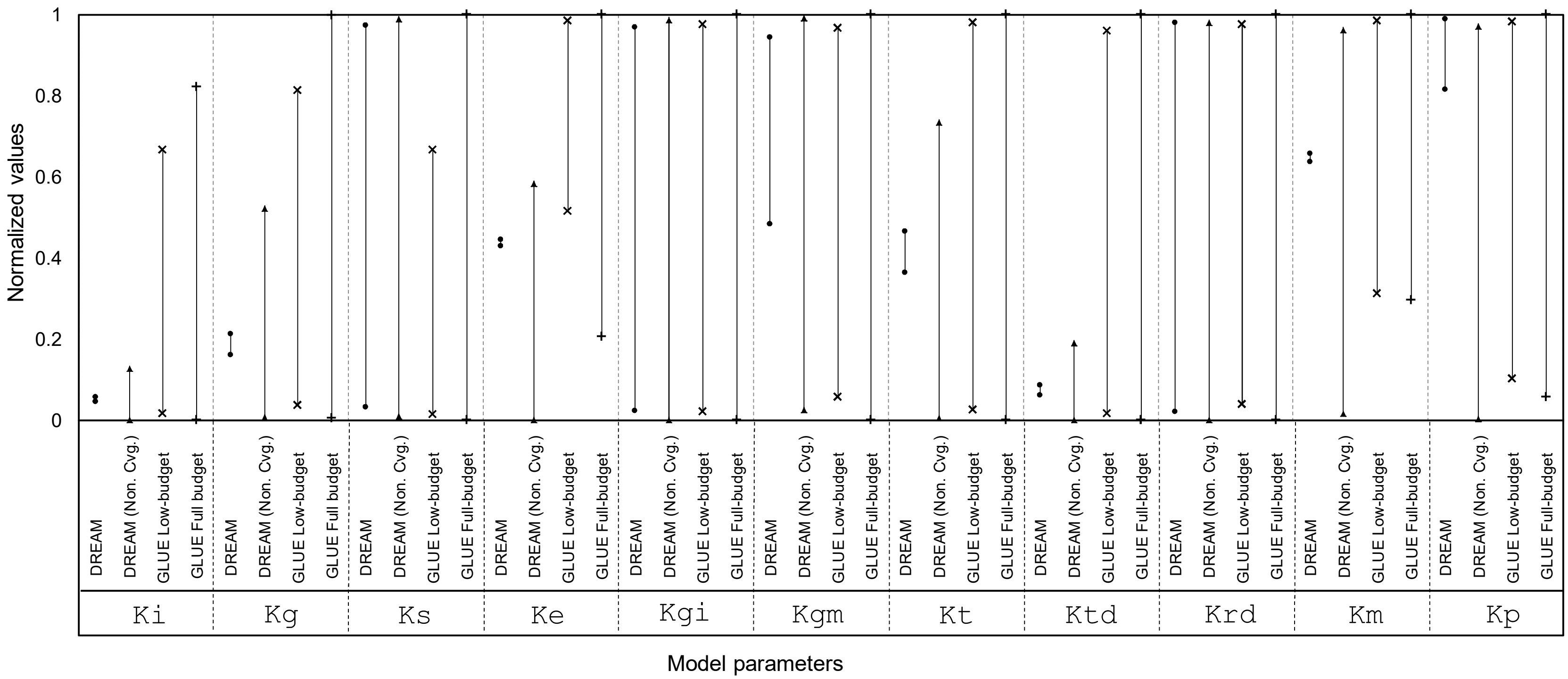

Figure 7. Posterior ranges of WetSpa parameters derived by different uncertainty-based calibration techniques.

