


## Randomized block quasi-Monte Carlo sampling for generalized likelihood uncertainty estimation

Charles Onyutha 

Department of Civil and Environmental Engineering, Kyambogo University, P.O. Box 1, Kyambogo, Kampala, Uganda  
E-mail: conyutha@gmail.com; conyutha@kyu.ac.ug

 CO, 0000-0002-0652-3828

### ABSTRACT

Although hydrological model forecasts aid water management decisions, they normally have predictive uncertainties. Generalized likelihood uncertainty estimation (GLUE) is popular for constructing predictive uncertainty bounds (PUBs). GLUE is based on simple Monte Carlo sampling (SMCS), a technique known to be ineffective in establishing behavioural simulations. This study introduced randomized block quasi-Monte Carlo sampling (RBMC). In RBMC, each parameter's range is divided into a stipulated number of sub-blocks ( $S_{nb}$ ). Parameters' values are separately generated in each sub-block. Finally, the sub-blocks are shuffled while maintaining the sequence of generated values in each sub-block. When  $S_{nb}$  is equal to the number of simulations, RBMC reduces to SMCS. Otherwise, each  $S_{nb}$  leads to a separate RBMC configuration or sampling scheme. The number of RBMC-based behavioural solutions was often found to be greater than that of SMCS, in some cases, by up to 33.6%. The width of the 90% confidence interval on 95th percentile flow based on SMCS was often larger than those of RBMC, sometimes by up to 23.4%. PUBs were found to vary in widths among RBMC configurations, thereby revealing the influence of the choice of a sampling scheme. Thus, GLUE based on RBMC is recommended to take into account the said influence.

**Key words:** generalized likelihood uncertainty estimation, hydrological model uncertainty, Latin hypercube sampling, Monte Carlo sampling, predictive uncertainty, randomized block quasi-Monte Carlo sampling

### HIGHLIGHTS

- GLUE uses simple Monte Carlo sampling (SMCS) which is ineffective in establishing behavioral simulations.
- This study introduced randomized block quasi-Monte Carlo Sampling (RBMC) for GLUE and SMCS becomes one of the various RBMC configurations.
- RBMC improved the number of retained solutions by up to 33.6% in some cases.
- RBMC improved the width of 90% confidence interval on a flow event by up to 23.4%.
- RBMC takes into account the influence of the choice of a sampling scheme as a sub-source of calibration uncertainty.

## GRAPHICAL ABSTRACT

Step 1: Stipulating lower and upper limits of each parameter

Parameter	Lower limit	Upper limit
$\theta_1$	$l_{\theta(1)}$	$u_{\theta(1)}$
$\theta_2$	$l_{\theta(2)}$	$u_{\theta(2)}$
$\theta_3$	$l_{\theta(3)}$	$u_{\theta(3)}$
$\theta_4$	$l_{\theta(4)}$	$u_{\theta(4)}$
$\theta_5$	$l_{\theta(5)}$	$u_{\theta(5)}$



Step 2: Dividing the range of each parameter into sub-blocks

Parameter	Sub-block 1	Sub-block 2	Sub-block 3	Sub-block 4
$\theta_1$	$z_{1,1}$	$z_{1,2}$	$z_{1,3}$	$z_{1,4}$
$\theta_2$	$z_{2,1}$	$z_{2,2}$	$z_{2,3}$	$z_{2,4}$
$\theta_3$	$z_{3,1}$	$z_{3,2}$	$z_{3,3}$	$z_{3,4}$
$\theta_4$	$z_{4,1}$	$z_{4,2}$	$z_{4,3}$	$z_{4,4}$
$\theta_5$	$z_{5,1}$	$z_{5,2}$	$z_{5,3}$	$z_{5,4}$



Step 3: Generating each parameter's values in every sub-block

$$F_{z_{i,j}}(z_{i,j}) = P[X \leq x] = \int_{-\infty}^x f_{z_{i,j}}(z_{i,j}') \quad \text{and} \quad z_{i,j} = F_X^{-1}(u_{i,j})$$



Step 4: Separately shuffling sub-blocks of each parameter

Parameter	Sub-block 1	Sub-block 2	Sub-block 3	Sub-block 4
$\theta_1$	$z_{1,2}$	$z_{1,1}$	$z_{1,3}$	$z_{1,4}$
$\theta_2$	$z_{2,4}$	$z_{2,2}$	$z_{2,1}$	$z_{2,3}$
$\theta_3$	$z_{3,2}$	$z_{3,1}$	$z_{3,4}$	$z_{3,3}$
$\theta_4$	$z_{4,3}$	$z_{4,4}$	$z_{4,2}$	$z_{4,1}$
$\theta_5$	$z_{5,1}$	$z_{5,3}$	$z_{5,4}$	$z_{5,2}$

## 1. INTRODUCTION

The indispensability of modelling as a tool for understanding different hydrological processes obligates scientific researchers to continually focus on ways of understanding and/or reducing model uncertainties. Uncertainty can be epistemic (Der Kiureghian & Ditlevsen 2009; Beven 2016; Nearing *et al.* 2016; Gupta & Govindaraju 2023) or aleatoric (Hora 1996; Gong *et al.* 2013; Nearing *et al.* 2016). Epistemic uncertainty can be linked to the model parameters and comprises a modeller's ignorance of his or her choice of the best-performing model. Aleatoric uncertainty is due to the noise inherent in the observations. To reduce epistemic (or reducible) uncertainty, a modeller has to make use of additional data (Nearing *et al.* 2016; Zhou *et al.* 2022) for training or calibration. Aleatoric uncertainty can be reduced by enhancing the precision and accuracy of measuring equipment. Aleatoric and epistemic uncertainties combine to yield predictive uncertainty in the model output (Beven & Binley 1992; Gupta & Govindaraju 2023). Although predictive uncertainty reduction does not mean simplification of decision-making, it is to increase the 'trustworthiness' of the inferences made on forecasts from scenario results.

