Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling

Blasone, Roberta-Serena; Vrugt, Jasper A.; Madsen, Henrik; Rosbjerg, Dan; Robinson, Bruce A.; Zyvoloski, George A.

Published in:
Advances in Water Resources

Link to article, DOI:
10.1016/j.advwatres.2007.12.003

Publication date:
2008

Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov chain Monte Carlo sampling

Roberta-Serena Blasone[1], Jasper A. Vrugt[2], Henrik Madsen[3], Dan Rosbjerg[1], Bruce A. Robinson[2] and George A. Zyvoloski[2]

(1) *Institute of Environment & Resources, Technical University of Denmark, Bygningstorvet 115, DK-2800 Kgs. Lyngby, Denmark*
(2) *Earth and Environmental Sciences Division, Los Alamos National Laboratory, Mail Stop T003, Los Alamos, NM, 87545*
(3) *DHI Water · Environment · Health, Agern Allé 5, DK-2970 Hørsholm, Denmark*

Submitted

ABSTRACT

In the last few decades hydrologists have made tremendous progress in using dynamic simulation models for the analysis and understanding of hydrologic systems. However, predictions with these models are often deterministic, and as such they focus on the most probable forecast, without an explicit estimate of the associated uncertainty. This uncertainty arises from incomplete process representation, uncertainty in initial conditions, input, output, and parameter error. The Generalized Likelihood Uncertainty Estimation (GLUE) framework was one of the first attempts to represent prediction uncertainty within the context of Monte Carlo (MC) analysis coupled with Bayesian estimation and propagation of uncertainty. Because of its flexibility, ease of implementation, and its suitability for parallel implementation on distributed computer systems, the GLUE method has been used in a wide variety of applications. However, the GLUE method has been criticized for not being formally Bayesian, and for often being implemented with a stratified MC parameter sampling scheme that does not properly sample the high probability density of the parameter space. In this paper we alleviate these problems through the development of an adaptive Markov Chain Monte Carlo sampling (the Shuffled Complex Evolution Metropolis, SCEM-UA, algorithm) scheme, and by determining the value of the cutoff threshold based on statistical arguments which allow for a better representation of the prediction uncertainty bounds. We demonstrate the superiority of this revised GLUE method with three different conceptual watershed models of increasing complexity, using both synthetic and real-world streamflow data from two different catchments with different hydrologic regimes.

Keywords: uncertainty assessment; rainfall-runoff modeling; GLUE; Markov chain Monte Carlo method; SCEM-UA.

1. INTRODUCTION AND SCOPE

It is an accepted fact that a hydrologic model prediction should not be deterministic, most-probable representation, but should also explicitly include an estimate of uncertainty. Uncertainty in model predictions arise from measurement errors associated with the system input (forcing) and output, from model structural errors arising from the aggregation of spatially distributed real-world processes into a mathematical model, and from problems with parameter estimation. Realistic assessment of these various sources of uncertainty is important for science-based decision making and will help direct resources towards model structural improvements and uncertainty reduction.

Recent years have seen an explosion of methods to derive meaningful prediction uncertainty bounds on our model predictions. Methods to represent model parameter, state and prediction uncertainty include classical Bayesian (*Kuczera and Parent*, 1998; *Thiemann et al.*, 2001; *Vrugt et al.*, 2003a), pseudo-Bayesian (*Beven and Binley*, 1992; *Freer et al.*, 1996), set-theoretic (*Keesman*, 1990; *Klepper et al.*, 1991; *van Straten and Keesman*, 1991; *Vrugt et al.*, 2001), multiple criteria (*Gupta et al.*, 1998; *Yapo et al.*, 1998; *Boyle et al.*, 2000, *Madsen*, 2000; *Madsen*, 2003; *Vrugt et al.*, 2003b) and sequential data assimilation methods (*Madsen et al.*, 2003; *Vrugt et al.*, 2005; *Moradkhani et al.*, 2005). These methods all have strengths and weaknesses, but differ in their underlying assumptions and how the various sources of error are being treated and made explicit. Among these methods, the Generalized Likelihood Uncertainty Estimation (GLUE) methodology of *Beven and Binley* (1992), inspired by the *Hornberger and Spear* (1981) method of sensitivity analysis was one of the first attempts to represent prediction uncertainty. This method maps the uncertainty in the modeling process onto the parameter space, and operates within the context of Monte Carlo analysis coupled with Bayesian estimation and propagation of uncertainty. The GLUE approach calls for rejecting the concept of a unique global optimum parameter set within some particular model structure, instead recognizing the acceptability, within a model structure, of different parameter sets that are similarly good in producing fit model predictions. This concept, defined as equifinality, is directly addressed by the evaluation of different sets of parameters within a Bayesian MC framework. The outputs of the GLUE procedure are posterior parameter distributions and associated prediction uncertainty bounds.

Since its introduction in 1992, the GLUE framework has found widespread application for uncertainty assessment in environmental modeling, including rainfall-

runoff modeling (*Beven and Binley*, 1992; *Freer et al.*, 1996; *Lamb et al.*, 1998), soil erosion modeling (*Brazier et al.*, 2001), modeling of tracer dispersion in a river reach (*Hankin et al.*, 2001), groundwater modeling and well capture zone delineation (*Feyen et al.*, 2001; *Jensen* 2003), unsaturated zone modeling (*Mertens et al.*, 2004), flood inundation modeling (*Romanowicz et al.*, 1996; *Aronica et al.*, 2002), land-surface–atmosphere interactions (*Franks et al.*, 1997), soil freezing and thawing modeling (*Hansson and Ludin*, 2006), crop yields and soil organic carbon modeling (*Wang et al.*, 2005), ground radar-rainfall estimation (*Tadesse and Anagnostou*, 2005), and distributed hydrological modeling (*McMichael et al.*, 2006; *Muleta and Nicklow*, 2005). The popularity of GLUE is probably best explained by its conceptual simplicity, relative ease of implementation and use, and its ability to handle different error structures and models without major modifications to the method itself.

Despite this progress made, various contributions to the hydrologic literature have criticized GLUE for not being formally Bayesian, requiring subjective decisions on the likelihood function and cutoff threshold separating behavioral from non-behavioral models, and for not implementing a statistically consistent error model (*Montanari*, 2005; *Christensen*, 2004). Moreover, those implementing the GLUE method typically use a rather simplistic stratified MC sampling scheme (Latin Hypercube Sampling – LHS – *McKay et al.*, 1979) to sample from the prior parameter distributions and to derive estimates of the posterior parameter probability density functions (PDFs) and associated model output prediction uncertainty bounds. While this approach may be adequate for low-dimensional sampling problems, it is unlikely to result in stable and consistent estimates of the posterior PDF for high-dimensional estimation problems. To compensate for this drawback, the LHS method typically requires many thousands of model simulations to result in a statistically sufficient number of behavioral parameter sets to draw inferences from (*Pappenberger et al.*, 2005; *Montanari*, 2005). Even at this extreme, it may be difficult to obtain a statistically significant number of behavioral models. In those situations, one should be particularly careful not to infer erroneous conclusions about parameter identifiability and equifinality (*Boyle et al*. 2000; *Vrugt et al*., 2003a).

In a separate line of research, Markov Chain Monte Carlo (MC^2) methods have been developed to locate the high probability density (HPD) region of the parameter space efficiently. These methods generate a random walk through the parameter space and successively visit solutions with frequency proportional to their weight in the posterior PDF. To do so, MC^2 methods use information from accepted solutions in the past to improve their search efficiency and converge to the posterior PDF of the parameters. For example, the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm, recently developed by *Vrugt et al.* (2003a), is a general purpose optimization algorithm that uses adaptive MC^2 sampling to provide an efficient search of the parameter space.

In this paper, we examine the use of adaptive MC^2 sampling within the GLUE methodology to improve the sampling of the HPD region of the parameter space. The concept is to construct the initial sample using the SCEM-UA algorithm, and derive the associated model output estimates (as the median of the distribution) and prediction uncertainty bounds (as percentiles of the output prediction) using the GLUE method. By using an algorithm designed to find the global optimum in the parameter space, we believe that this revised GLUE method should locate behavioral models much more efficiently, thereby improving the computational efficiency and statistical validity of the predictive uncertainty results.

This paper is structured as follows. Section 2 briefly describes the GLUE methodology and discusses the LHS and SCEM-UA methods for sampling of the prior parameter distribution. In section 3, we discuss the three conceptual watershed models and catchments used to test the revised GLUE methodology. Section 4 discusses the results of the analysis, comparing the LHS method and SCEM-UA algorithm for generating the initial sample, and examining the influence of model complexity on the sampling and GLUE-derived median forecasts and uncertainty bounds. Finally, section 5 summarizes the most important findings.

2. METHODS

In this section we briefly discuss the GLUE methodology, and describe the LHS and SCEM-UA algorithms for sampling of the prior parameter distribution.

2.1. The GLUE methodology

The GLUE procedure is a Monte Carlo method, the objective of which is to identify a set of behavioral models within the universe of possible model/parameter combinations. The term "behavioral" is used to signify models that are judged to be "acceptable," that is, not ruled out, on the basis of available data and knowledge. To implement GLUE, 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% confidence level in most of the published GLUE studies. These respective bounds are called 90% confidence bounds or prediction limits.

2.2. Parameter Sampling Strategy

To sample the prior parameter distribution, practitioners of the GLUE methodology generally implement a Latin Hypercube Sampling (LHS) strategy. This stratified random sampling method, though relatively simple to implement, is unlikely to densely sample the parameter space close to the global optimum with a dense distribution of points. Our conjecture is that considerable improvements in sampling can be made by using an adaptive sampling method that uses information from past draws to update the search direction. Such a method would probably result in parameter and prediction uncertainty estimates that are more reliable from a statistical point of view.

In this paper, we explore the use of the SCEM-UA algorithm to achieve this improvement. Instead of randomly sampling the prior parameter space, the SCEM-UA algorithm generates a random walk through the parameter space such that any individual state is visited with a frequency proportional to its weight in the posterior PDF. In contrast to LHS, the SCEM-UA algorithm is an adaptive sampler that periodically updates the covariance (size and direction) of the sampling or proposal distribution during the evolution of the sampler toward the HPD region of the parameter space, using information from the sampling history induced in the transitions of the Markov Chain. Experiments using synthetic mathematical test functions have demonstrated that the SCEM-UA algorithm has the appropriate ergodic properties, and provides a more efficient sampling of the HPD region of the parameter space than traditional Metropolis-Hastings samplers (*Vrugt et al.*, 2003a).