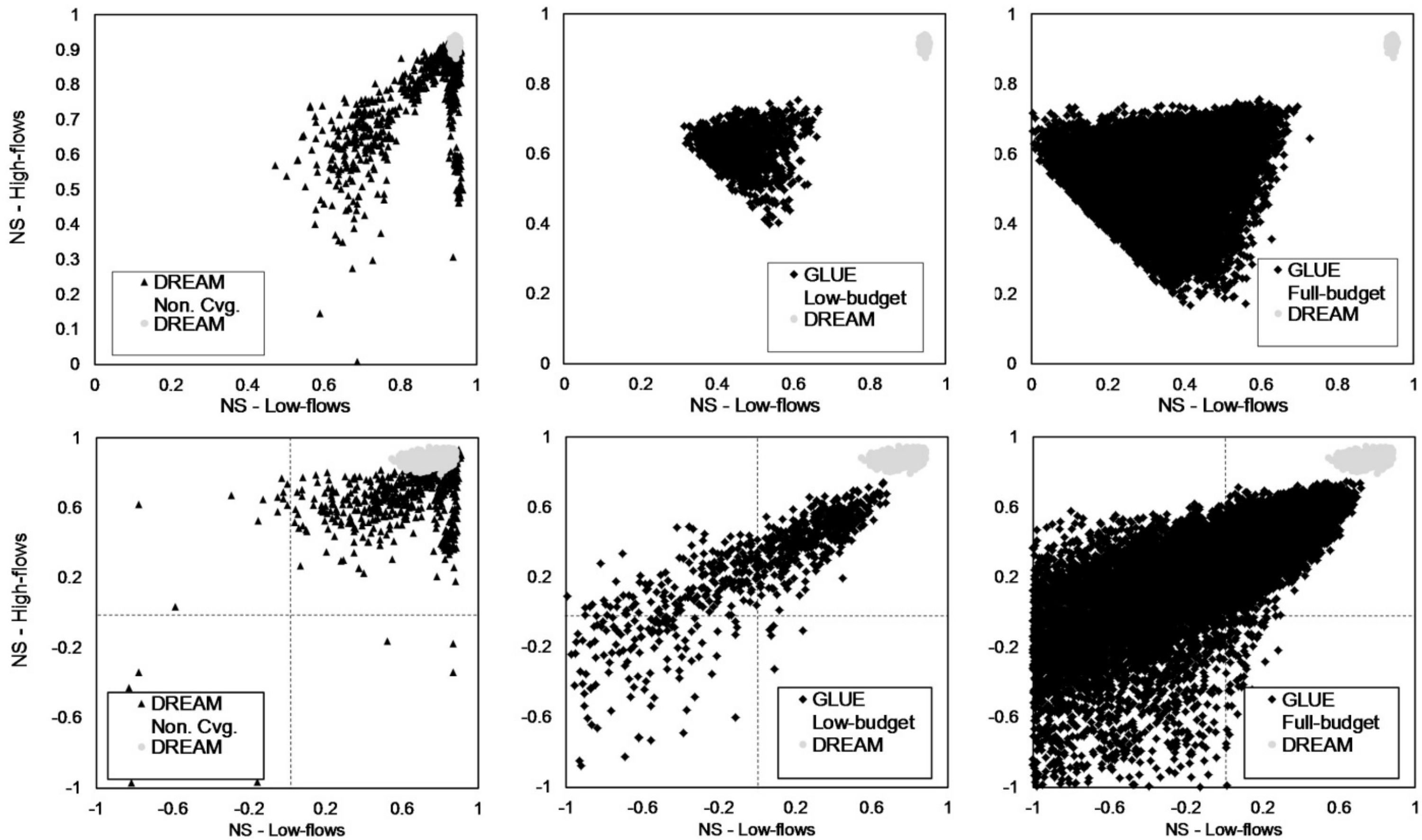

Figure 8. NS values of low-flows (horizontal axis) and high-flows (vertical axis) in calibration (upper panel) and validation (lower panel) period for WetSpa case study, derived from DREAM (light points) versus non-converged DREAM and GLUE methods (dark points).