Apart from model structure, parameters, and input data, calibration comprises an important area on which a modeller can focus to reduce model uncertainty (Beven & Binley 1992; Kavetski *et al.* 2006; Zhang *et al.* 2009). Calibration entails changing the values of a model's parameters to guarantee the minimal mismatch between the observed and modelled variables. In other words, calibration is performed to minimize parameter estimation uncertainties (Eckhardt *et al.* 2005) and maximize a model's reliability.

Changing a model's parameters can be done manually or automatically. The idea behind manual calibration is that when certain parameters are adjusted (for instance, by increasing their magnitudes), some predictable changes in the modelled outputs can be obtained. When several interacting parameters are adjusted, the changes in the outputs can be unpredictable (Gupta *et al.* 1999). Furthermore, manual calibration tends to be time-consuming and requires expert knowledge of the modeller (Boyle *et al.* 2000). For complex models with many parameters, manual calibration can be inefficient and frustrating. In fact, a strict adherence to model calibration using manual procedure would inhibit the widespread use of complex and sophisticated models (Gupta *et al.* 1999). To overcome the challenges of changing parameters manually, automation of the model calibration started way back in the mid-1960s (see for instance, Dawdy & O'Donnell 1965). Automatic changing of parameters involves applications of mathematical and statistical approaches to minimize the model residuals based on mainly optimization functions. Even in the automatic calibration, it remains unlikely, given the various model uncertainties, that a modeller can obtain only one set of optimal parameters (Qi *et al.* 2019). This notion led to equifinality (Beven & Binley 1992; Beven & Freer 2001), a concept which recognizes the fact that there can be several sets of model parameters to yield a highly comparable model performance. The recognition and acceptance of equifinality steered the development of the generalized likelihood uncertainty estimation (GLUE) framework by Beven & Binley (1992).

GLUE is the Bayesian approach (which is, actually, an automatic calibration strategy) and consists of (i) stipulation of each parameter's upper and lower limits, (ii) randomization of many (e.g. 10,000) sets of model parameters from the prior distribution, and (iii) inferring posterior distribution using simulations. Apart from the GLUE framework, several approaches exist for establishing bounds of uncertainty on model predictions such as sequential data assimilation (Moradkhani *et al.* 2005; Vrugt *et al.* 2005), multi-model averaging methods (Georgekakos *et al.* 2004; Vrugt & Robinson 2007), classical Bayesian (Kuczera & Parent 1998; Thieman *et al.* 2001), and pseudo-Bayesian (Freer *et al.* 1996). Other uncertainty analysis methods include the Bayesian total error analysis (Kavetski *et al.* 2003), parameter estimation code (PEST) (Doherty 2010), multi-objective analysis (Hadka & Reed 2013), and differential evolution adaptive metropolis (DREAM) (Vrugt 2016). While these methods have different underlying assumptions, each of them also has its own advantages and disadvantages. Nevertheless, among the various uncertainty analysis methods, GLUE is very popular due to its conceptual simplicity, ease of implementation, and capacity to handle various error structures and models (Blasone *et al.* 2008).

It can be argued that the randomization of parameter values in the GLUE framework with regard to Monte Carlo analysis is mainly ineffective in establishing behavioural simulations (Blasone *et al.* 2008). This problem can be compounded by the substantial computational time required to obtain the stipulated behavioural modelled series from complex models, and the hardship in dealing with high-dimensional parameter estimation problems (Blasone *et al.* 2008). Furthermore, the use of less formal likelihood by GLUE can lead to very flat posterior distributions (Mantovan & Todini 2006; Stedinger *et al.* 2008; Liu *et al.* 2022). To obtain better posterior distributions, one would think of other Markov chain Monte Carlo (MCMC) methods. Actually, through the use of simple Monte Carlo random sampling (hereinafter denoted as RND), GLUE does not take into account the influence of the choice of parameter sampling technique on calibration results and this is the gap on which this paper focused.

This paper introduced randomized block quasi-Monte Carlo sampling (RBMC) while investigating its suitability to support the GLUE framework. The RBMC starts by dividing each parameter range into a stipulated number of intervals or sub-blocks. The parameter's values are separately generated in each interval. The final step consists of shuffling the sub-blocks while the sequence of the generated values in each sub-block is not affected. In this way, RND is nested within the RBMC in terms of the chosen number of sub-blocks.

## 2. MATERIALS AND METHODS

### 2.1. RBMC sampling

Monte Carlo sampling relies on a pseudo-random generator with the underlying concept of using randomness to solve tasks which may, in principle, comprise deterministic problems. An important note is that the approach of using a

pseudo-random generator leads to values which are chaotic or extremely in disarray. Thus, the generated random numbers are not equidistant and they have uneven differences in their magnitudes. In a quasi-Monte Carlo sampling, we can use the sequence of low discrepancy to bring about a faster rate of convergence. A number of low-discrepancy sequences exist such as the Faure sequence, Halton sequence, and Sobol sequence. A key challenge of low-discrepancy sequences is that they are deterministic and not random and this means that quasi-Monte Carlo sampling can be considered a de-randomized or deterministic approach. De-randomization in the quasi-Monte Carlo sampling leads to error bound which makes the estimation of the error hard. Since we want a method which can allow us to estimate the variance, randomization becomes a plausible technique to modify the quasi-Monte Carlo sampling into the randomized quasi-Monte Carlo sampling. One technique is the shuffling of sub-blocks of the generated values. Eventually, this study introduced RBMC.