In the SCEM-UA algorithm, a predefined number of different Markov Chains are initialized from the highest likelihood values of the initial population. These chains independently explore the search space, but communicate with each other through an external population of points, which are used to continuously update the size and shape of the proposal distribution in each chain. The MC^2 evolution is repeated until the *R*-statistic of *Gelman and Rubin* (1992) indicates convergence to a stationary posterior distribution. An extensive description and explanation of the method appears in *Vrugt et al.* (2003a) and so will not be repeated here.

The rationale for adopting this sampling strategy in the GLUE methodology rests on arguments of the generation of statistically representative results, as well as on computational efficiency. Because the SCEM-UA algorithm provides an adequate sampling of the HPD region of the parameter space, it will find a greater number of

behavioral solutions, thereby yielding more statistically valid estimates of parameter and prediction uncertainty. Also, because the SCEM-UA method is well suited for searching high-dimensional parameter spaces, far fewer model evaluations will be needed to provide a good approximation of the posterior PDF. Finally, although the equifinality method that inspired the GLUE method downplays the importance of finding the global optimum in a global search procedure (e.g. Beven, 2006), we believe that it is logical to take steps to ensure that the global optimum is contained in the family of behavioral models. The SCEM-UA algorithm is designed to find this optimal parameter set.

2.3. Choice of the likelihood function

Various likelihood functions have been proposed in the literature (e.g. *Beven and Binley*, 1992; *Romanowicz et al.*, 1994; *Christensen*, 2004; *Montanari*, 2005) as measures that quantify the closeness between model simulations and observations. Most of these functions are considered pseudo-likelihood functions because they do not adhere to formal Bayesian statistics, but instead are designed to implicitly account for errors in model structure and input data, and to avoid over-conditioning to a single parameter set. In this study we implement the following commonly used likelihood function:

$$L(\theta_i \mid Y) = \exp\left\{- N \cdot \sigma_i^2 / \sigma_{obs}^2 \right\} \quad (1)$$

where $L(\theta_i|Y)$ is the likelihood measure for the i-th model conditioned on the observations Y, σ_i^2 is the error variance for the i-th model (i.e. the combination of the model and the i-th parameter set) and σ_{obs}^2 is the variance of the observations. The exponent N is an adjustable parameter that sets the relative weightings of the better and worse solutions: higher N values have the effect of giving more weight to the best simulations, thus increasing the difference between good and bad solutions (*Freer et al.*, 1996). Small values for N result in a flat likelihood function with significant probability mass extending over a large part of the parameter space. On the contrary, relatively high values for N will result in a peaked likelihood function, with a well-defined global optimal solution.

This likelihood function was chosen principally because it is commonly used within the GLUE methodology, so using it facilitates comparison with other studies. Furthermore, varying N in Eq. (1) is a simple and flexible way to test the influence of the shape of the likelihood function on the efficiency of the LHS and SCEM-UA algorithm for sampling of the prior distribution. In this paper we provide a comparison

assessment of LHS and the SCEM-UA algorithm for different N values ranging from 1 to 100.

2.4. Choice of the cutoff threshold for the behavioral simulations

One criticism of the GLUE methodology is that the prediction uncertainty bounds are subjective, based on an arbitrary cutoff to differentiate between behavioral and non-behavioral simulations. Ideally, the prediction uncertainty spread should be as small as possible, but consistent with observations, so that the predictive PDF is as sharp as possible. Stated differently, if the model is required to generate a probabilistic forecast at a given confidence level, say, 95%, then the predictions should encompass 95% of the observations. Unfortunately, most formulations of the GLUE methodology do not guarantee that the appropriate percentage of the observations lies within the uncertainty bounds. In this study, instead of using predefined quantiles from the GLUE derived output CDF, we tune the uncertainty bounds so they exhibit the appropriate coverage. For all case studies we use 90% prediction intervals. These intervals are found by a trial-and-error method in which the acceptance criterion is adjusted and the coverage is computed over a fixed calibration period.

3. CASE STUDIES

In this section we describe the three conceptual watershed models used in our comparison analysis, and discuss the synthetic and measured streamflow data used.

3.1. Models Used and Prior Uncertainty Ranges

Three conceptual watershed models of increasing complexity are used in the present study: HYMOD (*Boyle*, 2000), NAM (*Nielsen and Hansen*, 1973; *Havnø et al*., 1995) and the Sacramento Soil Moisture Accounting Model (SAC-SMA: *Burnash et al.*, 1973; *Burnash*, 1995). Brief descriptions of each model are presented in the following three sections. These models differ in their structure, simulated hydrologic processes, and number of calibration parameters, thereby allowing us to examine how model complexity affects the results of our sampling and uncertainty assessment analysis.

3.1.1. The HYMOD model

The HYMOD model consists of a relatively simple rainfall excess model, associated with two series of linear reservoirs: three identical reservoirs generating the quick flow response and a single reservoir for the slow response. A slightly different version of HYMOD is employed in this study: two identical reservoirs in series for the

quick response, and two reservoirs in parallel for the slow response. The 5 model parameters (summarized in Table 1) assessed in this work are the same as those considered in the studies by *Vrugt et al.* (2003b) and *Montanari* (2005). The last column in Table 1 lists the prior uncertainty ranges used to generate the initial sample.

Table 1. Parameters of the models used and their prior uncertainty ranges.

HYMOD

Parameter	*Unit*	*Range*	*Description*
C_{max}	[mm]	1 - 500	maximum storage capacity in the catchment
b_{exp}	[-]	0.1 - 2	degree of spatial variability of soil moisture capacity within the catchment
A	[-]	0. - 0.99	factor distributing the flow between the two series of reservoirs
R_s	[day]	0 - 0.1	residence time of the linear slow response reservoir
R_q	[day]	0.1 - 0.99	residence time of the linear quick response reservoir

NAM

Parameter	*Unit*	*Range*	*Description*
U_{max}	[mm]	1- 50	maximum water content (size) of the surface storage
L_{max}	[mm]	50 - 1000	maximum water content (size) of the root zone storage
CQOF	[0,1]	0 - 1	fraction of excess rainfall that contributes to the overland flow
CKIF	[hours]	0.01 - 2000	time constant for drainage of interflow
CK_{12}	[hours]	3 - 100	time constant for routing interflow and overland flow; it determines the shape of hydrograph peaks
TOF	[-]	0 - 0.99	threshold value for overland flow, which is generated only for relative moisture content of the lower zone higher than TOF
TIF	[-]	0 - 0.99	threshold value for interflow (similar effect on interflow as TOF has on overland flow)
TG	[-]	0 - 0.99	root zone threshold value for recharge (similar effect on recharge as TOF on overland flow)
CK_{BF}	[hours]	0.01 - 5000	time constant for baseflow, it determines the shape of the hydrograph in dry periods (exponential decay)
C_{snow}	[mm/°C/day]	0.5 - 10	degree-day coefficient for determining snow melting

SAC-SMA

Parameter	*Unit*	*Range*	*Description*
UZTWM	[mm]	1 - 150	upper zone tension water capacity
UZFWM	[mm]	1 - 150	upper zone free water capacity
UZK	[day^{-1}]	0.1 - 0.5	upper zone free water lateral depletion rate
PCTIM	[-]	0.000001 - 0.1	fraction of the impervious area
ADIMP	[-]	0 - 0.4	fraction of the additional impervious area
ZPERC	[-]	1 - 250	maximum percolation rate coefficient
REXP	[-]	0 - 5	exponent of the percolation equation
LZTWM	[mm]	1 - 500	lower zone tension water capacity
LZFSM	[mm]	1 - 1000	lower zone supplementary free water capacity
LZFPM	[mm]	1 - 1000	lower zone primary free water capacity
LZPK	[day^{-1}]	0.0001 - 0.25	lower zone primary free water depletion rate
LZSK	[day^{-1}]	0.01 - 0.25	lower zone supplementary free water depletion rate
PFREE	[-]	0 - 0.6	fraction percolating from upper to lower zone free water storage
RTCOEF	[day^{-1}]	0 - 1	retention coefficient of routing linear reservoirs

3.1.2. The NAM model

The NAM model is a deterministic, lumped, conceptual rainfall-runoff model originally developed at the Technical University of Denmark (*Nielsen et al.*, 1973; *Havnø et al*., 1995). It has been used in many different applications and studies (*Storm et al.*, 1988; *Lorup et al.*, 1998; *Madsen*, 2000; *Khu and Madsen*, 2005). The NAM model describes, in a simplified quantitative form, the behavior of the different land phase of the hydrological cycle, accounting for the water content in different mutually interrelated storages. These storages are the surface zone storage (water content intercepted by vegetation, in surface depression and in the uppermost few centimeters of the ground), the root-zone storage, the ground-water storage and the snow storage. The river routing is done through linear reservoirs that represent the overland flow (two identical liner reservoirs in series), the interflow (a single reservoir) and the baseflow (a single reservoir), each characterized by a specific time constant. The NAM model specifies 10 parameters that need to be determined by calibration against a historical record of streamflow data. A description of these parameters, including their prior uncertainty ranges is given in Table 1.

3.1.3. The Sacramento Soil Moisture Accounting (SAC-SMA) model

The Sacramento soil moisture accounting model, SAC-SMA, is a lumped conceptual watershed model developed by *Burnash et al.* (1973; – see also *Burnash*, 1995). It is currently used by the National Weather Service River Forecast System (NWSRFS) center to perform real-time river and flood forecasts as well as long term predictions.

The SAC-SMA model distributes soil moisture in various depths and energy states of the soil with a network of interconnected tanks. It is constituted by an upper and a lower zone, each including tension and free-water storages. These storages interact with each other and with the other catchment components through the processes of evapotranspiration, vertical drainage (percolation), and generation of five different runoff components. In the original Sacramento model, the runoff components combine to produce the river runoff through a unit hydrograph routing. In the version of the Sacramento model used in this study, the routing module is replaced with a series of 3 linear Nash-Cascade reservoirs, all characterized by the same retention coefficient, *RTCOEF*. This formulation of the SAC-SMA model does not require independent derivation of the unit hydrograph, and therefore provides a more flexible formulation for application in different watersheds. In this study, the parameters *SIDE*, *RSERV* and *RIVA* were fixed at values recommend in *Peck* (1976); this leaves a total of 14 parameters in our analysis. Table 1 provides a condensed overview and description of the SAC-SMA calibration parameters, including their prior uncertainty ranges.

