Available online at www.sciencedirect.com

ScienceDirect

Advances in Water Resources 31 (2008) 630–648

Advances in Water Resources

www.elsevier.com/locate/advwatres

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

Roberta-Serena Blasone[a], Jasper A. Vrugt[b,*], Henrik Madsen[c], Dan Rosbjerg[a], Bruce A. Robinson[d], George A. Zyvoloski[e]

[a] *Department of Environmental Engineering, Technical University of Denmark, Kongens Lyngby, Denmark*
[b] *Center for Nonlinear Studies, Los Alamos National Laboratory, Los Alamos, NM, USA*
[c] *DHI Water, Environment and Health, Hørsholm, Denmark*
[d] *Civilian Nuclear Program Office (SPO-CNP), Los Alamos National Laboratory, Los Alamos, NM, USA*
[e] *Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, NM, USA*

Received 23 May 2007; received in revised form 10 December 2007; accepted 17 December 2007
Available online 11 February 2008

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 MC based sampling strategy of the prior parameter space typically utilized in GLUE is not particularly efficient in finding behavioral simulations. This becomes especially problematic for high-dimensional parameter estimation problems, and in the case of complex simulation models that require significant computational time to run and produce the desired output. In this paper we improve the computational efficiency of GLUE by sampling the prior parameter space using an adaptive Markov Chain Monte Carlo scheme (the Shuffled Complex Evolution Metropolis (SCEM-UA) algorithm). Moreover, we propose an alternative strategy to determine the value of the cutoff threshold based on the appropriate coverage of the resulting 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 catchments with different hydrologic regimes.

Keywords: Hydrologic modeling; Uncertainty estimation; Generalized likelihood uncertainty estimation (GLUE); Markov Chain Monte Carlo; Rainfall-runoff models; Calibration

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.

[*] Corresponding author. Tel.: +1 505 667 0404; fax: +1 505 665 8737.
E-mail address: vrugt@lanl.gov (J.A. Vrugt).

0309-1708/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.advwatres.2007.12.003

Recent years have seen an explosion of methods to derive meaningful uncertainty bounds on our model predictions. Methods to represent model parameter, state and prediction uncertainty include classical Bayesian [30,54,57], pseudo-Bayesian [3,15], set-theoretic [26,28, 52,56], multiple criteria [19,62,7,34,35,58], sequential data assimilation [36,59,42] and multi-model averaging methods [17,1,60]. 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 [3], inspired by the Hornberger and Spear [23] 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 (MC) 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 pseudo-Bayesian MC framework. The outputs of the GLUE procedure are parameter distributions conditioned on the available observational data and associated uncertainty bounds.

Since its introduction in 1992, the GLUE framework has found widespread application for uncertainty assessment in environmental modeling, including rainfall-runoff modeling [3,15,31], soil erosion modeling [8], modeling of tracer dispersion in a river reach [20], groundwater modeling and well capture zone delineation [13,24], unsaturated zone modeling [40], flood inundation modeling [50,2], land-surface–atmosphere interactions [14], soil freezing and thawing modeling [21], crop yields and soil organic carbon modeling [61], ground radar-rainfall estimation [53] and distributed hydrological modeling [39,44]. 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 [54,38,41,11]. Moreover, in most GLUE applications a rather simplistic MC sampling scheme is used to sample from the prior parameter distributions (with some notable exceptions that will be discussed later) and to find a well-distributed set of behavioral models and their associated predictive simulation uncertainty. While this approach may be adequate for relative simple low-dimensional sampling problems, it is unlikely to result in stable and consistent estimates of the set of behavioral models (and thus parameter distributions) for relatively high-dimensional and complex estimation problems. To compensate for this drawback, the LHS method typically requires many thousands of model simulations [39,43,41,46]. However, various contributions to the hydrologic literature have demonstrated that even at this extreme only very few behavioral models are found. In those situations, one should be particularly careful not to infer erroneous conclusions about parameter identifiability and equifinality [7,57].

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 found in the past to improve their search efficiency and converge to the posterior PDF of the parameters.

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. A few papers do discuss the use of MC^2 sampling in GLUE for generating the initial sample (e.g. [37]), but this approach has not become common practice. The concept is to construct the initial sample using the Shuffled Complex Evolution Metropolis (SCEM-UA) global optimization algorithm and derive the associated model output estimates (as the median of the distribution) and 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 more efficiently, thereby improving the computational efficiency and robustness of the so-derived uncertainty bounds.

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 sometimes 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, based on the fact that they are created by using the retained solutions covering 90% of the posterior probability. It should be pointed out that these so-derived GLUE uncertainty bounds are not confidence bounds in a statistical sense, i.e. they are not expected to include a given percentage of the observations.

While the GLUE method has found widespread implementation for predictive uncertainty analysis in environmental modeling, the method has several drawbacks that have been well pointed out and discussed in the literature [55,38,41,11]. Perhaps most importantly, the GLUE derived parameter distributions and uncertainty bounds are entirely subjective and have no clear statistical meaning. This is because the method uses an informal likelihood to extract information from the observational data and implements a subjective cutoff threshold to separate behavioral from non-behavioral models. So, strictly speaking, the method is incoherent and inconsistent from a statistical point of view, although some easy adjustments to GLUE can be made so that the method adheres to formal Bayesian theory [5].

Although GLUE can be criticized for not being properly Bayesian, it is not particularly easy to develop a likelihood measure that properly accounts for input, output, parameter and model structural error within a single performance measure [32,29]. This has been the subject of much research in many different fields of study, but no universal applicable likelihood measure has yet been developed (and perhaps might not exist!) that properly extracts information from observational data in the presence of different sources of uncertainty. Approaches that do implement formal likelihood measures for inference such as sequential data assimilation methods [59,42], multi-model averaging approaches [1,60] and recent extensions to BATEA (Bayesian total error analysis) [25] have their own weaknesses. For instance, they essentially rely on Gaussian or Gamma probability distributions to characterize and propagate various sources of errors. These distributions seem incomplete and inappropriate, particularly for real-world applications.

In the absence of a formal likelihood measure that incorporates all sources of uncertainty, or error models that appropriately characterize input, output, parameter and model structural error within a Bayesian framework (and within the context of streamflow forecasting), we here focus instead on improving the computational efficiency of GLUE to increase applicability of the method to relative high-dimensional parameter estimation problems and complex simulation models that require significant computational time to run and produce the desired output. One might argue that the main problem with GLUE is not to find an efficient sampling strategy and that it would be more productive if we would focus our research efforts on the development of statistically proper likelihood measures. Nevertheless, we believe that improving computational efficiency is a necessary developmental step to further increase applicability of uncertainty estimation methods such as GLUE to complex inference problems [43]. Hence, irrespective of the likelihood measure used, successful application of GLUE essentially relies on the identification of a well-distributed set of behavioral solutions in the parameter space that appropriately captures and reports uncertainty.

2.2. Parameter sampling strategy

To sample the prior parameter distribution, practitioners of the GLUE methodology generally implement a simple random sampling, or in some cases the more efficient Latin hypercube sampling (LHS) strategy [12,44]. Random sampling methods, though relatively simple to implement, are 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 robust.

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 so that the posterior PDF is approximated with a sample of parameter sets. 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 provides a close approximation of the HPD region of the parameter space, but is significantly more efficient than traditional Metropolis–Hastings samplers [57].

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 [16] indicates convergence to a stationary posterior distribution. An extensive description and explanation of the method appears in [57] and so will not be repeated here. The SCEM-UA method was implemented using standard values for the algorithmic parameters presented in [57]. Moreover, in the absence of reliable prior information about the location of the HPD region in the parameter space, we assume a uniform prior for each of the individual parameters in all our SCEM-UA optimizations. This is a typical assumption in hydrologic modeling and frequently used in the SCEM-UA algorithm to approximate the posterior PDF of the model parameters.