To explain the RBMC, we can start from the RND. Let  $n$  be the total number of model parameters. In the Monte Carlo simulation approach, which makes use of RND, the values of the  $u_j$ 's get drawn from the entire range 0–1. Let  $f(x)$  and  $F(x)$  be the probability density function (pdf) and cumulative distribution function (CDF), respectively, of a certain random variable  $X$ . Using RND, a realization  $x_j$  of  $X$  can be obtained through the calculation from the pdf (Equation (1)) based on a uniform deviate  $u_j$  over the range 0–1 using a pseudo-random number generator such that

$$F_Z(z) = P[X \leq x] = \int_{-\infty}^x f_Z(z') \quad (1)$$

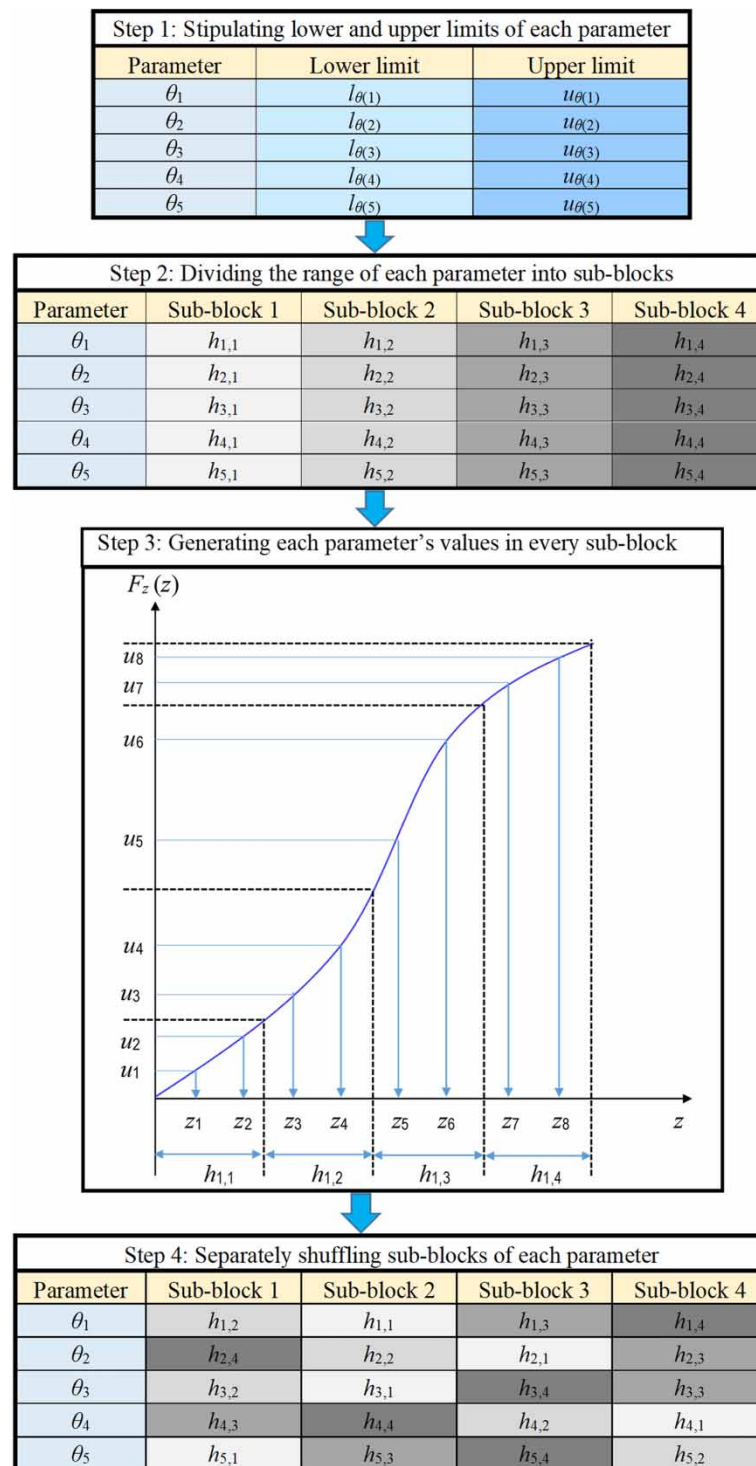
and  $u_j$  can be converted using the inverse function  $z_j = F_X^{-1}(u_j)$ . In RND, values of a parameter are generated in a purely random way over the specified domain.

In RBMC, the first step comprises the specification of the range of each parameter value. In other words, for  $1 \leq i \leq n$ , we provide the upper ( $u_{\theta(i)}$ ) and lower ( $l_{\theta(i)}$ ) limits of the  $i$ th parameter. In the second step, we choose the term  $b$  that is the number of sub-blocks of values of each parameter. In the third step, we draw numbers from a distribution function based on the term  $b$ . Consider that  $n_{\text{sim}}$  denotes the number of simulations. The number of values to be generated from each sub-block is given by the ratio of  $n_{\text{sim}}$  to  $b$ . A good idea is that  $n_{\text{sim}}$  should be chosen such that it is divisible by  $b$ . This is possible for the case when a single RBMC configuration is being considered. When a modeller chooses to consider several RBMC configurations, it means that the term  $b$  will have several values. Thus, it becomes impossible to choose a particular  $n_{\text{sim}}$  that can be divisible by each of the values of  $b$  for the respective RBMC configurations. In other words, in cases where  $n_{\text{sim}}$  is not divisible by  $b$  (or when  $\text{mod}(n_{\text{sim}}, b) > 0$ ), we introduce another term  $n_{\text{wa}}$  (or the modified number of simulations) such that  $n_{\text{wa}} = (n_{\text{sim}} - \text{mod}(n_{\text{sim}}, b) + b)$  where  $\text{mod}(n_{\text{sim}}, b)$  stands for  $n_{\text{sim}} \bmod b$ . Therefore, the number of values of each parameter to be generated within each interval or sub-block is given by  $(n_{\text{sim}}/b)$  and  $(n_{\text{wa}}/b)$  for the cases when  $\text{mod}(n_{\text{sim}}, b) = 0$  and  $\text{mod}(n_{\text{sim}}, b) > 0$ , respectively. Figure 1 shows an illustration of four sub-blocks ( $b = 4$ ) for a model with five parameters. The values ( $m_{(i,j)}$ ) denoting the bounds of the sub-blocks of values of each parameter can be given by

$$m_{(i,j)} = l_{\theta(i)} + (j-1) \times \frac{(u_{\theta(i)} - l_{\theta(i)})}{b} \quad \text{for } 1 \leq i \leq n \quad \text{and} \quad 1 \leq j \leq (b+1) \quad (2)$$

For the initial parameter (or when  $i = 1$ ), the first ( $h_{1,1}$ ), second ( $h_{1,2}$ ), third ( $h_{1,3}$ ), and fourth ( $h_{1,4}$ ) sub-blocks are bounded by  $(m_{(1,1)}, m_{(1,2)}]$ ,  $(m_{(1,2)}, m_{(1,3)}]$ ,  $(m_{(1,3)}, m_{(1,4)}]$ , and  $(m_{(1,4)}, m_{(1,5)}]$ , respectively. It should be apparent that  $l_{\theta(1)} = m_{(1,1)}$  and  $m_{(1,5)} = u_{\theta(1)}$ . The next step entails determining the number of values to be generated from each sub-block of every parameter.