3.2. Hydrologic Systems and Data Used

We compare the usefulness and power of our revised GLUE method (using SCEM-UA) to the traditional GLUE approach (using LHS) by application to two different catchments with significantly different hydrologic regimes. The first is the Tryggevælde catchment, located in the eastern part of Denmark. This catchment, which has an area of approximately 130.2 km^2, consists of predominantly clayey soils, and has an average daily river discharge of about 1 m^3/s. For the period between January 1, 1975 and December 31, 1984, available data for this catchment includes the mean areal precipitation (mm/d), potential evapotranspiration (mm/d), daily average temperature (°C) and discharge (m^3/s). To reduce sensitivity to state value initialization, a one-year warm up period was used in which no updating of the likelihood function was performed.

The second system studied is the Leaf River catchment, located in southern Mississippi. It is a principal tributary of the Pascagoula River, which flows to the Gulf of Mexico. It is a humid watershed, with an area of about 1944 km^2. The available data record consists of mean daily precipitation (mm/d), potential evapotranspiration (mm/d), and daily streamflow (m^3/s). The Leaf River data have been discussed and used extensively in previous studies. In the present study, data in the period between July 28, 1952 and September 30, 1962 are used, with a warm-up period of 65 days.

The Tryggevælde and Leaf River watersheds have quite different hydrologic regimes, thereby providing diverse data sets for testing the revised GLUE method. For example, the average daily runoff of the Leaf River (27.13 m^3/s) is much higher than that of the Tryggevælde catchment. In addition, the Leaf River data set includes a relatively large number of significant rainfall-runoff events, with streamflow values up to about 800 m^3/s.

Before analyzing the measured data sets, described in this section, initial benchmarking analyses were performed using corrupted synthetic data to test the performance of our sampling methods in the presence of data error only. This synthetic streamflow data was generated by calibrating the HYMOD, NAM and SAC-SMA model using the SCEM-UA algorithm, then using these parameter values in a forward model run to represent catchment behavior. This synthetic time series of streamflow data was then corrupted with a normally distributed error with an error deviation of 10% of the simulated value.

3.3. Implementation Details

For the data sets considered in this paper, the GLUE methodology is applied for the likelihood function defined in Eq. (1), using the initial sample of simulations derived with either the LHS and SCEM-UA sampling schemes. The analysis uses a total of

10,000 parameter combinations for the HYMOD model and 20,000 for the NAM and SAC-SMA models. Initial analyses have demonstrated that these numbers are sufficient and result in stabilized estimates of parameter and prediction uncertainty. After sampling, the GLUE-derived model prediction is then given by the median of the output distribution, and the associated uncertainty is derived from tuning the uncertainty bounds to obtain an approximate coverage of about 90% of the observations.

4. RESULTS AND DISCUSSION

This section presents analyses for the synthetic and measured data sets for the three different conceptual watershed models. The presentation is organized by starting with the synthetic data sets, discussing the GLUE results for (i) median prediction, (ii) prediction uncertainty bounds and (iii) parameter uncertainty and correlation. We then repeat this process for the measured data sets.

4.1. Synthetic data sets

4.1.1. Median GLUE prediction

Table 2 lists the likelihood values for different values of N of the best streamflow simulation from the initial sample generated with the LHS and SCEM-UA algorithm for the Tryggevælde watershed. Though we restrict attention to this catchment, similar results are found for the Leaf River watershed. The results in this Table clearly demonstrate the advantages of the SCEM-UA algorithm for sampling the prior parameter distribution. The algorithm generally finds better values of the likelihood function than LHS, with differences becoming larger with increasing N-values and model complexity. Small values of N result in a flat likelihood function with probability mass extending over a large range of the parameter space. Even with random sampling, it is then likely to find a parameter combination that reasonably fits the data. For increasing N values the likelihood function becomes peakier, and it is increasingly important to have the search capabilities of the SCEM-UA algorithm to find acceptable solutions. In addition, note that, as expected, increased modeling complexity will further reduce the chance of finding preferred solutions with random sampling (see, for example, the results of HYMOD, NAM and SAC-SMA for $N = 100$).

To verify whether the quality of the initial sample is influencing the deterministic forecast of the GLUE methodology, consider Table 3, which presents the likelihood value of the median prediction of the GLUE-derived CDF for the synthetic Tryggevælde data set using the HYMOD, NAM and SAC-SMA models. Consistent with the previous results, the median GLUE prediction derived from the initial samples created using the SCEM-UA algorithm is generally better than its counterpart derived using LHS. Adaptive MC^2 sampling improves the quality of the initial sample and

therefore the results derived with the GLUE method. Also notice that the GLUE derived median prediction is generally a better predictor than the best individual simulation in the initial sample (compare tables 2 and 3). This is particularly true for the SCEM-UA created initial sample and suggests that averaging of predictions of different parameter combinations increases predictive capabilities, something that is commonly observed with ensemble forecasting (*Raftery et al.*, 2005; *Vrugt and Robinson*, 2007). Again, differences between the LHS and SCEM-UA algorithm increase with increasing *N* value and complexity of the catchment model.

Table 2. Likelihood of the best runoff simulation from the initial sample generated with the LHS and SCEM-UA algorithm for different values of *N*: Tryggevælde watershed - synthetic data.

	SCEM-UA			*LHS*		
N	*HYMOD*	*NAM*	*SAC-SMA*	*HYMOD*	*NAM*	*SAC-SMA*
1	0.9798	0.9614	0.9760	0.9784	0.9527	0.9698
5	0.8995	0.7652	0.8552	0.8964	0.7847	0.8578
10	0.8177	0.6837	0.7324	0.8035	0.6158	0.7357
20	0.6827	0.4754	0.5982	0.6456	0.3792	0.5413
50	0.3888	0.1031	0.2659	0.3349	0.0885	0.2156
100	0.1609	0.0475	0.1681	0.1122	0.0078	0.0465

Table 3. Likelihood value of the median runoff estimate from the posterior CDF derived with the GLUE methodology: Tryggevælde watershed - synthetic data. The output CDF was tuned to contain 90% of the streamflow observations.

	SCEM-UA			*LHS*		
N	*HYMOD*	*NAM*	*SAC-SMA*	*HYMOD*	*NAM*	*SAC-SMA*
1	0.9815	0.9655	0.9771	0.9803	0.9276	0.9063
5	0.9112	0.8112	0.8883	0.9055	0.8263	0.8825
10	0.8246	0.7055	0.7995	0.8201	0.6854	0.7798
20	0.6899	0.5263	0.6370	0.6730	0.4752	0.6101
50	0.3826	0.2271	0.3515	0.3725	0.1514	0.2854
100	0.1546	0.0799	0.1420	0.1401	0.0018	0.0653

Next, the dependency of the goodness-of-fit of the GLUE-derived median streamflow estimate as function of the number of retained or behavioral solutions is analyzed. Plots of likelihood function versus the number of retained solutions are presented for the Tryggevælde and Leaf River data sets in Figures 1 and 2, respectively for the NAM model. First, note that accepting a relatively small number of solutions as behavioral generally produces the closest correspondence of the GLUE median output estimate with the observed streamflow data. On the order of 20 individual streamflow simulations (about 0.1% of the total sample) is required for accurate streamflow forecasting, whereas a larger sample of retained solutions decreases the goodness-of-fit of the median GLUE output estimate. However, a large sample improves the accuracy of the uncertainty bounds, as will be shown later. Thus, there is a considerable trade-off

between accuracy and precision when selecting the appropriate number of behavioral solutions. Given this situation, it is pertinent to point out that for the SCEM-UA sample, the likelihood value of the GLUE derived median output estimate appears to be less affected by the number of behavioral samples. The SCEM-UA algorithm provides a denser sampling in the vicinity of the HPD region of the parameter space, and thus yields a higher frequency of good solutions.

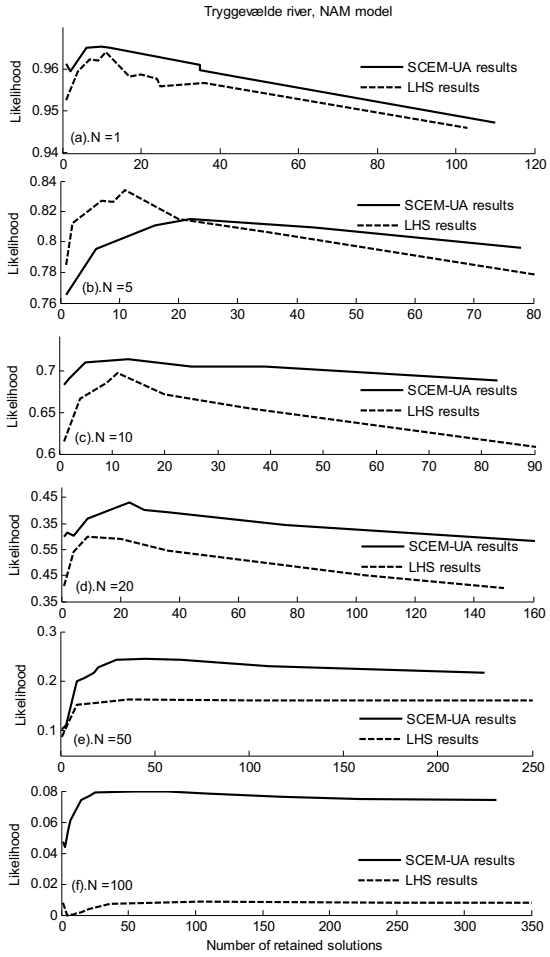

Figure 1. Tryggevælde watershed - NAM model.

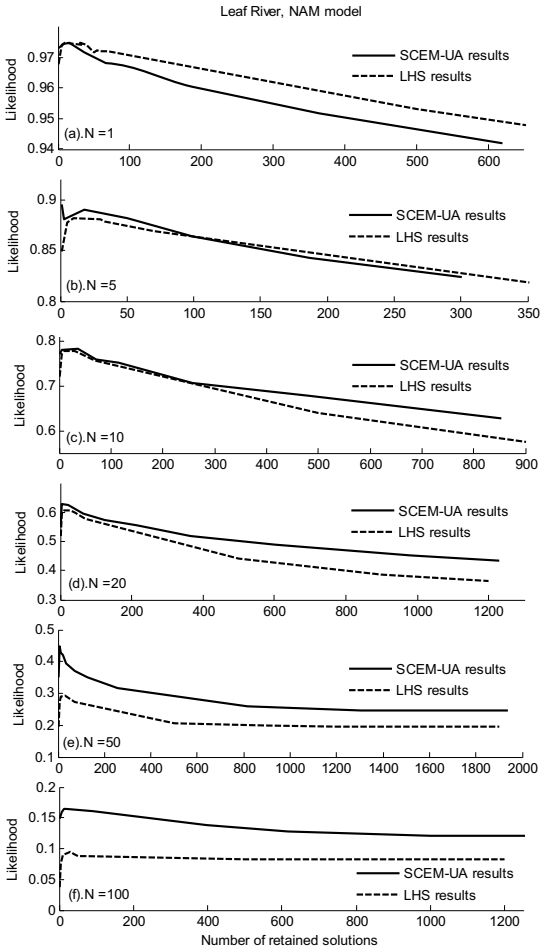

Figure 2. Leaf River watershed - NAM model.