The rationale for adopting this sampling strategy in the GLUE methodology rests on arguments of the generation of 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 robust 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. [4]), 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. [3,49,11,42]) 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:

$$L(\theta_i|Y) = \exp\{-N \cdot \sigma_i^2/\sigma_{\text{obs}}^2\} \qquad (1)$$

where $L(\theta_i|Y)$ is the likelihood measure for the ith model conditioned on the observations Y, σ_i^2 is the error variance for the ith model (i.e. the combination of the model and the ith 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 [15]. 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. It can assume values between 0 and 1. The closer to 1 the likelihood is, the better the simulations are, thus this quantity has to be maximized in selecting the behavioural GLUE solutions. Of course, we could have implemented a classical likelihood function with SCEM-UA so that this method operates within a formal Bayesian framework and statistically sound inferences can be made about parameter uncertainty, correlation and identifiability. However, it has been shown [57] that the use of a formal likelihood measure within the context of streamflow forecasting results in uncertainty bounds that are too small and do not appropriately capture the measured discharge data (compare the dark grey uncertainty bounds in Fig. 11 of their paper with the observed streamflow values). This is because the current generation of adaptive MC^2 samplers (such as the SCEM-UA algorithm) typically maps all the uncertainty in the modeling process onto the parameter space, effectively neglecting the influence of input and model structural errors. To avoid this overconditioning of the posterior PDF to a too small region in the parameter space, and thus implicitly handle the effect of other sources of error, we therefore implement the likelihood function in Eq. (1). 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. Note, that when using SCEM-UA with an informal likelihood measure (as is done here), the method does no longer adhere to classical statistics and like GLUE should now be referred to as pseudo-Bayesian.

In principle, there is no need to run the samples derived with the SCEM-UA method through the GLUE method, as the SCEM-UA method already results in an approximation of the posterior PDF for a given likelihood function. This distribution can be used to directly construct parameter distributions and associated uncertainty bounds. However, for a fair comparison between LHS and adaptive MC^2 sampling it is desirable to have all elements in the inference similar, so that we can directly compare the number of retained solutions and spread and sharpness of the uncertainty bounds derived with both sampling methods. We therefore postprocess both the LHS and SCEM-UA generated samples with the GLUE method. The flowchart in Fig. 1 illustrates how the different sampling procedures are incorporated in the GLUE framework.

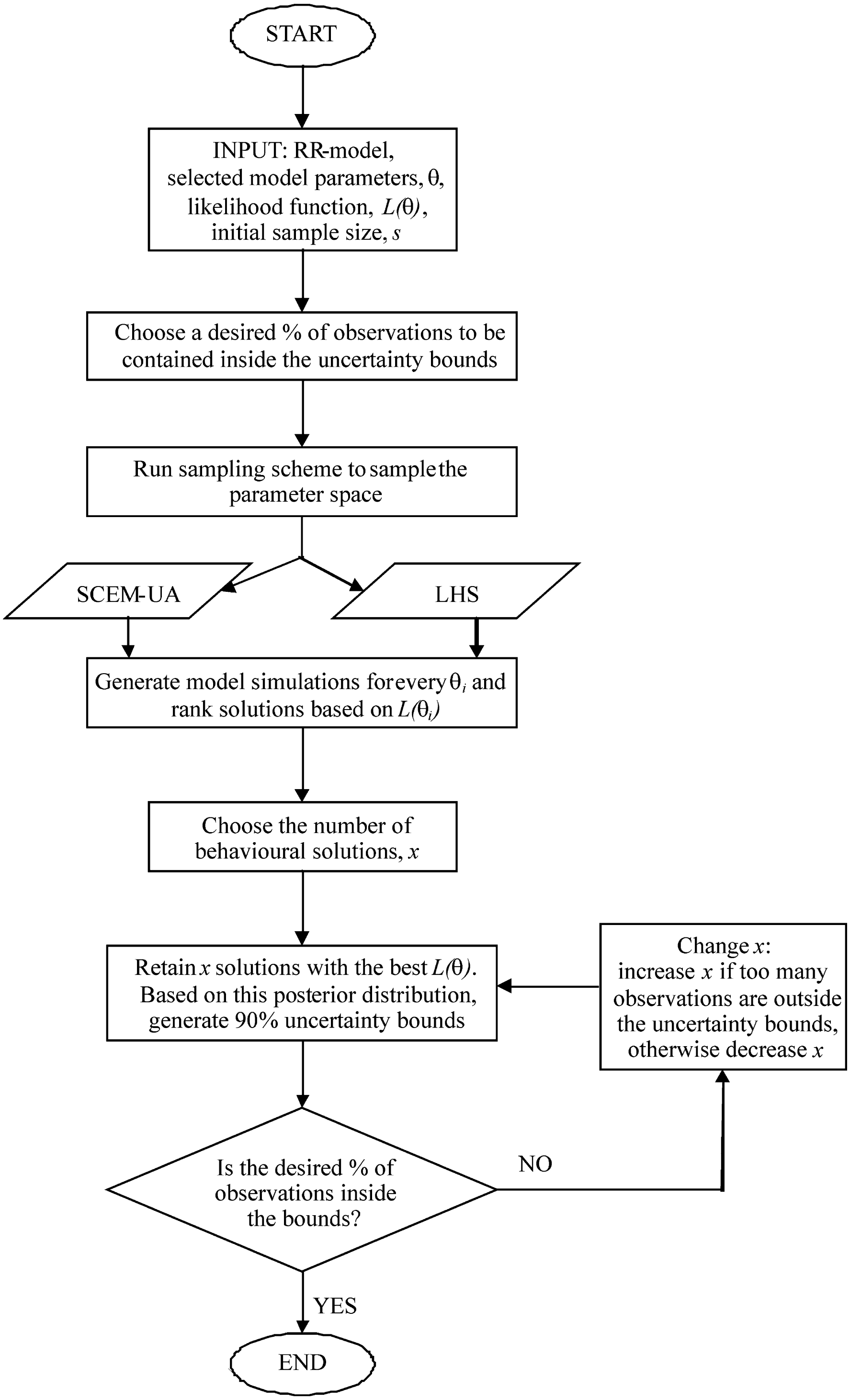
START
INPUT: RR-model, selected model parameters, θ, likelihood function, L(θ), initial sample size, s
Choose a desired % of observations to be contained inside the uncertainty bounds
Run sampling scheme to sample the parameter space
SCEM-UA
LHS
Generate model simulations for every θ_i and rank solutions based on L(θ_i)
Choose the number of behavioural solutions, x
Retain x solutions with the best L(θ). Based on this posterior distribution, generate 90% uncertainty bounds
Change x: increase x if too many observations are outside the uncertainty bounds, otherwise decrease x
Is the desired % of observations inside the bounds?
NO
YES
END

Fig. 1. GLUE flowchart with SCEM-UA and LHS sampling schemes.

2.4. Choice of the cutoff threshold for the behavioral simulations

One criticism of the GLUE methodology is that the 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 [18]. 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% uncertainty intervals. These intervals are found by a trial-and-error method in which the acceptance criterion (i.e. the number of behavioural solutions) is adjusted and the coverage is computed over a fixed calibration period.

It should be underlined that this approach, which is common in GLUE applications, estimates the total simulation error based on parameter variability only. In this way other sources of uncertainty are accounted for by the GLUE uncertainty bounds only implicitly.

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 [6], NAM [45,22] and the Sacramento Soil Moisture Accounting Model (SAC-SMA: [10,9]). 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 (HYdrologic MODel) 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 reported in [58] and [41]. The last column in Table 1 lists the prior uncertainty ranges used to generate the initial sample.

3.1.2. The NAM model

The NAM model (from the Danish: "Nedbor-Afstromnings-Model", which means precipitation-runoff-model) is a deterministic, lumped, conceptual rainfall-runoff model originally developed at the Technical University of Denmark [45,22]. It has been used in many different applications and studies [51,33,34,27]. 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 linear 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. [10] (see also [9]). It is currently used by the National Weather Service River Forecast System (NWSRFS) centers 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 three 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 recommended in [47]; 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.