Let us assume that in our case,  $b = 4$  and  $n_{\text{sim}} = 8$  such that we want to generate two (i.e.  $8/4 = 2$ ) values from each sub-block (as illustrated in Figure 1). Based on the initial model parameter ( $\theta_1$ ), we have  $h_{1,1} = [z_1, z_2]$ ,  $h_{1,2} = [z_3, z_4]$ ,  $h_{1,3} = [z_5, z_6]$  and  $h_{1,4} = [z_7, z_8]$  where  $z_j = F_Z^{-1}(u_j)$  and  $1 \leq j \leq 8$ . In other words, we separately apply Equation (1) to each sub-block of every parameter's values. The last step entails separately shuffling the sub-blocks of the values of each parameter. Here, the shuffling process should be carefully done to ensure that only the sub-blocks are shuffled while the sequence of generated values in each sub-block is not altered. At this point, the total number of generated values for each parameter should be  $n_{\text{sim}}$  and  $n_{\text{wa}}$  for the cases where  $\text{mod}(n_{\text{sim}}, b) = 0$  and  $\text{mod}(n_{\text{sim}}, b) > 0$ , respectively. For the first case (or when  $\text{mod}(n_{\text{sim}}, b) = 0$ ), we finally consider all the  $n_{\text{sim}}$  generated values of each parameter for model calibration. In the second case (or when  $\text{mod}(n_{\text{sim}}, b) > 0$ ), only the first  $n_{\text{sim}}$  (from the total  $n_{\text{wa}}$ ) generated values of each parameter are considered for model calibration or simulation. For  $1 \leq k \leq n_{\text{sim}}$ , a complete  $k$ th set of the model parameters is obtained by taking every  $k$ th value of each parameter. Therefore, a model can be run  $n_{\text{sim}}$  times to yield  $n_{\text{sim}}$  modelled series.



**Figure 1** | Procedure of RBMC for the case with four sub-blocks when a model has five parameters.

An important note is that there are two special cases when RBMC is the same as RND. The first case is when the value of the term  $b$  is set to one (1), meaning that the  $n_{\text{sim}}$  value of each parameter is generated once. This first case can rarely be applicable for simulation analysis given the need for large  $n_{\text{sim}}$  required to achieve realistic predictions in the context of Monte Carlo analysis. The second case is when the term  $b$  is set to  $n_{\text{sim}}$  (given that  $n_{\text{sim}}$  is far larger than 1) indicating



that one value of a given parameter is generated in each of the  $n_{\text{sim}}$  sub-blocks of every parameter's values. In other words, the number of every parameter's generated values becomes equal to  $n_{\text{sim}}$ . This means that RBMC is a composite Monte Carlo sampling method which comprises RND as one of its configurations.

## 2.2. Existing schemes for comparison with RBMC

### 2.2.1. Latin hypercube sampling

Latin square is a square grid with sampling points such that each row and every column include one point. If this concept is extended to large numbers of dimensions such that each axis-aligned hyperplane consists of one point, it means we are dealing with a Latin hypercube. To generate  $v$  sample values, we divide the total area under the pdf into  $v$  equal areas. From each area, one random value is generated. Considering all the areas, we obtain a sample of  $v$  values. For instance, let us consider that we are dealing with a uniform distribution whose pdf with domain  $[c, d]$  is to be divided into five equal areas. We divide the domain to obtain  $[c, t_1]$ ,  $[t_1, t_2]$ ,  $[t_2, t_3]$ ,  $[t_3, t_4]$  and  $[t_4, d]$ . We make use of the analytical inversion method to generate one point from each interval. Thus, the interval  $[0, 1]$  is split into  $v$  equal portions and from each portion, a random variable is generated to obtain  $u_1, u_2, \dots, u_v$ . The generated values become  $x_1, x_2, \dots, x_v$  based on  $x_1 = F^{-1}(u_1)$ ,  $x_2 = F^{-1}(u_2)$ ,  $\dots$ ,  $x_v = F^{-1}(u_v)$ .

### 2.2.2. Stratified Monte Carlo random sampling

Consider that a certain random variable  $Z$  has pdf and CDF denoted by  $f_Z(z)$  and  $F_Z(z)$ , respectively. For a stratified random sampling, we divide the range 0–1 into a finite number of intervals which can be of the same or equal width (such as 0–0.2, 0.2–0.4, 0.4–0.6, 0.6–0.8, 0.8–1.0 if there should be five intervals). Over each interval, we can generate the same number of deviates  $u_j$ 's and the associated  $z_j$ 's. The idea is that the procedure should lead to realizations of  $Z$  which are nearly evenly spread over the range with the advantage of achieving realistic estimates of  $f_Z(z)$  using few realizations. One may realize that Latin hypercube sampling (LHS) is comparable with the stratified random sampling. However, an important note is that if we draw  $v$  samples with LHS, we will have  $v$  equal sub-space, while stratified random sampling can have  $w$  sub-space and in each sub-space  $k$ , samples will be drawn resulting into a total of  $v$  values or  $w \times k = v$ .

## 2.3. Case study

### 2.3.1. Selected data and models

Quality-controlled daily precipitation and potential evapotranspiration (PET) over the River Mpanga catchment with an area of 1,484 km<sup>2</sup> in Uganda were adopted from [Onyutha et al. \(2021\)](#). To allow comparison of the new and existing methods, two lumped conceptual hydrological models were applied. The first model was the VHM ([Willems 2014](#)) and it makes use of catchment-wide average of precipitation and PET as time-variable model inputs to generate runoff as a function of the soil moisture storage. The model splits runoff into overland flow, interflow, and slow flow and these components are separately routed and later combined into the modelled flow. The other model was the Nedbør-Afstrømnings model (NAM) ([Nielsen & Hansen 1973](#)). Like the VHM, NAM also uses lumped (or catchment-wide averaged) precipitation and PET as the meteorological model inputs. In NAM, surface storage is obtained as a function of the precipitation and PET. Quick flow and slow flow components are generated from the surface storage and separately routed and the resultant outputs are combined into the modelled flow from NAM.