Likelihood of the median GLUE estimates obtained from the LHS and SCEM-UA samples versus number of retained solutions. Plots correspond to different values of the exponent of the likelihood function, N: (a) N=1; (b) N=5; (c) N=10; (d) N=20; (e) N=50; (f) N=100.

Finally, the plots show that the relative difference between the likelihood of the estimated median hydrograph from the LHS and SCEM-UA sampling methods increases with increasing value of the exponent N of the likelihood function. This trend, found for both data sets, can be explained by the increased performance of the SCEM-UA algorithm in cases with a well-defined HPD region. In contrast, the SCEM-UA algorithm will not have good convergence properties when a large part of the parameter

space exhibits similar performance in producing the observed data (i.e. for low values of *N*). Thus, in these situations LHS might suffice to generate the initial sample. However, increasingly peaked likelihood functions, require optimization-based algorithms to locate and visit solutions in the HPD region.

4.1.2. Prediction Uncertainty Bounds

In this section we address the uncertainty bounds derived with the GLUE methodology for the LHS and SCEM-UA sampling methods. Accurate probabilistic forecasting requires that the uncertainty bounds are statistically meaningful and exhibit the appropriate coverage. Instead of focusing on the goodness-of-fit of the median output estimate of the GLUE-derived CDF, we examine the statistical properties of the ensemble of retained solutions.

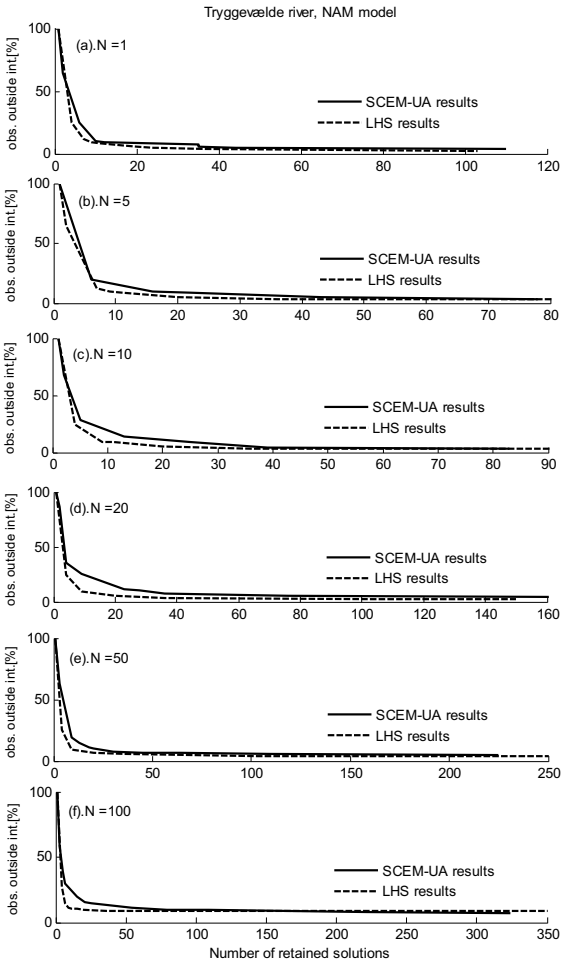

Figure 3. Tryggevælde watershed - NAM model.

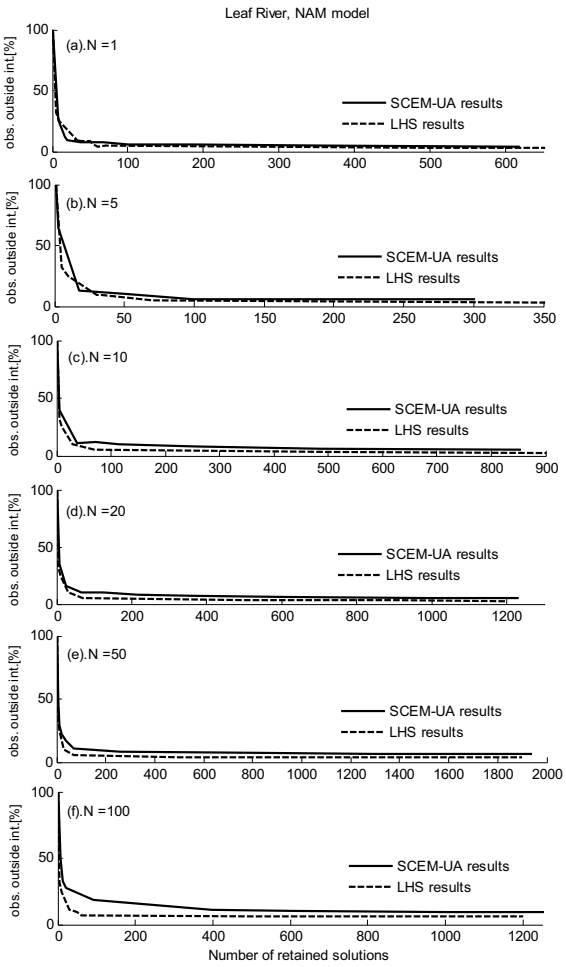

Figure 4. Leaf River watershed - NAM model.

Percentage of runoff observations outside GLUE LHS and SCEM-UA uncertainty intervals versus number of retained solutions. Plots correspond to different values of the exponent of the likelihood function, *N*: (a) *N*=1; (b) *N*=5; (c) *N*=10; (d) *N*=20; (e) *N*=50; (f) *N*=100.

Figures 3 and 4 are plots of the percentage of observations falling outside the prediction uncertainty bounds versus the number of retained parameter sets for the NAM model. For a given number of retained solutions, the GLUE-derived uncertainty bounds using LHS are generally larger than their counterparts derived from GLUE implemented with the SCEM-UA algorithm. The GLUE method implemented with SCEM-UA exhibits better predictive performance, resulting in less spread of the uncertainty bounds. This is further demonstrated in Figure 5, which depicts the average width of the streamflow uncertainty bounds as function of the number of retained solutions for different values of *N*.

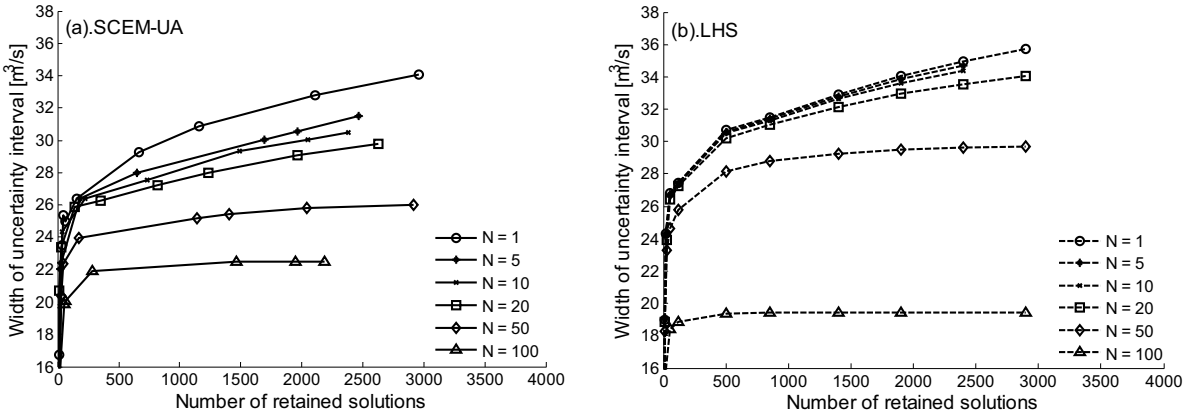

Figure 5. Leaf River watershed - SAC-SMA model: width of the uncertainty bounds as a function of the number of retained solutions: (a) SCEM-UA and (b) LHS results.

To examine this behavior further, consider Figures 6 (SAC-SMA model, Tryggevælde watershed) and 7 (NAM model, Leaf River Catchment), time-series plots of observed versus predicted streamflow data for a representative portion of the historical record. The top panels in both figures present the measured hyetograph, whereas the bottom two panels illustrate the GLUE-derived 90% uncertainty bounds for the predicted hydrographs for three different values of *N* (1, 20 and 100) using the (b) SCEM-UA and (c) LHS methods for sampling the prior parameter distribution.

The results for both sampling methods are qualitatively similar, and appear relatively unaffected by the choice of the value of the exponent *N* in the likelihood function. Although the uncertainty bounds exhibit the appropriate coverage and are generally centered on the observations, they appear to be unrealistically large, especially for the SAC-SMA model for low flows. This is a limitation of the GLUE method, and caused by the way the method treats uncertainty. The total uncertainty is mapped onto the parameters, without explicitly accounting for input and model structural errors. Much tighter uncertainty bounds that still exhibit the appropriate coverage can be obtained by using a formal Bayesian likelihood function (in the case of synthetic data)

or by accounting for input and model structural errors using state-space filtering methods such as the Ensemble Kalman Filter (*Vrugt et al.,* 2005).