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

Parameter	Unit	Range	Description
HYMOD			
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			
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	[h]	0.01–2000	Time constant for drainage of interflow
CK_{12}	[h]	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}	[h]	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			
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

3.2. 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 (0.99 m^3/s). 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. The 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 by adding a normally distributed white-noise error with standard deviation equal to 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 and robust 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) 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. When looking at Table 2, it should be noticed that increasing N will naturally result in lower likelihood functions. Therefore the likelihood function values calculated using different N are not directly comparable, unless a proper transformation is applied to get rid of the exponential effect of N. 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 posterior distribution 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 posterior distribution 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 $N = 100$). Similar results have been reported in [15].

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

N	SCEM-UA			LHS		
	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

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 [48,60]. Again, differences between the LHS and SCEM-UA algorithm increase with increasing N-value and complexity of the catchment model.

Table 3
Likelihood value of the median runoff estimate from the posterior CDF derived with the GLUE methodology: Tryggevælde watershed – synthetic data

N	SCEM-UA			LHS		
	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

The output CDF was tuned to contain 90% of the streamflow observations.

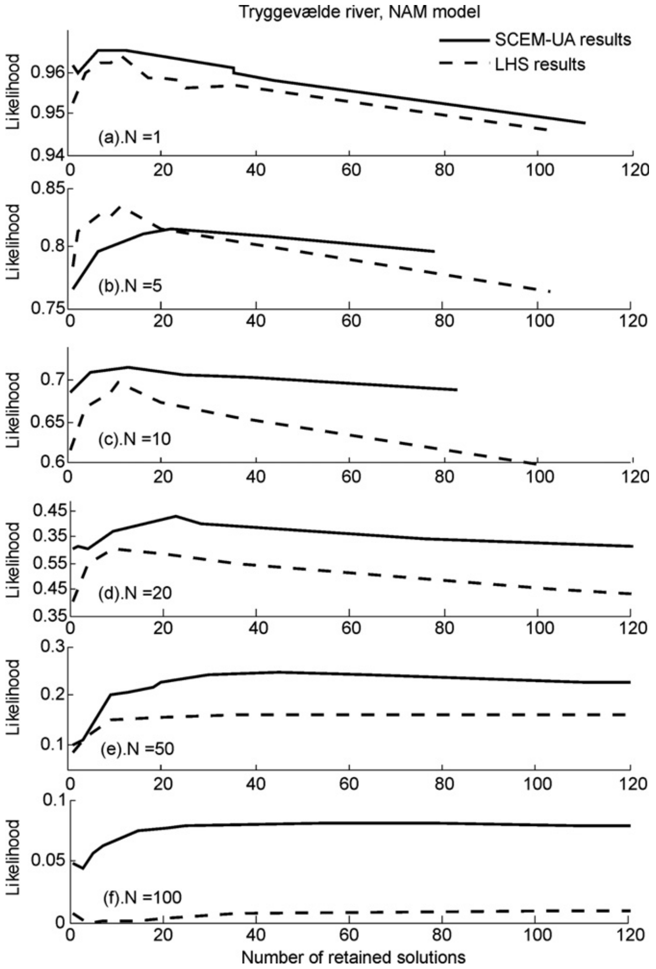

Fig. 2. Tryggevælde watershed – NAM model: likelihood of the median GLUE estimates obtained from the LHS and SCEM-UA samples versus number of retained solutions. Plots correspond to different values of the exponent of the likelihood function, N: (a) $N = 1$; (b) $N = 5$; (c) $N = 10$; (d) $N = 20$; (e) $N = 50$; (f) $N = 100$.

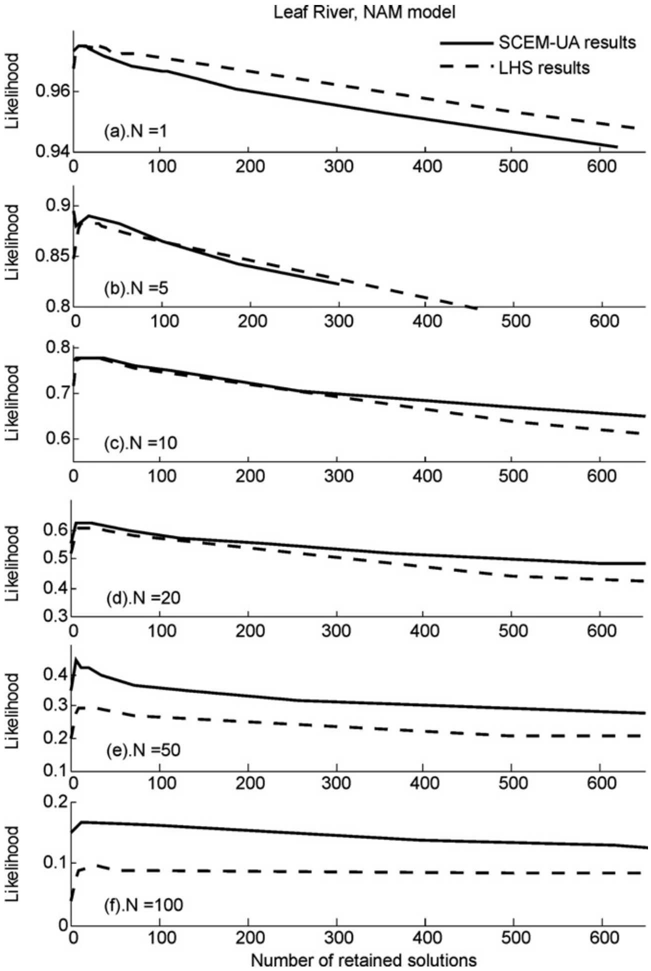

Fig. 3. Leaf River watershed – NAM model: likelihood of the median GLUE estimates obtained from the LHS and SCEM-UA samples versus number of retained solutions. Plots correspond to different values of the exponent of the likelihood function, N: (a) $N = 1$; (b) $N = 5$; (c) $N = 10$; (d) $N = 20$; (e) $N = 50$; (f) $N = 100$.

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 Figs. 2 and 3, 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.

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.

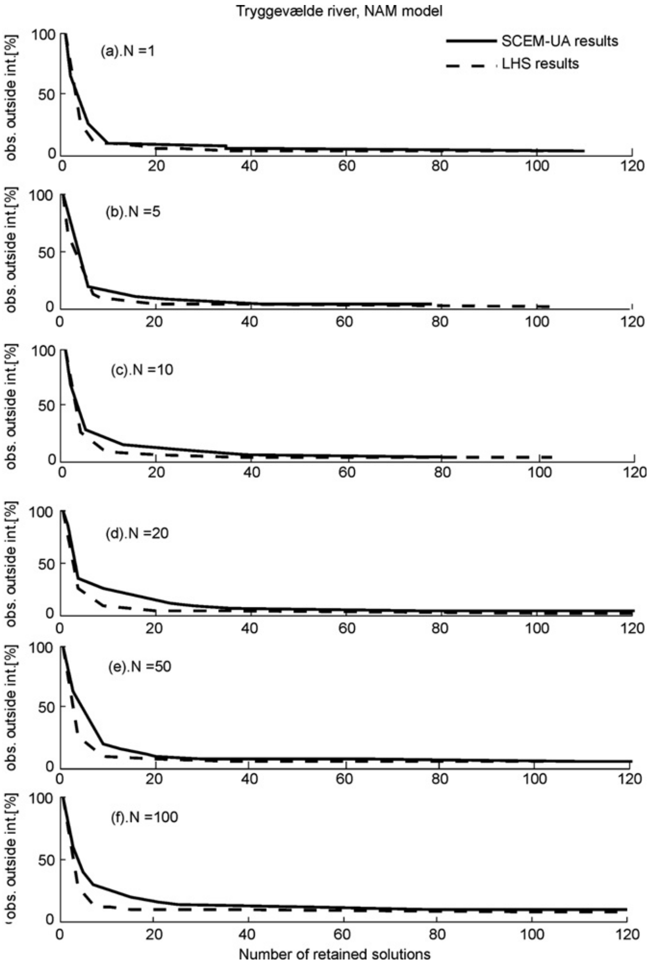

Fig. 4. Tryggevælde watershed – NAM model: percentage of runoff observations outside GLUE LHS and SCEM-UA uncertainty intervals versus number of retained solutions. Plots correspond to different values of the exponent of the likelihood function, N: (a) $N = 1$; (b) $N = 5$; (c) $N = 10$; (d) $N = 20$; (e) $N = 50$; (f) $N = 100$.