### 2.3.2. The choice of the number of parameter sub-blocks

The choice of the term  $b$  in Equation (2) can be made on a case-by-case basis. For a high-dimensional parameter space, the term  $b$  could be set to range from 2 up to, say,  $\tau$ . To consider the uncertainty of the choice of parameter sampling scheme on model output,  $\tau$  can be greater than 10. However, since this study focused on testing the acceptability of RBMC for GLUE,  $\tau$  was set to 10. Thus, qMC2, qMC3, ..., qMC10, were considered.

### 2.3.3. Number of retained modelled series

Values of both NAM and VHM parameters were generated using RND, LHS, and nine configurations of the introduced method from qMC2 to qMC10. In other words, there were a total of 11 sampling schemes considered in this study. In the next step,  $u_\theta$  and  $l_\theta$  of each parameter were specified as shown in Appendix A. The number of simulations ( $n_{\text{sim}}$ ) was at first set to 1,000. In other words, a total of  $n_{\text{sim}}$  values of each parameter were generated using every sampling scheme.

For a particular sampling scheme, each model was run  $n_{\text{sim}}$  times. This procedure was repeated with various values of  $n_{\text{sim}}$  varying from 10,000 to 90,000 at an interval of 10,000.

A likelihood function was required for obtaining behavioural solutions. Several likelihood functions exist in the literature (see for instance, Christensen 2004; Moradkhani *et al.* 2005; Beven & Binley 2014) for measuring mismatch between observations and modelled series. A likelihood function adopted for this study was given by

$$L(\theta_k|Y) = \exp(-N \times \sigma_k^2 / \sigma_Y^2) \quad (3)$$

such that  $L(\theta_k|Y)$  is the likelihood measure for the  $k$ th model conditioned on observations  $Y$ ,  $\sigma_k^2$  denotes the error variance for the  $k^{\text{th}}$  model, and  $\sigma_Y^2$  is the variance of observations. The term  $N$  is an adjustable parameter to assign weights for distinguishing between good and bad solutions (Freer *et al.* 1996). The likelihood function in Equation (3) was adopted in this study because it is commonly used for GLUE (Blasone *et al.* 2008). When  $N = 1$ , the basic form of Equation (3) or  $L(\theta_k|Y) = (1 - (\sigma_k^2 / \sigma_Y^2))^2$  becomes the commonly known Nash–Sutcliffe efficiency (NSE) (Nash & Sutcliffe 1970) and may indirectly be related to the other recently introduced forms of coefficient of determination such as the revised  $R$ -squared and hydrological model skill score (Onyutha 2022). Small  $N$  leads to a flat likelihood function which extends over a wider region of the parameter space (Freer *et al.* 1996). A large  $N$  leads to a peaked likelihood function with a precise optimal solution (Blasone *et al.* 2008). To allow investigation of the possible effect of the shape of the likelihood function on the efficiency of the various sample-generating schemes,  $N$  was varied from 1 to 20.

Further comparison was made in terms of the relative bias (RB, %) or the difference between the number of retained solutions of RND and those from other sampling schemes. Values of RB were computed using

$$\text{RB} = (S_i - S_{\text{RND}}) / S_{\text{RND}} \times 100 \quad (4)$$

where  $S_{\text{RND}}$  and  $S_i$  refer to the number of retained solutions based on RND and the  $i$ th other remaining sampling scheme (and  $1 < i \leq 11$  in this study). Large  $n_{\text{sim}}$  values from 30,000 to 90,000 were considered in the computation of RB to avoid exaggerated values of RB.

### 2.3.4. Comparison of uncertainty bounds of the various sampling schemes

To construct the confidence interval (CI),  $\alpha = 0.10$  was chosen. Normally, 95% CI (or the use of  $\alpha = 0.05$ ) is common in the application of GLUE. However, GLUE has a limitation such that it is difficult to ensure many observations fall within the 95% CI especially based on available formulations (Blasone *et al.* 2008). Therefore, the use of  $\alpha = 0.10$  was preferred to  $\alpha = 0.05$ .

The initial value of  $n_{\text{sim}}$  was set to 1,000 and each model was calibrated using the  $n_{\text{sim}}$  sets of parameters' values generated by each of the selected sampling schemes. Two quantiles were considered for analysis including the 95th and 2nd percentile modelled flow events. The uncertainty bounds on each selected flow quantile were obtained as the difference between the upper and lower limits (or width) of the 90% CI. This procedure was repeated with  $n_{\text{sim}}$  varying from 10,000 to 90,000 at an interval of 10,000.

Comparison was also made in terms of the relative difference (RD, %) in the CI widths for the various sampling schemes. To do so, RD was computed using