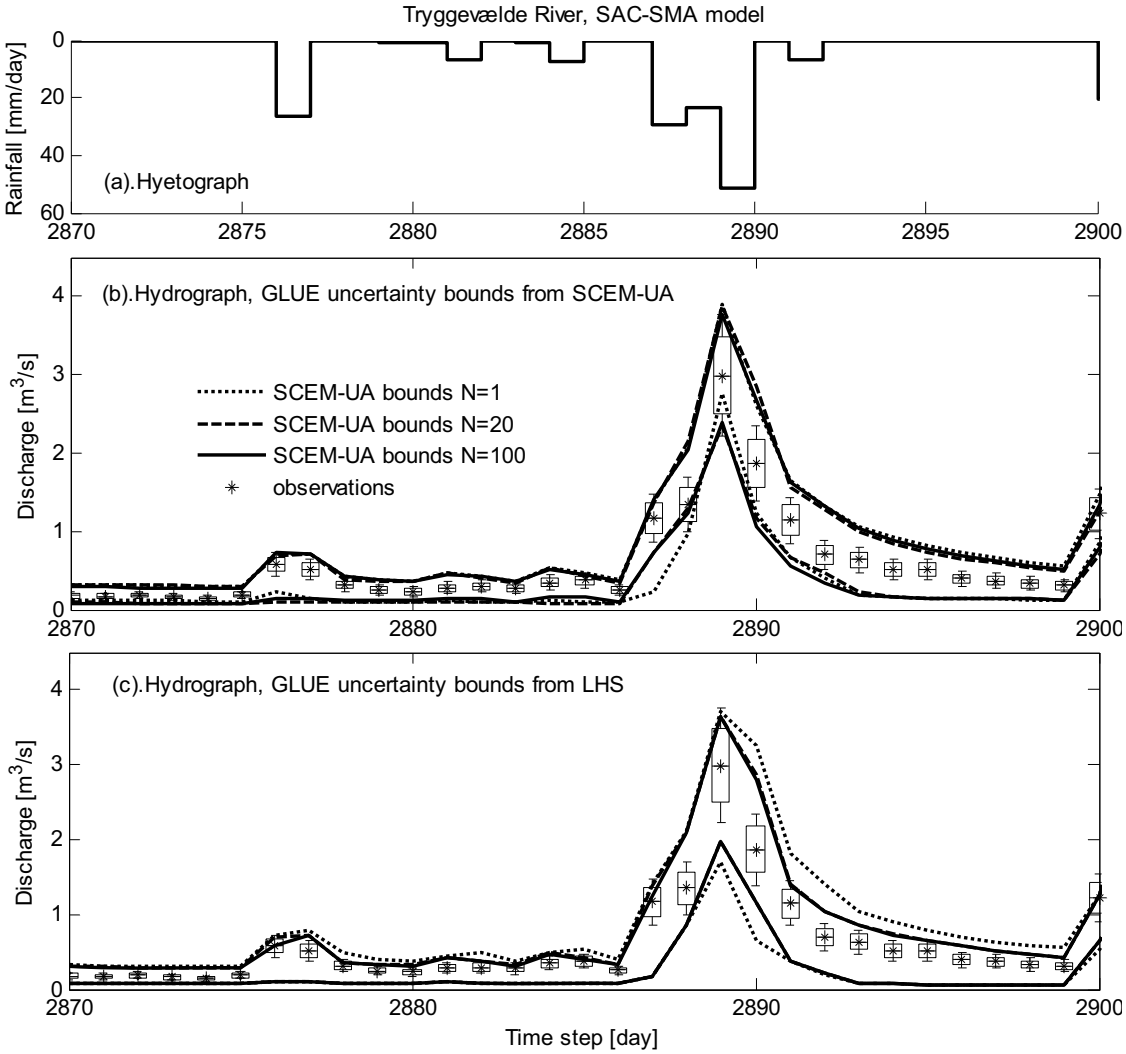

Figure 6. Tryggevælde watershed - SAC-SMA model: hyetograph (a) and hydrographs including the uncertainty bounds containing the 90% of the observations generated by GLUE from SCEM-UA (b) and LHS initial samples (c). The error bars in these plots represent the error properties of the streamflow data: the boxes correspond to the 5th and 95th percentiles of the error distribution, while the vertical lines extend up to the 0.5th and 99.5th percentiles.

4.1.3. Parameter Uncertainty and Correlation

In this section we compare the GLUE-derived posterior parameter PDFs from the LHS and SCEM-UA derived initial sample using the two sampling techniques. The GLUE-derived posterior parameter PDFs for different values of N are presented for the parameter L_{max} in the NAM model (Figure 8) and the parameter *LZFSM* in the SAC-

SMA model (Figure 9). This selection of parameters and models is representative of the entire set of results.

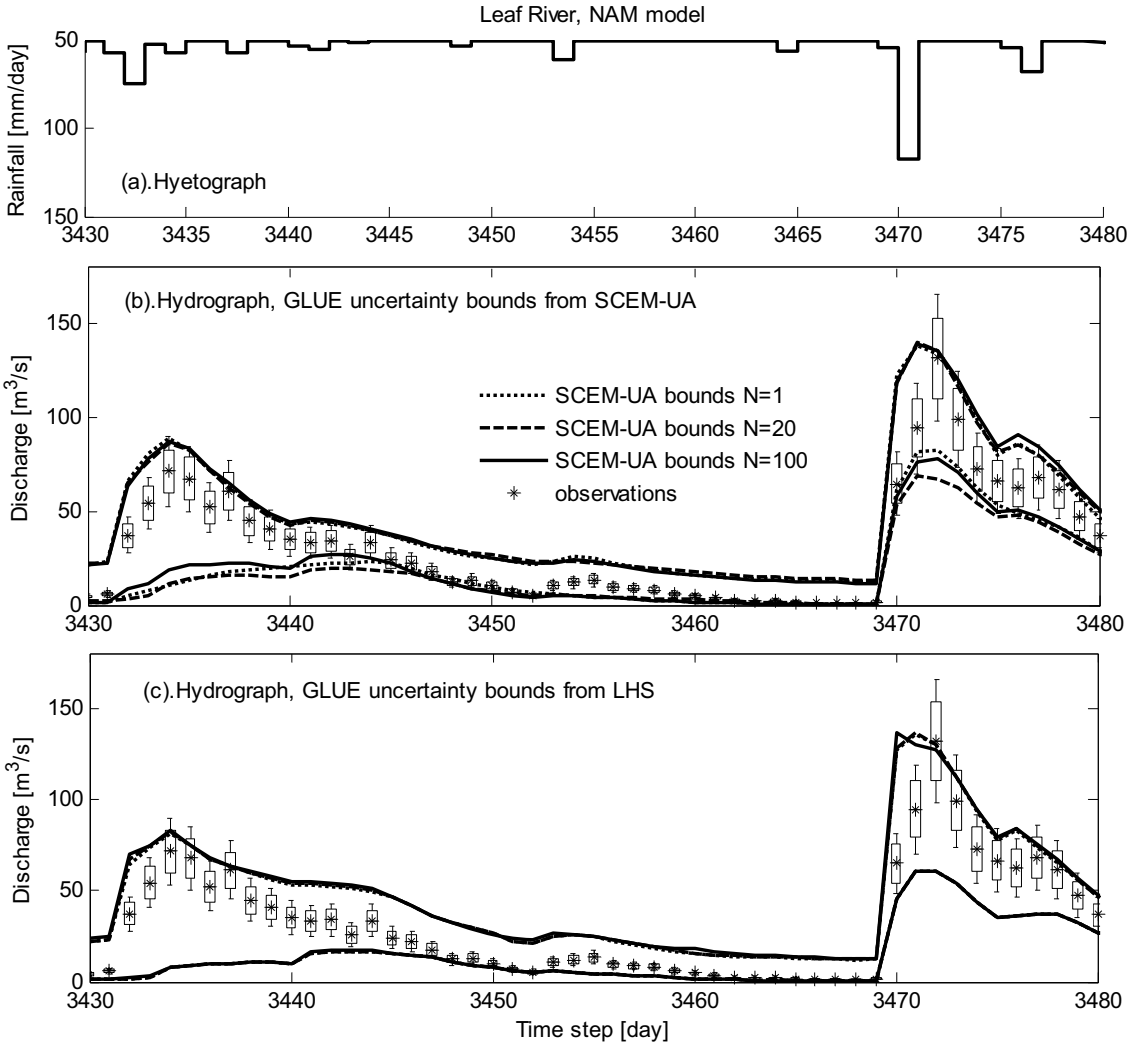

Figure 7. Leaf River watershed - NAM model: hyetograph (a) and hydrographs including the uncertainty bounds containing the 90% of the observations generated by GLUE from SCEM-UA (b) and LHS initial samples (c). The error bars in these plots represent the error properties of the streamflow data: the boxes correspond to the 5th and 95th percentiles of the error distribution, while the vertical lines extend up to the 0.5th and 99.5th percentiles.

First, note that the LHS and SCEM-UA derived posterior PDFs are qualitatively similar for the NAM model, but different for the SAC-SMA model. For models of higher dimensionality, random sampling does not provide a sufficiently large sample of solutions within the HPD region of the parameter space. Second, with respect to the parameter *N*, the posterior PDFs become narrower and peakier with increasing *N*-values. But even for increasing *N*-value the LHS derived posterior PDFs remain multi-modal, while the SCEM-UA derived histograms become Gaussian-like with a single