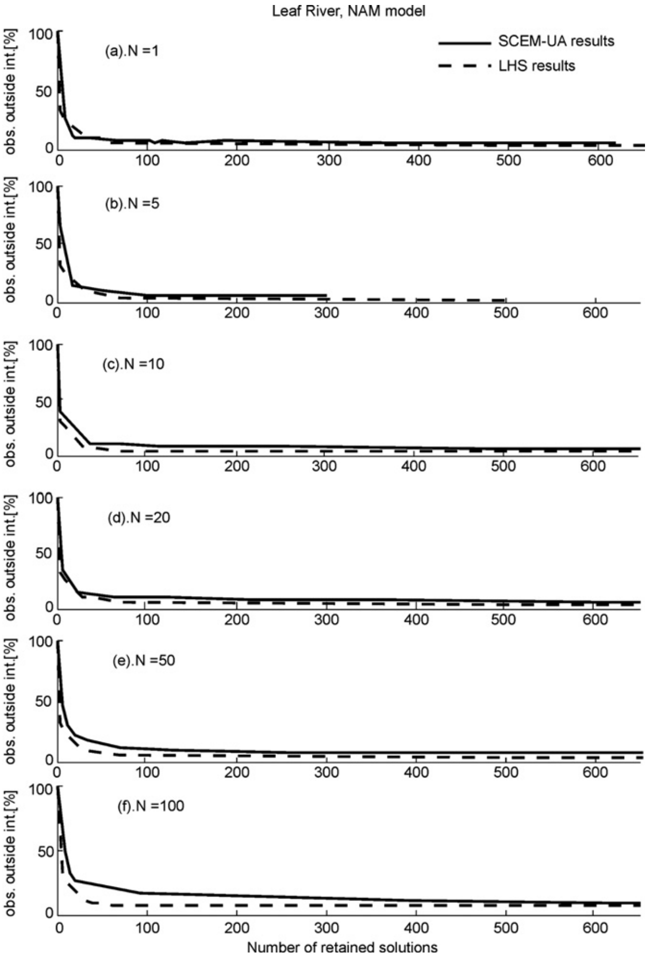

Fig. 5. Leaf River watershed – NAM model: percentage of runoff observations outside GLUE LHS and SCEM-UA uncertainty intervals versus number of retained solutions. Plots correspond to different values of the exponent of the likelihood function, N: (a) $N = 1$; (b) $N = 5$; (c) $N = 10$; (d) $N = 20$;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

4.1.2. GLUE 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.

Figs. 4 and 5 are plots of the percentage of observations falling outside the 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 Fig. 6, which depicts the average width of the streamflow uncertainty bounds as function of the number of retained solutions for different values of N.

To examine this behavior further, consider Fig. 7 (SAC-SMA model, Tryggevælde watershed) and Fig. 8 (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 do not accurately reproduce the real uncertainty

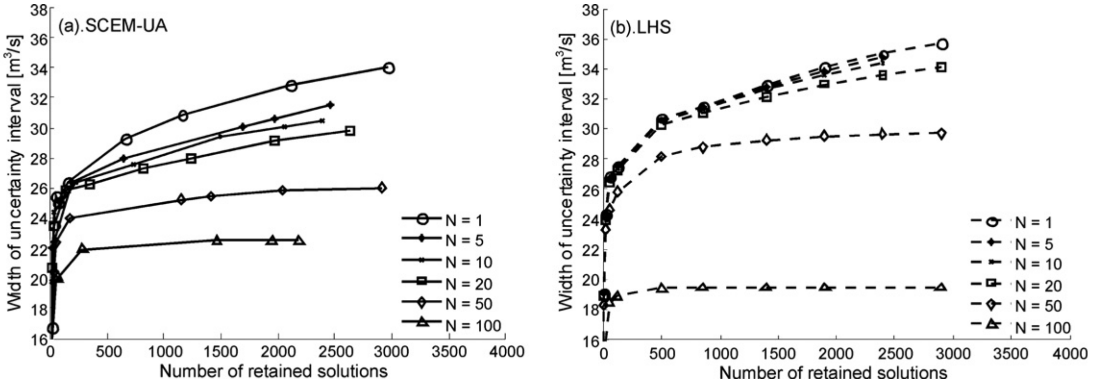

Fig. 6. Leaf River watershed – SAC-SMA model: width of the uncertainty bounds as a function of the number of retained solutions: (a) SCEM-UA and (b) LHS results.

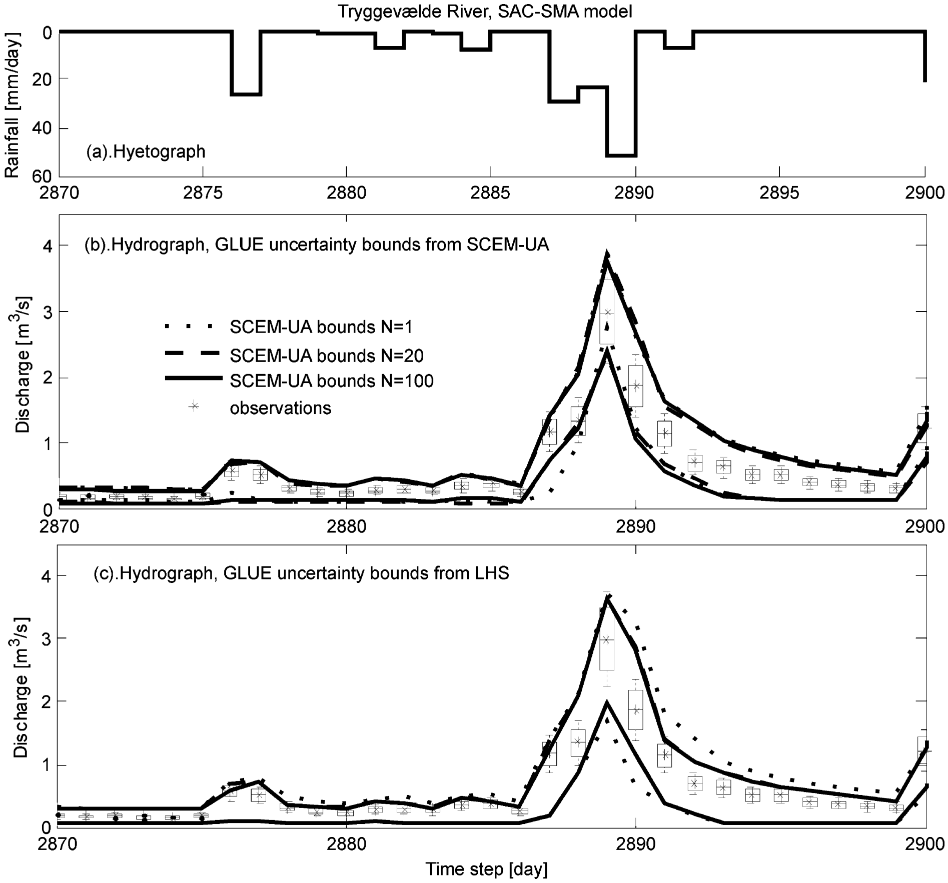

Fig. 7. Tryggevælde watershed – SAC-SMA model: hyetograph (a) and hydrographs including the uncertainty bounds containing the 90% of the observations generated by GLUE from SCEM-UA (b) and LHS initial samples (c). The error bars in these plots represent the error properties of the streamflow data: the 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.

and they appear to be too large, especially for the SAC-SMA model for low flows. This is a limitation of the GLUE method, 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. More accurate and 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 [59], or by fitting weighted probability distributions around the predictions of individual models (Bayesian model averaging [60]).