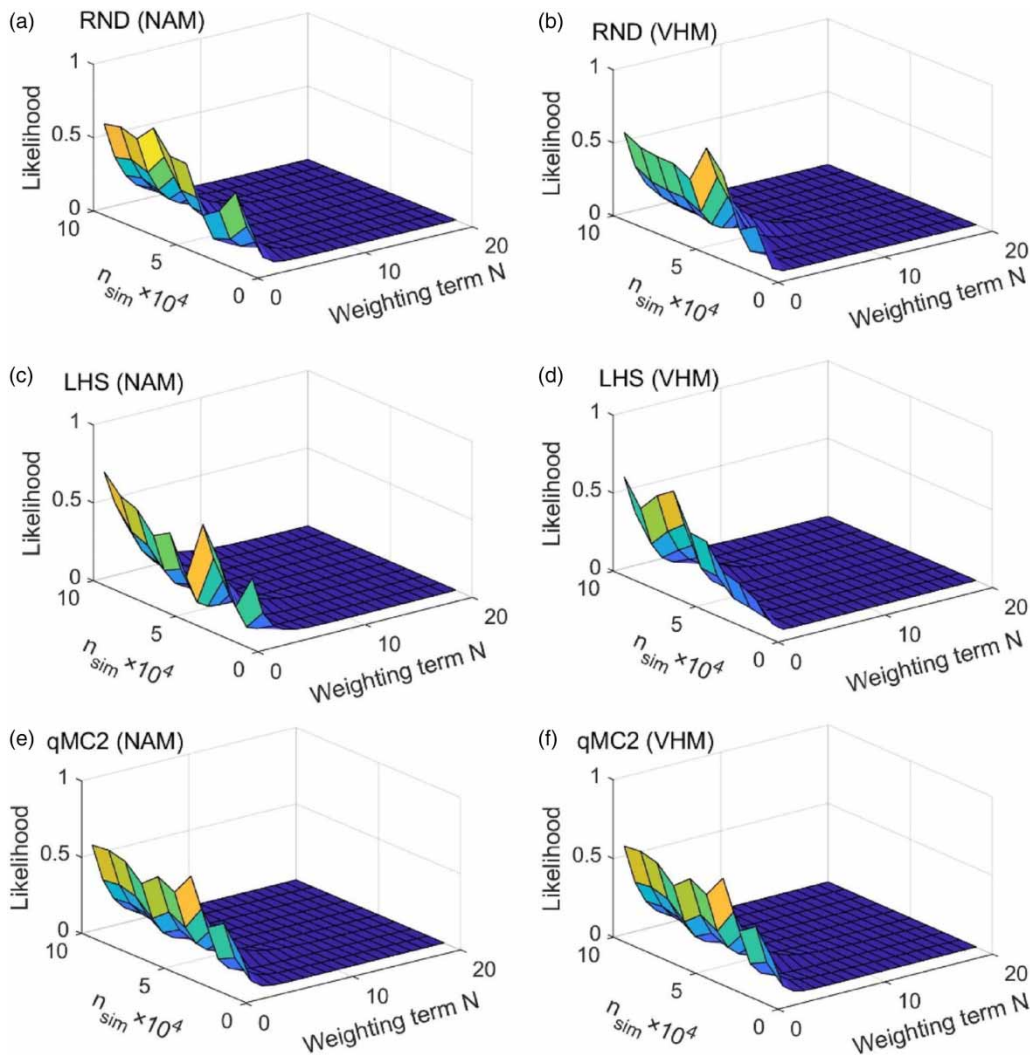
$$\text{RD} = (C_i - C_{\text{RND}}) / C_{\text{RND}} \times 100 \quad (5)$$

where  $C_{\text{RND}}$  and  $C_i$  refer to the widths of the CI based on RND and the  $i$ th other remaining sampling scheme (and  $1 < i \leq 11$  in this study).

## 3. RESULTS

### 3.1. Likelihood

Figure 2 shows likelihood based on various sampling schemes. Generally, the likelihood for a particular  $n_{\text{sim}}$  decreases exponentially with increasing  $N$  (Figure 2(a)–2(f)). On a cursory look, patterns of the variation of likelihood under various  $n_{\text{sim}}$  and  $N$  appear comparable for the different sampling schemes (Figure 2(a)–2(f)). This was for both NAM (Figure 2(a), 2(c) and

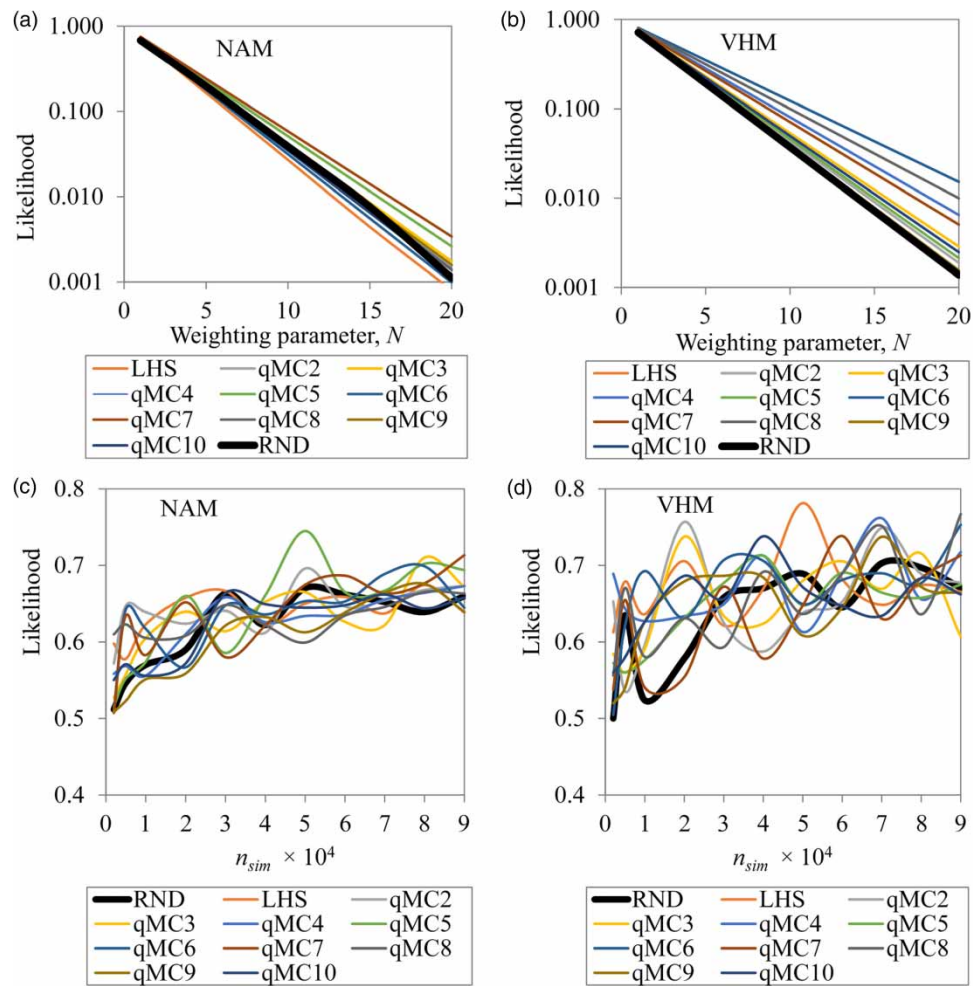


**Figure 2** | Likelihood for the various schemes applied to (a,c,e) NAM and (b,d,f) VHM.

2(e)) and VHM (Figure 2(b), 2(d) and 2(f)). These results depict the acceptability of the RBMC. Besides, patterns of the likelihood obtained by applying all the selected RBMC configurations (though shown only for qMC2, Figure 2(e)–2(f)) were also comparable.