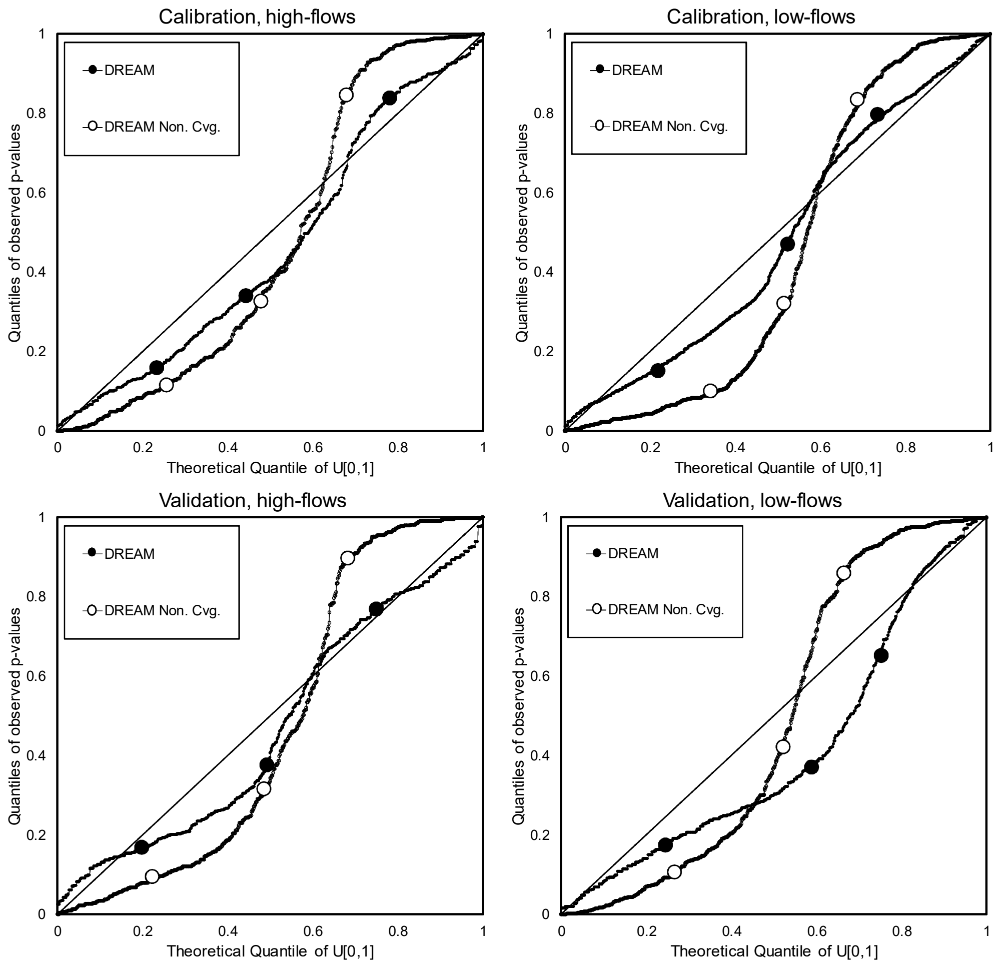

Figure 9. QQ plot of Bayesian p-values for high- and low-flows derived from converged and non-converged DREAM, for the calibration (upper) and validation (bottom) periods of the WetSpa Hornad River case study.