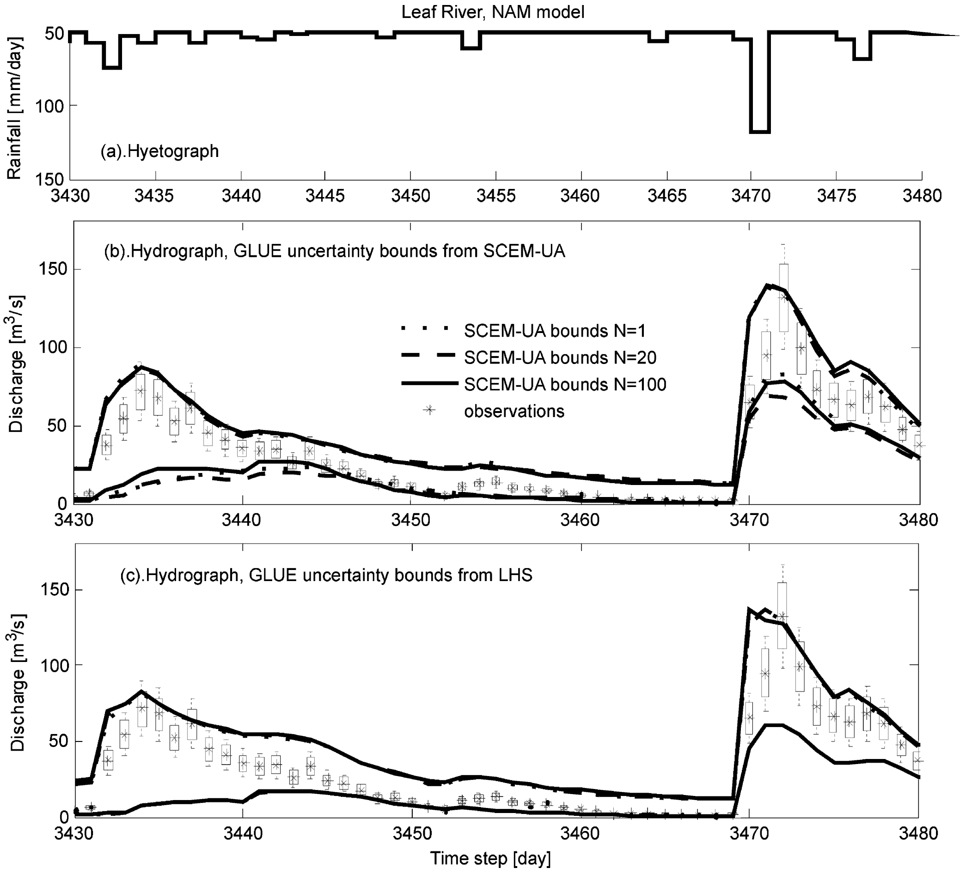

Fig. 8. Leaf River watershed – NAM model: hyetograph (a) and hydrographs including the uncertainty bounds containing the 90% of the observations generated by GLUE from SCEM-UA (b) and LHS initial samples (c). The error bars in these plots represent the error properties of the streamflow data: the boxes 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.

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 parameter PDFs for different values of N are presented for the parameter L_{max} in the NAM model (Fig. 9) and the parameter LZFSM in the SAC-SMA model (Fig. 10). This selection of parameters and models is representative of the entire set of results.

First, note that the LHS and SCEM-UA derived 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 PDFs become narrower and peakier with increasing N-values. However, the LHS derived PDFs remain multi-modal, while the SCEM-UA derived histograms become Gaussian-like with a single well-defined mode (the desired result). Finally, note that the mode of the LHS and SCEM-UA derived 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.

As illustration, Fig. 11 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 [58]. Correlations between parameters in other models were typically low, but increase with increasing N-value for the SCEM-UA sampling.

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, by consid-

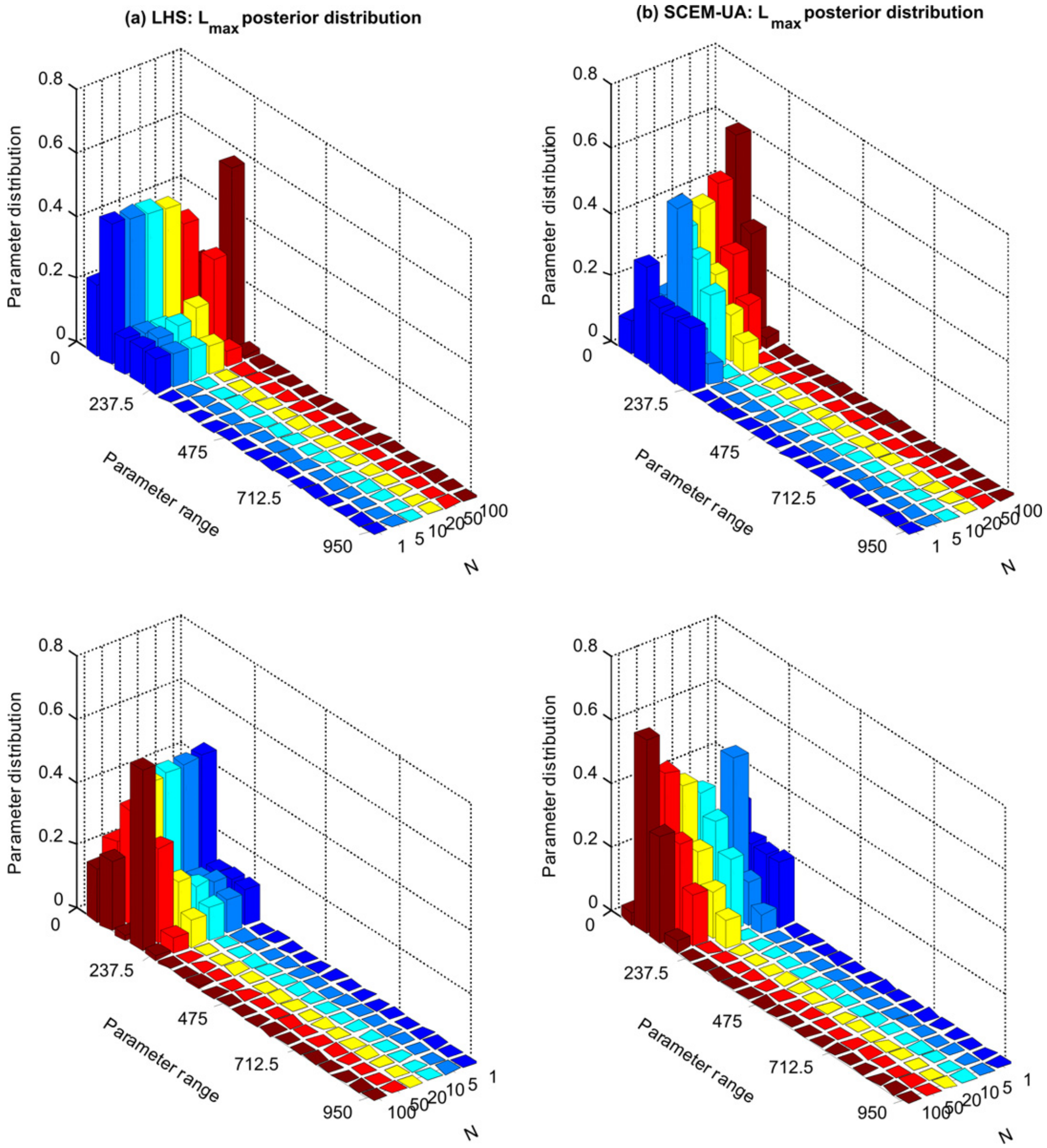

Fig. 9. Posterior distribution of parameter L_{max} for Tryggevælde watershed – NAM model obtained from SCEM-UA (a) and LHS dataset (b). Real value: $L_{max} = 121.1$.