Figure 3 shows the variation of likelihood with the term  $N$ . The plots in Figure 3 were based on values in Table 1 for  $n_{sim}$  equal to 90,000. The vertical axis of each of the plots in Figure 3(a) and 3(b) was made on a logarithmic scale for clarity. For all the values of  $n_{sim}$  (though only shown for 90,000 in Figure 3 for illustration), the likelihood value decreased with an increase in  $N$  and this was due to the exponential effect of  $N$ . For NAM, some values were above and others below that of RND (Figure 3(a)). However, the likelihood of RND based on VHM was systematically below the values of other sampling schemes (Figure 3(b)). The likelihood of the best simulation was obtained with the term  $N$  equal to 1 (Figure 3(c) and 3(d)). Generally, the likelihood tended to increase with increasing  $n_{sim}$ . For a particular  $n_{sim}$ , there were a number of RBMC configurations which yielded higher likelihood values than that of RND and this was for both models (Figure 3(c) and 3(d)). Considering results from both models, there was no sampling scheme which consistently yielded the highest (or best) likelihood value for the various values of  $n_{sim}$  and  $N$ . This demonstrates the uncertainty in calibration due to the choice of the sampling scheme. To quantify such an uncertainty, results from a large array of sampling schemes are required. Thus, the concept of having several configurations of RBMC, as demonstrated in this study, offers an important step for insights on quantifying influence from the said sub-source of calibration-related uncertainty.





**Figure 3** | Likelihood for the various values of (a,b)  $N$  and (c,d) best simulation.

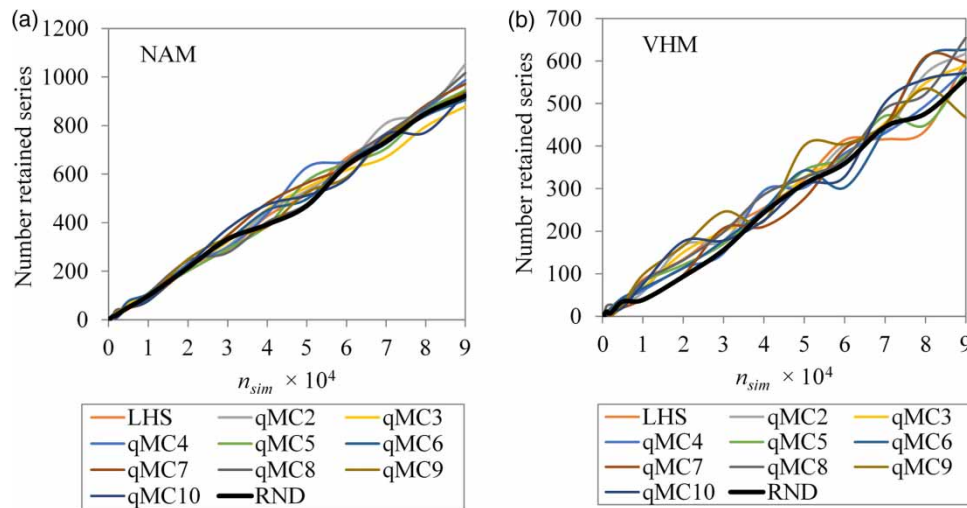
**Table 1** | Likelihood values for  $n_{sim}$  set to 90,000

Sampling scheme	$N$									
	1	5	10	15	20	1	5	10	15	20
	VHM					NAM				
RND	0.7194	0.1927	0.0371	0.0072	0.0014	0.6800	0.1967	0.0387	0.0076	0.0011
LHS	0.7166	0.1890	0.0357	0.0068	0.0013	0.6967	0.1641	0.0269	0.0044	0.0008
qMC2	0.7306	0.2082	0.0433	0.0090	0.0019	0.7248	0.2000	0.0400	0.0080	0.0016
qMC3	0.7462	0.2313	0.0535	0.0124	0.0029	0.7278	0.2042	0.0417	0.0085	0.0017
qMC4	0.7772	0.2837	0.0805	0.0228	0.0065	0.7210	0.1949	0.0380	0.0074	0.0014
qMC5	0.7351	0.2147	0.0461	0.0099	0.0021	0.7427	0.2260	0.0511	0.0115	0.0026
qMC6	0.8111	0.3511	0.1233	0.0433	0.0152	0.7074	0.1772	0.0314	0.0056	0.0010
qMC7	0.7676	0.2666	0.0711	0.0189	0.0050	0.7526	0.2415	0.0583	0.0141	0.0034
qMC8	0.7942	0.3160	0.0998	0.0315	0.0100	0.7194	0.1927	0.0371	0.0072	0.0014
qMC9	0.7236	0.1984	0.0394	0.0078	0.0015	0.7242	0.1991	0.0397	0.0079	0.0016
qMC10	0.7408	0.2231	0.0498	0.0111	0.0025	0.7147	0.1865	0.0348	0.0065	0.0012

### 3.2. Number of retained modelled series

Figure 4 shows variation in the number of solutions that were behavioural with  $n_{sim}$ . Note that the first value of  $n_{sim}$  in Figure 4 is 1,000 and not 0. The difference between any two subsequent  $n_{sim}$  values greater than 1,000 is 10,000. However, the difference between the first two  $n_{sim}$  values considered was less than 10,000. This explains why the horizontal axis starts from 0. The spread of the number of behavioural solutions increased with increasing  $n_{sim}$  although this was smaller for NAM (Figure 4(a)) than VHM (Figure 4(b)).