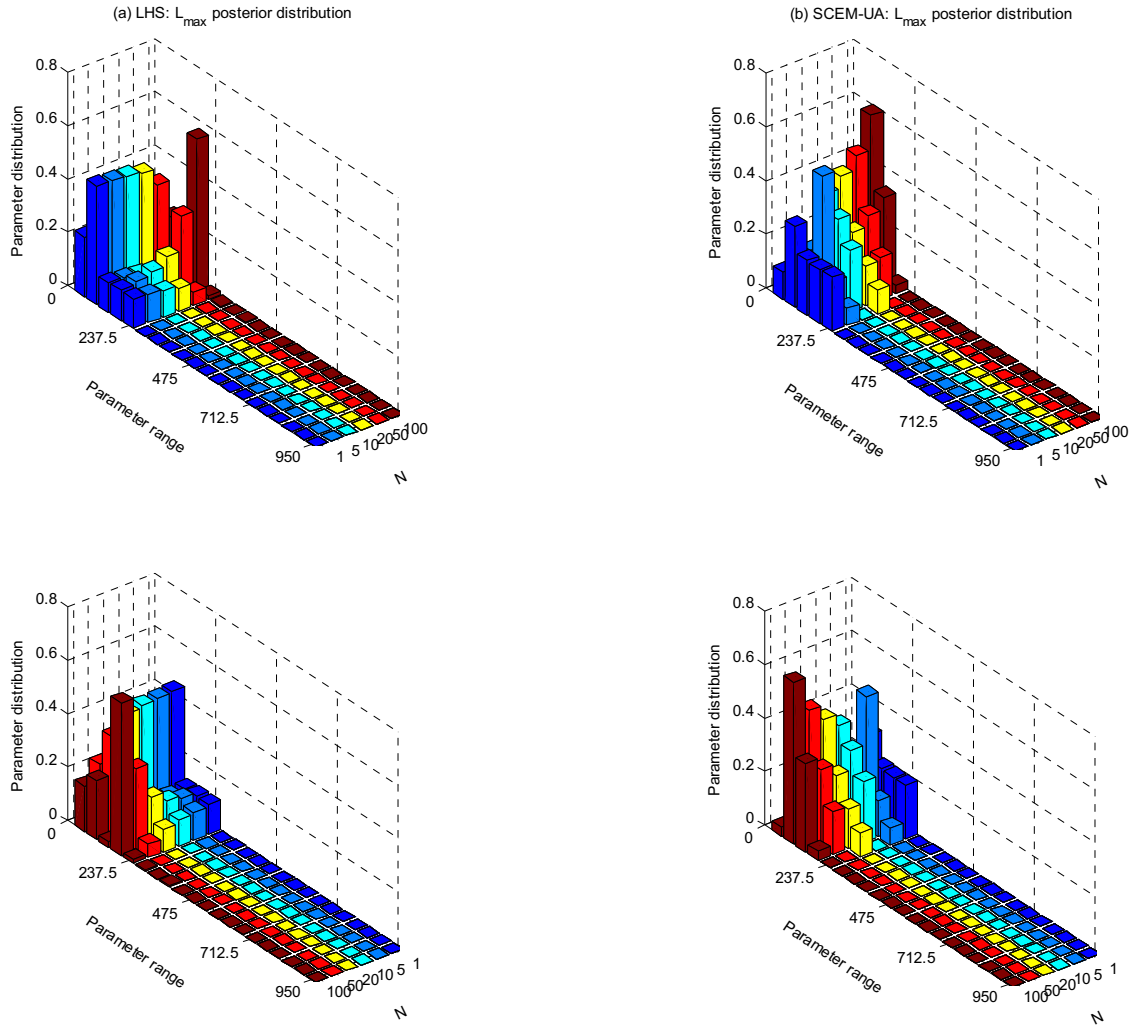

Figure 8. Posterior distribution of parameter L_{max} for Tryggevælde watershed - NAM model obtained from SCEM-UA (a) and LHS dataset (b). Real value: L_{max} = 121.1.

well-defined mode (the desired result). Finally, note that the mode of the LHS and SCEM-UA derived posterior PDFs are different, with the SCEM-UA result converging to the true value of the parameter used to generate the synthetic data, but the LHS-derived result deviating from the true value. This outcome is also reflected in the correlation coefficients between the true parameter sets used to generate the synthetic data sets, and the modal values of the posterior distributions derived with the LHS and SCEM-UA methods. For example, for the NAM and SAC-SMA models, the correlation between the true parameter sets and mode of the posterior PDF is 0.12 and 0.44 respectively for the LHS method, and 0.88 and 0.66 for the SCEM-UA sampling. This finding is consistent with our previous results.

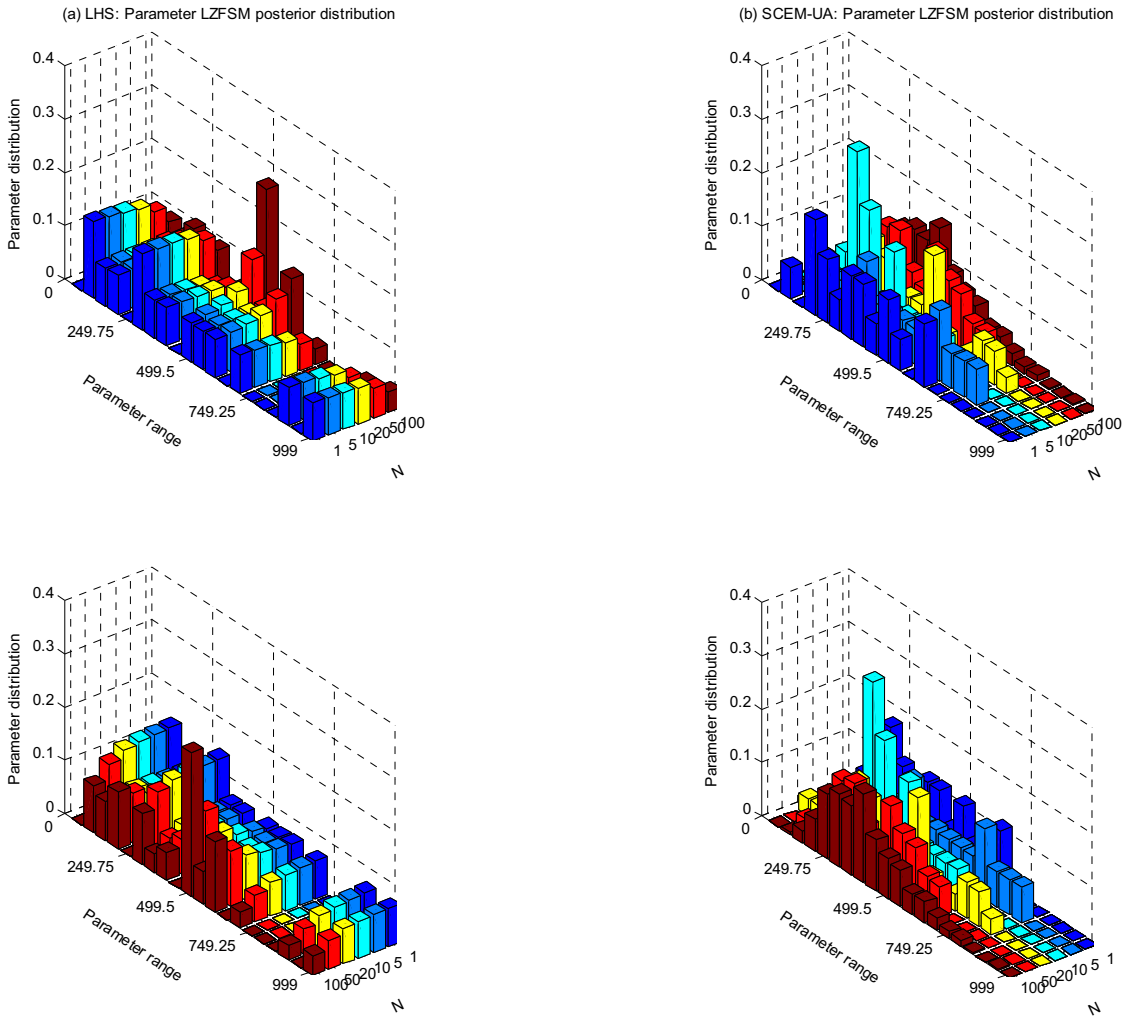

Figure 9. Posterior distribution of parameter LZFSM for Tryggevælde watershed - Sacramento model: obtained from SCEM-UA (a) and LHS dataset (b). Real value: LZFSM = 438.85.

As illustration, Figure 10 presents correlation plots between the parameters in the HYMOD model using synthetic streamflow data for the Tryggevælde watershed. These plots correspond to the GLUE-derived posterior PDF using the SCEM-UA derived initial sample for N = 100. Most plots show very low correlations, with the exception of the $\{C_{max}, b_{exp}\}$ panel, which exhibits a linear dependency, with correlation coefficient of about 0.75. This correlation plot is consistent with previous results presented in *Vrugt et al.* (2003b). Correlations between parameters in other models were typically low, but increase with increasing N-value for the SCEM-UA sampling.

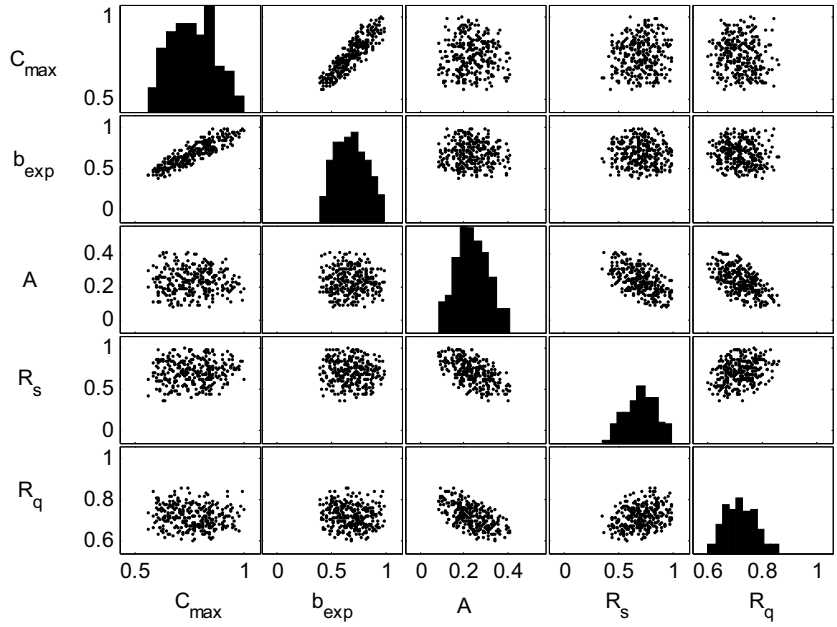

Figure 10. HYMOD model - Tryggevælde river: correlation plots of normalized parameters from posterior distributions obtained from SCEM-UA sample with likelihood function exponent N=100. Diagonal: histograms of parameter distribution.

4.2. Measured data sets

4.2.1. Median GLUE prediction

When measured streamflow observations are used, the presence of model error and forcing input error adds additional uncertainty into the modeling process. The main effects of these errors become apparent when deriving uncertainty bounds that contain a prescribed percentage of the streamflow observations (90% in this study). A much larger number of solutions need to be retained for real applications, compared to the synthetic data cases previously discussed. This is true regardless whether the LHS or SCEM-UA method is used for sampling of the prior distribution, and reflects an inability of the GLUE method to properly treat input and model structural error. Table 4 summarizes these results for N = 1, and lists the percentage of observations included within the confidence bounds and the associated number of retained solutions.

For increasing N values, the narrowing down of the bounds causes depletion of the coverage of the observations by the uncertainty intervals. Table 5 compares likelihood values of the median deterministic GLUE forecast between the LHS and SCEM-UA sampling for different values of N for the Tryggevælde catchment. As in the synthetic data experiment, the predictive capability of the median GLUE forecast is generally higher when sampling the prior distribution with the SCEM-UA algorithm than when using LHS to derive the initial sample. Also note that the relative differences

in likelihood values between the methods become larger with increasing values of *N*. As mentioned earlier, the reason for the latter tendency is explained by the better performance of the SCEM-UA method in sampling the HPD region of the parameter space, when using a peakier probability distribution. Finally, note that when explicitly dealing with model and input errors, the likelihood values of the median deterministic GLUE forecast are significantly lower than for the synthetic experiment. Similar tendencies are found for the Leaf River dataset.