ering only parameter variability. Table 4 summarizes these results for $N = 1$ and lists the percentage of observations included within the uncertainty 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

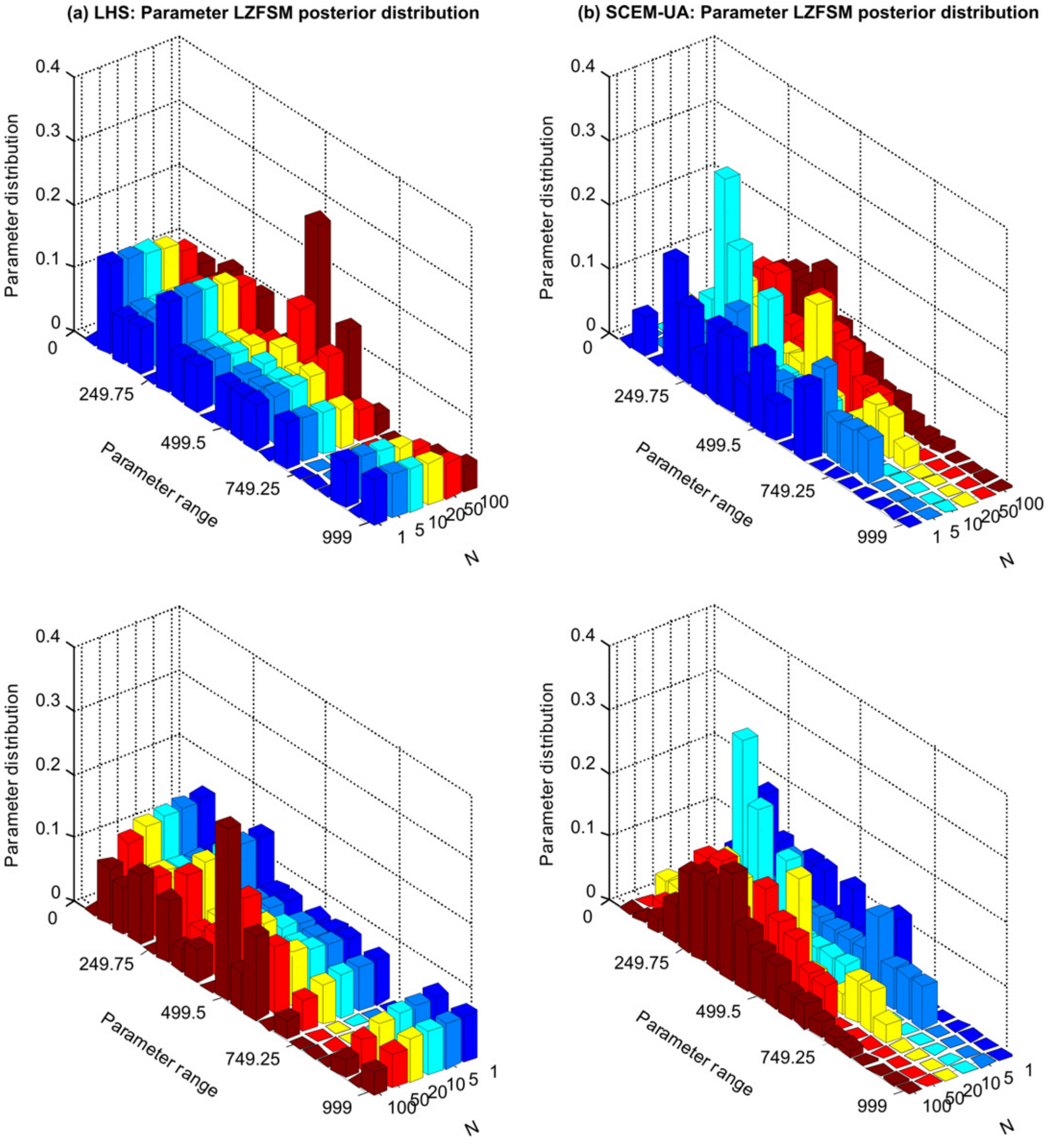

Fig. 10. Posterior distribution of parameter LZFSM for Tryggevælde watershed – Sacramento model: obtained from SCEM-UA (a) and LHS dataset (b). Real value: LZFSM = 438.85.

number of retained solutions when deriving the initial sample with the SCEM-UA algorithm.

4.2.2. GLUE 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 meaningful probabilistic 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. Figs. 12 and 13 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 (Fig. 13).

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.

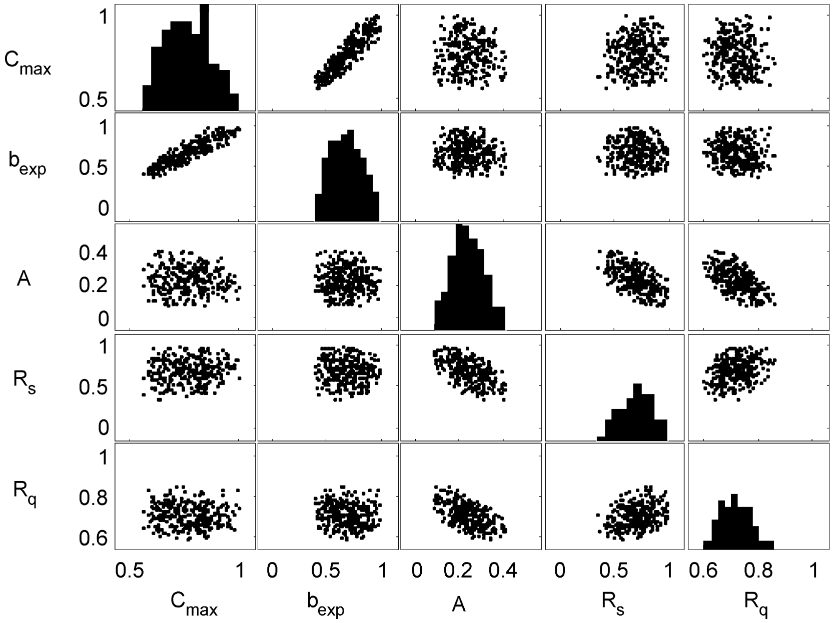

Fig. 11. HYMOD model – Tryggevælde river: correlation plots of normalized parameters from posterior distributions obtained from SCEM-UA sample with likelihood function exponent $N = 100$. Diagonal: histograms of parameter distribution.

Table 4
Percentage of observations contained within the GLUE uncertainty intervals and number of retained solutions (in parentheses)

Model	Tryggevælde		Leaf River	
	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)

Results correspond to the Tryggevælde and Leaf River data sets using the LHS and SCEM-UA methods (likelihood exponent $N = 1$).