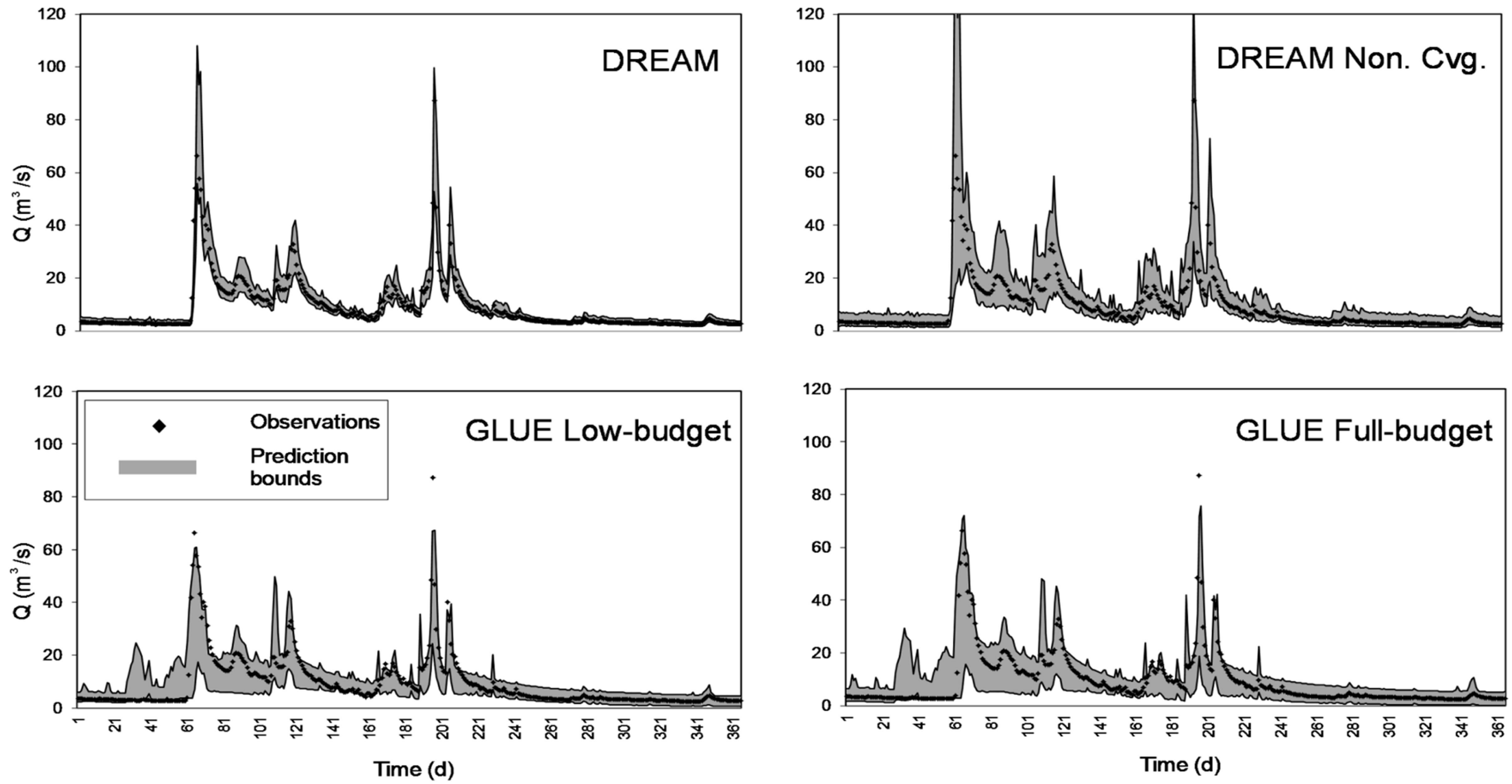

Figure 10. Prediction bounds and observations for the year 1999 of validation period for the WetSpa Hornad River case study.

List of Figures:

Figure 1. Schematic of the predictive QQ plot based on Thyer et al. (2009)

Figure 2. Posterior ranges of HYMOD parameters for the Leaf River case study; The parameter ranges correspond to 95% posterior intervals for different uncertainty analysis methods.

Figure 3. NS values of low-flows (horizontal axis) and high-flows (vertical axis) in calibration (upper panels) and validation (lower panels) period for HYMOD case study, derived from DREAM (light points) versus non-converged DREAM and GLUE methods (dark points).

Figure 4. Validation period reliability and sharpness for low-flows (upper panels) and high-flows (lower panels) in application of different techniques (shown in different shapes) to the HYMOD and WetSpa simulation models.

Figure 5. QQ plot of Bayesian p-values for high- and low-flows derived from converged and non-converged DREAM, for calibration (upper) and validation (bottom) periods of the HYMOD Leaf River case study.

Figure 6. Prediction bounds and observations for the validation period in the HYMOD case study.

Figure 7. Posterior ranges of WetSpa parameters derived by different uncertainty-based calibration techniques.

Figure 8. NS values of low-flows (horizontal axis) and high-flows (vertical axis) in calibration (upper panel) and validation (lower panel) period for WetSpa case study, derived from DREAM (light points) versus non-converged DREAM and GLUE methods (dark points).

Figure 9. QQ plot of Bayesian p-values for high- and low-flows derived from converged and non-converged DREAM, for the calibration (upper) and validation (bottom) periods of the WetSpa Hornad River case study.

Figure 10. Prediction bounds and observations for the year 1999 of validation period for the WetSpa Hornad River case study.