The dependency of the likelihood value of the GLUE-derived median estimate of the hydrograph on the number of retained solutions shows similar patterns as previously found and discussed in our synthetic experiment. A similar trade-off between the predictive quality of the median GLUE estimate of the runoff, and the number of retained solutions is also visible when analyzing measured streamflow data. Furthermore, the GLUE-derived median estimate of the hydrograph appears less affected by the number of retained solutions when deriving the initial sample with the SCEM-UA algorithm.

Table 4. Percentage of observations contained within the GLUE uncertainty intervals and number of retained solutions (in parentheses). Results correspond to the Tryggevælde and Leaf River data sets using the LHS and SCEM-UA methods (likelihood exponent *N*=1).

	Tryggevælde		*Leaf River*	
Model	*SCEM-UA*	*LHS*	*SCEM-UA*	*LHS*
HYMOD	71.1 (2596)	74.0 (2800)	84.6 (2032)	87.9 (2000)
NAM	86.3 (2143)	87.7 (2600)	88.8 (2507)	90.8 (2200)
SAC-SMA	78.2 (5148)	77.8 (2800)	87.7 (1822)	89.6 (1800)

Table 5. Likelihood of the best runoff simulation from the initial sample generated with the LHS and SCEM-UA algorithm: Tryggevælde watershed – measured data set.

	SCEM-UA			*LHS*		
N	*HYMOD*	*NAM*	*SAC-SMA*	*HYMOD*	*NAM*	*SAC-SMA*
1	0.7024	0.7138	0.7170	0.7025	0.7196	0.7175
5	0.1712	0.1944	0.1919	0.1711	0.1929	0.1901
10	0.0298	0.0379	0.0369	0.0293	0.0372	0.0361
20	0.00088	0.00159	0.00134	0.00086	0.00139	0.00131
50	2.326E-08	7.169E-08	8.341E-08	2.149E-08	7.152E-08	6.163E-08
100	5.810E-16	1.310E-14	7.164E-15	4.618E-16	5.115E-15	3.799E-15

4.2.2. Prediction Uncertainty Bounds

As previously mentioned, the presence of input and model structural error reduces the coverage of the observations by the uncertainty intervals, thus making it more difficult to produce statistically meaningful predictions. However, Table 4 shows that percentages of observations close to the 80% can be included within the bounds, if

a very large number of solutions are retained. Figures 11 and 12 show the percentage of solutions included within the uncertainty bounds, and the width of these bounds, respectively, as functions of the number of retained solutions. Given a pre-specified number of retained solutions, the GLUE-derived uncertainty bounds are generally smaller for the SCEM-UA algorithm than for LHS. For the SCEM-UA-derived sample, the average distance to the optimal model is small, resulting in relatively small uncertainty bounds. In contrast, the inability of the LHS method to adequately sample the HPD region of the parameter space results in a rapid increase in average width of the uncertainty bounds with increasing number of retained solutions (Figure 12).

In addition, Note that the slopes of the curves decrease with increasing *N* values. While retaining more solutions will extend the extreme tails of the GLUE CDF streamflow output distribution, it hardly affects the size of the 95% uncertainty bounds, as most of the probability mass is located within the desired confidence interval. Furthermore, smaller values of *N* result in larger uncertainty bounds, because the likelihood function causes the probability mass to be spread out over a large part of the parameter space, resulting in a wide variety of simulations that are considered to be behavioral.

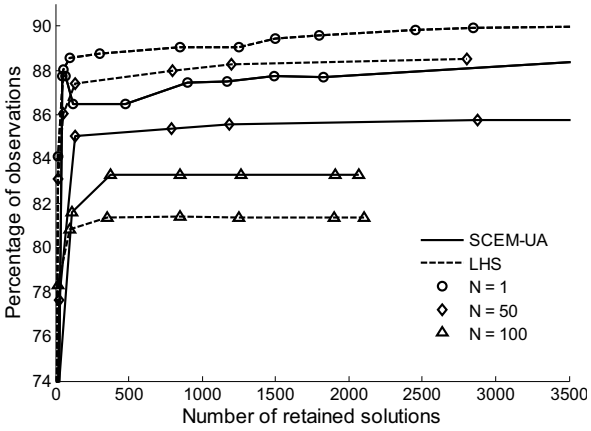

Figure 11. Leaf River watershed – SAC-SMA model: percentage of solutions included within the uncertainty bounds as a function of the number of retained solutions.

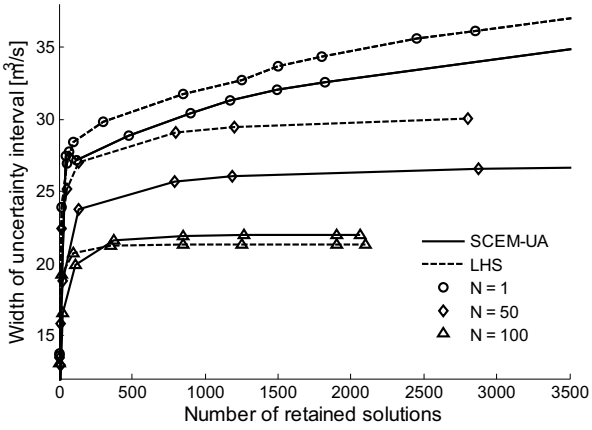

Figure 12. Leaf River watershed – SAC-SMA model: width of the uncertainty bounds as a function of the number of retained solutions.

4.2.3. Parameter Uncertainty and Correlation

As mentioned earlier, fewer observations are covered by the uncertainty intervals when the measured streamflow data is used. Moreover, there is a decrease in the coverage for increasing value of *N*. While the uncertainty intervals generated with the LHS and SCEM-UA samples include between 75-90% of the observations for *N* = 1, these percentages, for all the models and data sets considered, range between 83%

and 40% for N = 100. Thus, it is not always possible to generate uncertainty intervals with a reliable statistical meaning. Nevertheless, the analysis of the posterior distributions of the available parameters fully confirms the results for the artificially generated data sets. This is also exemplified in Figure 13, a plot of the GLUE-derived posterior PDFs obtained from the LHS and SCEM-UA samples for the parameter *LZFSM* of SAC-SMA model applied to the Leaf River watershed. First, note that the posterior distributions get narrower and peakier for increasing N value. Moreover, while the PDFs inferred from the SCEM-UA sample show a well-defined peak, those from the LHS dataset generally exhibit multimodality. This feature, caused by the peculiarities of the initial random sample, reduces the reliability of the parameter estimates. Also, similar to what was found for the synthetic streamflow data, the difference between the PDFs obtained from the LHS and the SCEM-UA initial samples increases with increasing model complexity.

Finally, no relevant correlations were found among the parameters of the various models, with the exception of the parameters C_{max} and b_{exp} of the HYMOD model, which have a correlation coefficient of about 0.78 when the model is applied to the Tryggevælde watershed. In this case, similarly as before, this correlation is found within all the LHS datasets as well as from the SCEM-UA sample, but, in this last case, only when N = 100.

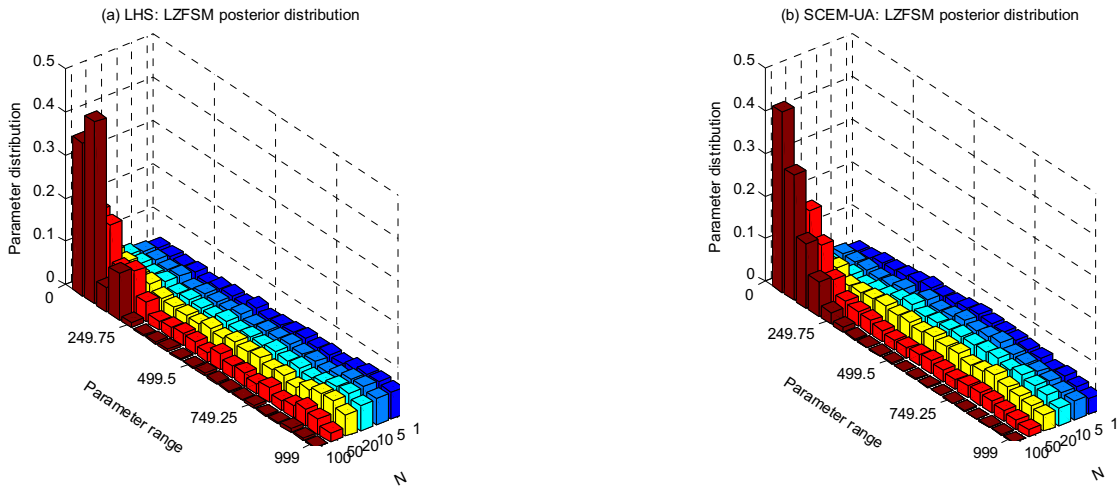

Figure 13. Posterior distribution of parameter LZFSM for Leaf River and Sacramento model obtained from LHS (a) and SCEM-UA dataset (b). The number of observations contained within the uncertainty interval ranges from 82% to 90% in this case.

5. SUMMARY AND CONCLUSIONS

This paper demonstrates the potential of improving the GLUE method by employing the Shuffled Complex Evolution Metropolis (SCEM-UA) global optimization algorithm for sampling the prior distribution of the model parameters. The SCEM-UA algorithm is an adaptive Markov Chain Monte Carlo (MC^2) sampler that periodically updates the size and direction of the proposal distribution. This feature enables it to visit solutions in the HPD region of the parameter space with higher frequency than a random sampling scheme. Through a comparison of the GLUE results using LHS and SCEM-UA sampling for creating the initial sample, we demonstrated the following conclusions:

1. The combined SCEM-UA – GLUE method provides better predictions of the model output than a classical GLUE procedure based on random sampling. This improvement is obtained for the median GLUE estimates and best parameter estimates from the initial sample. At the same time, the Markov Chain sampler yields a reduction in the uncertainty of the output estimate, providing narrower confidence intervals than those obtained from the LHS dataset. The differences in the results from the two sampling methods increase with the model complexity and with *N*, the exponent of the likelihood function.
2. When using SCEM-UA sampling, the GLUE-derived median output estimate and associated prediction uncertainty bounds are less affected by the number of retained solutions in the analysis. The SCEM-UA-derived initial sample contains numerous solutions in the HPD region of the parameter space, so that the average distance of the various parameter combinations to the optimal model is small. This results in uncertainty bounds that are less dependent on the number of retained solutions. In contrast, the inability of random sampling to closely sample the HPD region of the parameter space results in a widening of the uncertainty bounds when a larger number of solutions are retained.
3. The SCEM-UA algorithm will likely be able to find the global optimum in the parameter space. In contrast, random sampling can require an unmanageably large number of model simulations to attain a statistically sufficient number of behavioral parameter sets. The LHS scheme, used frequently in the GLUE method and implemented in this paper, finds solutions well removed from the best attainable model. Therefore, the GLUE method with SCEM-UA sampling should be superior for making valid conclusions about parameter identifiability and equifinality.
4. The efficiency of the SCEM-UA algorithm is controlled by the shape of the likelihood function used in the GLUE analysis. Likelihood functions for which significant probability extends over a large range of the prior parameter space will adversely affect the search and explorative capabilities of the SCEM-UA algorithm. The

sampler will have difficulty converging under these circumstances. On the contrary, in situations in which the likelihood function is peaked and significant probability mass is associated with a small interior region of the parameter space, the SCEM-UA method will significantly improve the quality of the GLUE results. This conclusion has been demonstrated in this paper through comparisons of results for different values of the parameter *N*.