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% and 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 probabilistic meaning and appropriate coverage of the observations. Nevertheless, the analysis of the distributions of the available parameters fully confirms the results for the artificially generated data sets. This is also

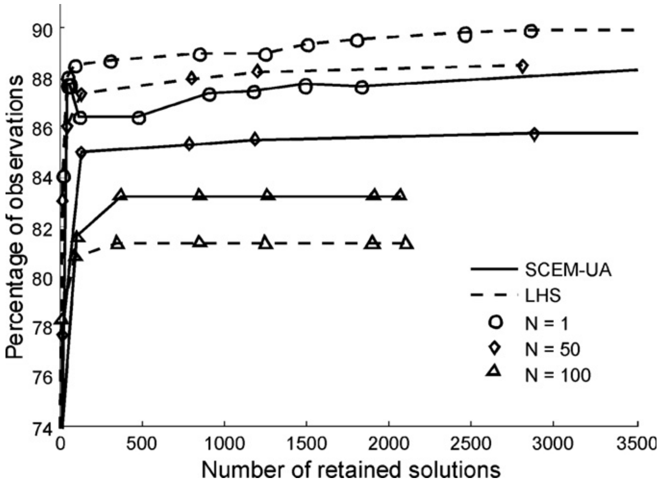

Fig. 12. Leaf River watershed – SAC-SMA model: percentage of solutions included within the uncertainty bounds as a function of the number of retained solutions.

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

N	SCEM-UA			LHS		
	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

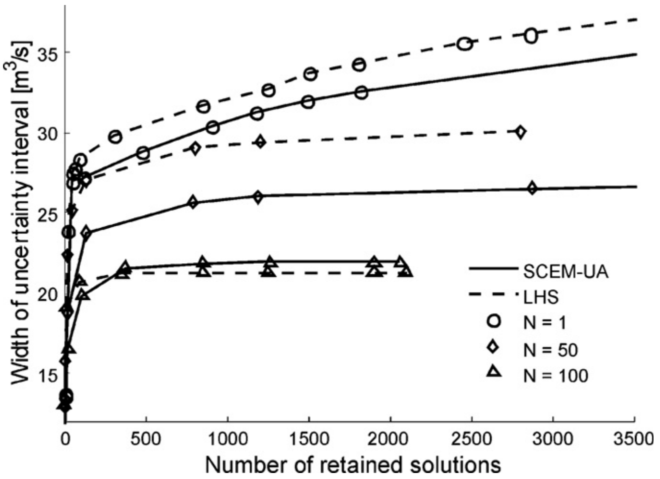

Fig. 13. Leaf River watershed – SAC-SMA model: width of the uncertainty bounds as a function of the number of retained solutions.

exemplified in Fig. 14, 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 parameter 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 $N = 100$.

5. Summary and conclusions

Most applications of GLUE reported in the literature implement a rather simplistic MC sampling scheme to sample from the prior parameter distributions and to find the set of behavioral models and their associated predictive simulation uncertainty. While this approach may be adequate for relative simple low-dimensional sampling problems, it is unlikely to result in stable and consistent estimates of the set of behavioral models (and thus parameter distributions) for relatively high-dimensional and complex inference problems. In this paper we have demonstrated 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

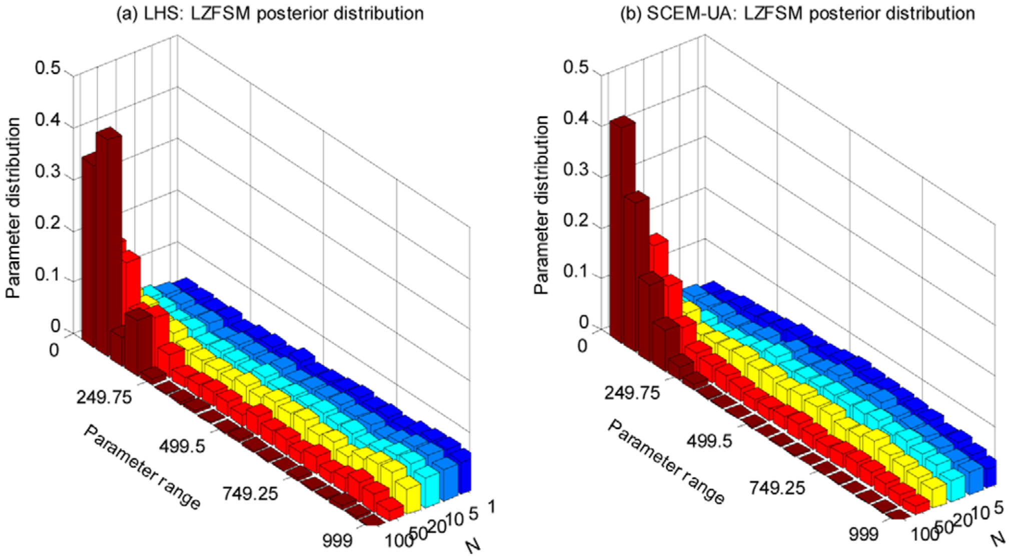

Fig. 14. Posterior distribution of parameter LZFSM for Leaf River and Sacramento model obtained from LHS (a) and SCEM-UA dataset (b). The number of observations contained within the uncertainty interval ranges from 82% to 90% in this case.

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 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 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 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 representative 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. SCEM-UA reached convergence with less model simulations than the maximum number chosen for each model. On the other side, a denser sampling of the parameter space would have been necessary for LHS to effectively sample the solution space and obtain similar posterior distributions than SCEM-UA. This result is especially important 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.

Acknowledgements

The second author is supported by a J. Robert Oppenheimer Fellowship from the LANL Postdoctoral program. The authors thank Torben Dolin (Department of Environmental Engineering) for formatting the figures included in this paper.

References

[1] Ajami NK, Duan Q, Sorooshian S. An integrated hydrologic bayesian multi-model combination framework: confronting input, parameter and model structural uncertainty. Water Resour Res 2007;43:W01403. doi:10.1029/2005WR004745.

[2] Aronica G, Bates PD, Horritt MS. Assessing the uncertainty in distributed model predictions using observed binary pattern information within GLUE. Hydrol Process 2002;16:2001–16.

[3] Beven KJ, Binley AM. The future of distributed models: model calibration and uncertainty prediction. Hydrol Process 1992;6: 279–98.

[4] Beven K. A manifesto for the equifinality thesis, The Model Parameter Estimation Experiment – MOPEX. J Hydrol 2006;320(1):18–36.

[5] Beven K, Smith P, Freer J. Comment on "Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology" by P. Mantovan, and E. Todini. J Hydrol 2007;338:315–8. doi:10.1016/j.jhydrol.2007.02.023.

[6] Boyle DP. Multicriteria calibration of hydrological models. Ph.D. thesis. Department of Hydrology and Water Resources, University of Arizona, Tucson; 2000.

[7] Boyle DP, Gupta HV, Sorooshian S. Toward improved calibration of hydrological models: combining the strengths of manual and automatic methods. Water Resour Res 2000;36:3663–74.

[8] Brazier RE, Beven KJ, Anthony SG, Rowan JS. Implications of model uncertainty for the mapping of hillslope-scale soil erosion predictions. Earth Surf Proc Land 2001;26:1333–52.

[9] Burnash RJC. The NWS river forecast system-catchment modeling. In: Singh VP, editor. Computer models of watershed hydrology. Colorado: Water Resources Publication; 1995. p. 311–66.

[10] Burnash RJC, Ferral RL, McGuire RA. A generalized streamflow simulation system: conceptual modeling for digital computers. Joint Federal-State River Forecast Center, Sacramento (CA); 1973.

[11] Christensen S. A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals. Nordic Hydrol 2004;35(1): 45–59.

[12] Christiaens K, Feyen J. Constraining soil hydraulic parameter and output uncertainty of the distributed hydrological MIKE SHE model using the GLUE framework. Hydrol Process 2002;16(2):373–91.

[13] Feyen L, Beven KJ, De Smedt F, Freer J. Stochastic capture zone delineation within the generalized likelihood uncertainty estimation methodology: conditioning on head observations. Water Resour Res 2001;37(3):625–38.

[14] Franks SW, Beven KJ, Quinn PF, Wright IR. On the sensitivity of soil–vegetation–atmosphere transfer (SVAT) schemes: equifinality and the problem of robust calibration. Agr Forest Meteorol 1997;86:63–75.

[15] Freer J, Beven KJ, Ambroise B. Bayesian estimation of uncertainty in runoff prediction and the value of data: an application of the GLUE approach. Water Resour Res 1996;32:2161–73.

[16] Gelman A, Rubin DB. Inference from iterative simulation using multiple sequences. Stat Sci 1992;7:457–72.

[17] Georgekakos KP, Seo DJ, Gupta H, Schaake J, Butts MB. Characterizing streamflow simulation uncertainty through multimodel ensembles. J Hydrol 2004;298(1–4):222–41.

[18] Gneiting T, Raftery AE, Balabdaoui F, Westveld A. Verifying probabilistic forecasts: calibration and sharpness In: Proceedings of the workshop on ensemble forecasting. Val-Morin (QC, Canada); 2003.

[19] Gupta HV, Sorooshian S, Yapo PO. Toward improved calibration of hydrologic models: multiple and noncommensurable measures of information. Water Resour Res 1998;34:751–63.

[20] Hankin BG, Hardy R, Kettle H, Beven KJ. Using CFD in a GLUE framework to model the flow and dispersion characteristics of a natural fluvial dead zone. Earth Surf Proc Land 2001;26(6):667–87.

[21] Hansson K, Lundin C. Equifinality and sensitivity in freezing and thawing simulations of laboratory and in situ data. Cold Reg Sci Technol 2006;44:20–37.

[22] Havnø K, Madsen MN, Dørge J. MIKE 11 – a generalized river modelling package. In: Singh VP, editor. Computer models of watershed hydrology. Colorado: Water Resources Publication; 1995. p. 733–82.

[23] Hornberger GM, Spear RC. An approach to the preliminary analysis of environmental systems. J Environ Manag 1981;12:7–18.

[24] Jensen JB. Parameter and uncertainty estimation in groundwater modelling. Ph.D. thesis. Department of Civil Engineering, Aalborg University, Series Paper No. 23; 2003.

[25] Kavetski D, Kuczera G, Franks SW. Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory. Water Resour Res 2006;42:W03407. doi:10.1029/2005WR004368.

[26] Keesman KJ. Set theoretic parameter estimation using random scanning and principal component analysis. Math Comput Simul 1990;32:535–43.

[27] Khu ST, Madsen H. Multiobjective calibration with Pareto preference ordering: an application to rainfall-runoff model calibration. Water Resour Res 2005;41(3):1–14.

[28] Klepper O, Scholten H, van de Kamer JPG. Prediction uncertainty in an ecological model of the Oosterschelde Estuary. J Forecasting 1991;10:191–209.

[29] Kuczera G, Kavetski D, Franks SW, Thyer MT. Towards a Bayesian total error analysis of conceptual rainfall-runoff models: characterising model error using storm-dependent parameters. J Hydrol 2006;331(1–2):161–77.

[30] Kuczera G, Parent E. Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the Metropolis algorithm. J Hydrol 1998;211:69–85.

[31] Lamb R, Beven K, Myrabø S. Use of spatially distributed water table observations to constrain uncertainty in a rainfall-runoff model. Adv Water Resour 1998;22(4):305–17.

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

[33] Lorup JK, Refsgaard JC, Mazvimavi D. 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 1998;205(3):147–63.

[34] Madsen H. Automatic calibration of a conceptual rainfall-runoff model using multiple objectives. J Hydrol 2000;235(3–4):276–88.

[35] Madsen H. Parameter estimation in distributed hydrological catchment modelling using automatic calibration with multiple objectives. Adv Water Resour 2003;26:205–16.

[36] Madsen H, Rosbjerg D, Damgård J, Hansen FS. Data assimilation in the MIKE 11 flood forecasting system using Kalman filtering, water resources systems – hydrological risk, management and development. In: Proceedings of symposium HS02b, IUGG2003 at Sapporo, July 2003. IAHS Publ. No. 281; 2003. p. 75–81.

[37] Makowski D, Wallach D, Tremblay M. Using a Bayesian approach to parameter estimation; comparison of the GLUE and MCMC methods. Agronomie 2002;22:191–203. doi:10.1051/agro.2002007.

[38] Mantovan P, Todini E. Hydrological forecasting uncertainty assessment: incoherence of the GLUE methodology. J Hydrol 2006;330:368–81.

[39] McMichael CE, Hope AS, Loaiciga HA. Distributed hydrological modeling in California semi-arid shrublands: MIKE SHE model calibration and uncertainty estimation. J Hydrol 2006;317:307–24.

[40] Mertens J, Madsen H, Feyen L, Jacques D, Feyen J. Including prior information in the estimation of effective soil parameters in unsaturated zone modeling. J Hydrol 2004;294(4):251–69.

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

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

[43] Mugunthan P, Shoemaker CA. Assessing the impacts of parameter uncertainty for computationally expensive groundwater models. Water Resour Res 2006;42:W10428. doi:10.1029/2005WR004640.

[44] Muleta MK, Nicklow J. Sensitivity and uncertainty analysis coupled with automatic calibration for a distributed watershed model. J Hydrol 2005;306:127–45.

[45] Nielsen SA, Hansen E. Numerical simulation of the rainfall runoff processes on a daily basis. Nordic Hydrol 1973;4:171–90.

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

[47] Peck EL. 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; 1976.

[48] Raftery AE, Gneiting T, Balabdaoui F, Polakowsk M. Using Bayesian model averaging to calibrate forecast ensembles. Monthly Weather Review 2005;133(5):1155–74.

[49] Romanowicz R, Beven KJ, Tawn J. Evaluation of predictive uncertainty in non-linear hydrological models using a Bayesian approach. In: Barnett V, Turkman KF, editors. Statistics for the

environment. Water related issues, vol. 2. Hoboken (NJ): John Wiley. p. 297–317.

[50] Romanowicz RJ, Beven KJ, Tawn J. Bayesian calibration of flood inundation models. In: Anderson MG, Walling DE, editors. Floodplain processes. Wiley; 1996. p. 333–60.

[51] Strom B, Jensen KH, Refsgaard JC. Estimation of catchment rainfall uncertainty and its influence on runoff prediction. Nordic Hydrol 1988;19(2):77–88.

[52] van Straten G, Keesman KJ. Uncertainty propagation and speculation in projective forecasts of environmental change: a lake eutrophication example. J Forecasting 1991;10:163–90.

[53] Tadesse A, Anagnostou EN. A statistical approach to ground radar-rainfall estimation. J Atm Ocean Technol 2005;22(11):1055–71.

[54] Thiemann M, Trosset M, Gupta H, Sorooshian S. Bayesian recursive parameter estimation for hydrological models. Water Resour Res 2001;37(10):2521–35.

[55] Vogel RM, Batchelder R, Stedinger RJ. Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. Water Resour Res [in press].

[56] Vrugt JA, Weerts AH, Bouten W. Information content of data for identifying soil hydraulic parameters from outflow experiments. Soil Sci Soc Am J 2001;65:19–27.

[57] Vrugt JA, Gupta HV, Bouten W, Sorooshian S. A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. Water Resour Res 2003;39(8):1201. doi:10.1029/2002WR001642.

[58] Vrugt JA, Gupta HV, Bastidas LA, Bouten W, Sorooshian S. Effective and efficient algorithm for multi-objective optimization of hydrologic models. Water Resour Res 2003;39(8):1214. doi:10.1029/2002WR001746.

[59] Vrugt JA, Diks CGH, Gupta HV, Bouten W, Verstraten JM. Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. Water Resour Res 2005;41(1):1–17.

[60] Vrugt JA, Robinson BA. Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. Water Resour Res 2007;43:W01411. doi:10.1029/2005WR004838.

[61] Wang X, He X, Williams JR, Izaurralde RC, Atwood JD. Sensitivity and uncertainty analyses of crop yields and soil organic carbon simulated with EPIC. Trans Am Soc Agric Eng 2005;48(3):1041–54.

[62] Yapo PO, Gupta HV, Sorooshian S. Multi-objective optimization for hydrologic models. J Hydrol 1998;204:83–97.