5. The results presented in this paper, along with additional analyses not presented, show strong consistency between results derived for synthetic and measured data sets, for models of two watersheds with significantly different hydrologic characteristics. This result demonstrates that our findings on the usefulness of our revised GLUE method are quite general.

6. Our approach for discriminating between behavioral and non-behavioral solutions using information from the coverage of the uncertainty bounds results in statistically meaningful uncertainty intervals. This approach therefore provides an adequate and satisfactory solution to the often criticized subjectivity involved in the choice of an appropriate cutoff value on the retained solutions (or on the likelihood function value). Nevertheless, even with the implementation of a more objective approach to separate between behavioral and non-behavioral solutions, a strong trade-off appears between the accuracy of the median GLUE forecast and precision of the uncertainty bounds. It is shown that the best output estimates are obtained when a relatively small number of solutions are retained, whereas a large number of solutions must be retained to generate uncertainty bounds with a sufficient coverage of the observations.

7. Adaptive MC^2 sampling of the prior parameter distribution improves the efficiency and robustness of the GLUE methodology. This result is especially true for complex environmental models with a relatively large number of model parameters, and likelihood functions that assign significant probability to a relatively small region interior to the plausible model or parameter space.

ACKNOWLWDGEMENTS

The second author is supported by the LANL Director's Funded Postdoctoral program.

REFERENCES

Aronica, G., P. D. Bates, and M. S. Horritt (2002), Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE, *Hydrol. Processes*, *16*, 2001–2016.

Beven, K. (2006), A manifesto for the equifinality thesis, *The Model Parameter Estimation Experiment - MOPEX*, *J. Hydrol.*, *320*(1), 18–36.

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

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

Boyle, D. P., H. V. Gupta, and S. Sorooshian (2000), Toward improved calibration of hydrological models: Combining the strengths of manual and automatic methods, *Water Resour. Res.*, *36*, 3663–3674.

Brazier, R E., K. J. Beven, S. G. Anthony, and J. S. Rowan (2001), Implications of model uncertainty for the mapping of hillslope-scale soil erosion predictions, *Earth Surf. Processes Landforms*, *26*, 1333–1352.

Burnash, R. J. C. (1995), The NWS river forecast system-catchment modeling. In Singh, V. J., (Ed.), *Computer Models of Watershed Hydrology*, Water Resources Publication, Highlands ranch, Colorado, 311–366.

Burnash, R. J. C., R. L. Ferral, and R. A. McGuire (1973), A Generalized Streamflow Simulation System: Conceptual Modeling for Digital Computers, Joint Federal-State River Forecast Center, Sacramento, CA.

Christensen, S. (2004), A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals, *Nordic Hydrol.*, *35*(1), 45–59.

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

Franks, S. W., K. J. Beven, P. F. Quinn, and I. R. Wright (1997), On the sensitivity of soil–vegetation–atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration, *Agric. For. Meteorol.*, *86*, 63–75.

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

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

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

Hankin, B. G., R. Hardy, H. Kettle, and K. J. Beven (2001), Using CFD in a GLUE framework to model the flow and dispersion characteristics of a natural fluvial dead zone, *Earth Surf. Processes Landforms*, *26*(6), 667–687.

Hansson, K., and C. Lundin (2006), Equifinality and sensitivity in freezing and thawing simulations of laboratory and in situ data, *Cold Regions Science and Technology*, *44*, 20–37.

Havnø, K., M. N. Madsen, and J. Dørge (1995), MIKE 11 – a generalized river modelling package, *Computer Models of Watershed Hydrology* (ed. Singh, V.P.), Water Resources Publications, Colorado, 733–782.

Hornberger, G. M., and R. C. Spear (1981), An approach to the preliminary analysis of environmental systems, *J. Environ. Manag.*, *12*, 7–18.

Jensen, J. B. (2003), Parameter and Uncertainty Estimation in Groundwater Modelling., PhD thesis, Department of Civil Engineering, Aalborg University, Series Paper No. 23.

Keesman, K. J. (1990), Set theoretic parameter estimation using random scanning and principal component analysis, *Math. Comput. Simul.*, *32*, 535–543.

Khu, S. T., and H. Madsen (2005), Multiobjective calibration with Pareto preference ordering: An application to rainfall-runoff model calibration, *Water Resour. Res.*, *41*(3), 1–14.

Klepper, O., H. Scholten, and J. P. G. van de Kamer (1991), Prediction uncertainty in an ecological model of the Oosterschelde Estuary, *J. Forecasting*, *10*, 191–209.

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

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

Lorup, J. K., J. C. Refsgaard, and D. Mazvimavi (1998), Assessing the effect of land use change on catchment runoff by combined use of statistical tests and hydrological modelling: case studies from Zimbabwe, *J. Hydrol.*, *205*(3), 147–163.

Madsen, H. (2003), Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives, *Adv. Water Resour.*, *26,* 205–216.

Madsen, H. (2000), Automatic calibration of a conceptual rainfall-runoff model using multiple objectives, *J. Hydrol.*, *235*(3-4), 276–288.

Madsen, H., D. Rosbjerg, J. Damgård, and F. S. Hansen (2003), Data assimilation in the MIKE 11 Flood Forecasting system using Kalman filtering, *Water Resources Systems - Hydrological Risk, Management and Development* (Proceedings of symposium HS02b held during IUGG2003 at Sapporo, July 2003). IAHS Publ. no. 281, 75–81.

McKay, M. D., W. J. Conover, and R. J. Beckman (1979), A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, *21*, 239–245.

McMichael, C. E., A. S. Hope, and H. A. Loaiciga (2006), Distributed hydrological modeling in California semi-arid shrublands: MIKE SHE model calibration and uncertainty estimation, *J. Hydrol.*, *317*, 307–324.

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

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.

Moradkhani, H., K.-L. Hsu, H. Gupta, and S. Sorooshian (2005), Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter, *Water Resour. Res.*, *41*(5), 1–17.

Muleta, M. K., and J. Nicklow (2005), Sensitivity and uncertainty analysys coupled with automatic calibration for a distributed watershed model, *J. Hydrol.*, *306*, 127–145.

Nielsen, S. A. and E. Hansen (1973), Numerical simulation of the rainfall runoff processes on a daily basis, *Nordic Hydrol.*, *4*, 171–190.

Pappenberger, F., K. Beven, M. Horritt, and S. Blazkova (2005), Uncertainty in the calibration of effective roughness parameters in HEC-RAS using inundation and downstream level observations, *J. Hydrol.*, *302*(1-4), 46–69.

Peck, E. L. (1976), Catchment modeling and initial parameter estimation for the National Weather Service river forecast system, Tech. Memo. NWS Hydro-31, Natl. Oceanic and Atmos. Admin., Silver Spring, Md..

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowsk (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Monthly Weather Review*, *133*(5), 1155–1174.

Romanowicz, R. J., K. J. Beven and J. Tawn (1996), Bayesian calibration of flood inundation models. In: M.G. Anderson and D.E. Walling (eds.), *Floodplain Processes*, Wiley, Chichester, pp. 333–360.

Romanowicz, R., K. J. Beven, and J. Tawn (1994), Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach, in *Statistics for the Environment*, vol. 2, Water Related Issues, edited by V. Barnett and K. F. Turkman, pp. 297–317, John Wiley, Hoboken, N. J.

Strom, B., K. H. Jensen, and J. C. Refsgaard (1988), Estimation of catchment rainfall uncertainty and its influence on runoff prediction, *Nordic Hydrol.*, *19*(2), 77–88.

Tadesse, A., and E. N. Anagnostou (2005), A statistical approach to ground radar-rainfall estimation, *Journal of Atmospheric and Oceanic Technology*, *22*(11), 1055–1071.

Thiemann, M., M. Trosset, H. Gupta, and S. Sorooshian (2001), Bayesian recursive parameter estimation for hydrological models, *Water Resour. Res.*, *37*(10), 2521–2535.

van Straten, G., and K. J. Keesman (1991), Uncertainty propagation and speculation in projective forecasts of environmental change: A lake eutrophication example, *J. Forecasting*, *10*, 163–190.

Vrugt, J. A., C. G. H. Diks, H. V. Gupta, W. Bouten, and J. M. Verstraten (2005), Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation, *Water Resour. Res.*, *41*(1), 1–17.

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

Vrugt, J. A., H. V. Gupta, L. A. Bastidas, W. Bouten, and S. Sorooshian (2003b), Effective and efficient algorithm for multi-objective optimization of hydrologic models, *Water Resour. Res.*, *39*(8), 1214, doi:10.1029/2002WR001746.

Vrugt, J. A., A. H. Weerts, and W. Bouten (2001), Information content of data for identifying soil hydraulic parameters from outflow experiments, *Soil Sci. Soc. Am. J.*, *65*, 19–27.

Vrugt, J. A., and B. A. Robinson (2007), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, *43*, W01411, doi:10.1029/2005WR004838.

Wang, X., X. He, J. R. Williams, R. C. Izaurralde, and J. D. Atwood (2005), Sensitivity and uncertainty analyses of crop yields and soil organic carbon simulated with EPIC, *Transactions of the American Society of Agricultural Engineers*, *48*(3), 1041–1054.

Yapo, P. O., H. V. Gupta, and S. Sorooshian (1998), Multi-objective optimization for hydrologic models, *J. Hydrol*., *204*, 83–97.