For certain values of  $n_{sim}$ , e.g. when  $n_{sim}$  was 50,000 and 20,000 considering NAM (Figure 4(a)) and VHM (Figure 4(b)), respectively, the numbers of retained solutions from RND were less than those from other sampling schemes. Nevertheless, the numbers of behavioural solutions generally tended to fluctuate around those for RND. Values of RB are summarized in



**Figure 4** | Number of retained solutions for (a) NAM and (b) VHM.

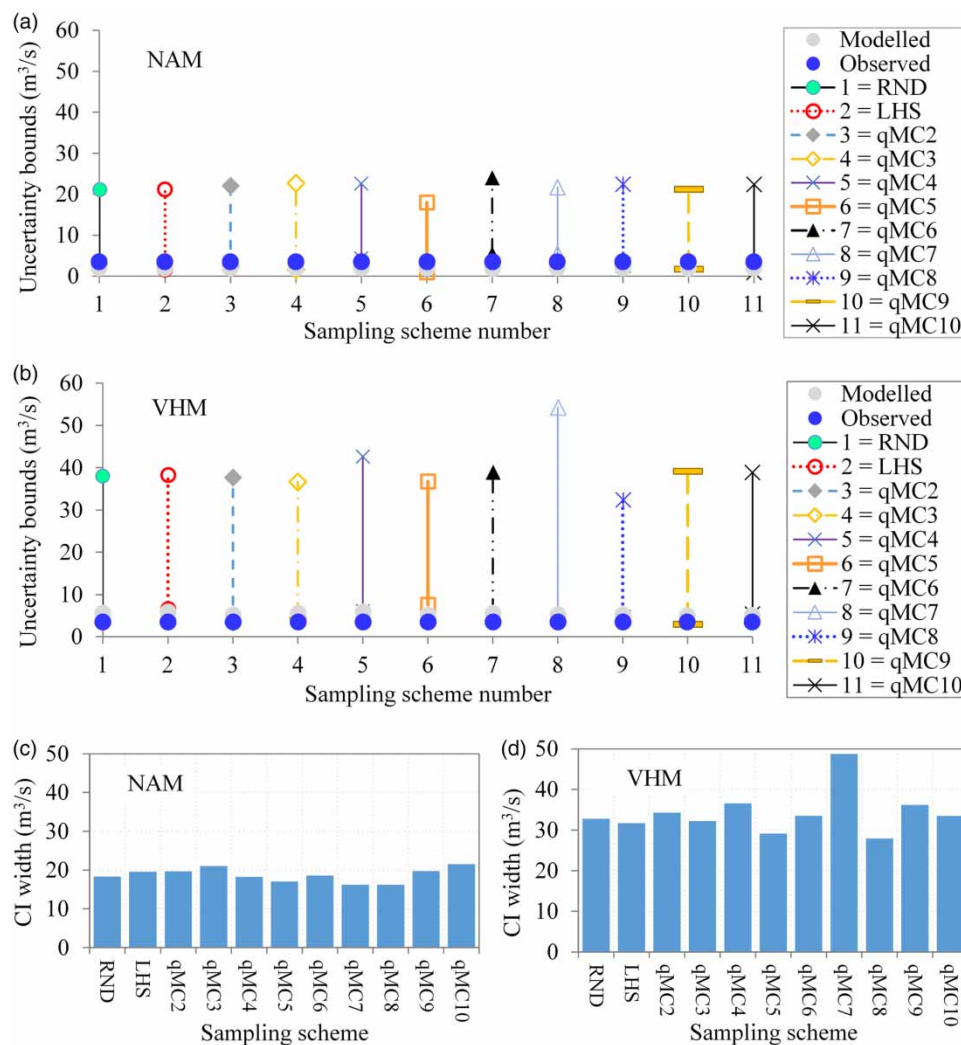
**Table 2** | Values of RB for the various sampling schemes

$n_{sim} \times 10^4$	LHS	qMC2	qMC3	qMC4	qMC5	qMC6	qMC7	qMC8	qMC9	qMC10
NAM										
3	-13.5	-9.0	-3.3	-10.5	-12.3	-1.8	4.5	-16.8	1.2	12.6
4	8.7	13.5	16.0	11.2	-1.0	14.8	22.4	2.5	-2.3	20.6
5	10.4	13.2	16.2	33.6	21.7	6.0	20.0	1.7	11.7	8.9
6	4.9	0.2	-3.0	2.5	1.3	2.5	-0.9	2.7	-7.7	-9.0
7	3.0	10.4	-8.6	-1.1	-3.8	1.2	2.3	4.6	2.5	4.0
8	0.7	0.7	-6.3	1.9	0.9	-1.2	3.9	3.2	0.5	-9.3
9	1.4	14.3	-4.7	7.0	2.7	-1.7	5.4	10.3	1.8	1.2
VHM										
3	31.8	14.1	28.2	-4.7	9.4	14.1	32.9	27.1	57.6	14.1
4	4.5	21.1	-0.8	21.1	3.0	3.0	-13.5	17.3	-8.3	-7.5
5	3.5	0.6	4.7	-2.9	10.0	10.0	-11.2	4.7	28.8	-0.6
6	15.3	12.2	5.1	6.6	4.1	-15.8	9.7	3.1	12.8	-8.7
7	-6.6	-2.9	2.1	-3.3	5.8	-0.8	0.8	9.5	0.8	12.8
8	-8.5	19.6	15.0	3.8	-5.8	27.3	28.1	9.6	12.3	16.9
9	7.5	10.5	5.6	3.9	2.6	12.1	6.9	17.0	-16.4	2.3

**Table 2.** The smaller the RB, the more comparable the numbers of series retained based on the two sampling schemes being considered. The larger the number of retained series, the better the sampling scheme. Thus, positive RB indicates better performance of a given sampling scheme than that of RND. The maximum improvement in NAM was 33.6% (qMC4) followed by 22.4 (qMC7). However, improvements in VHM went up to 57.6 (qMC9) followed by 32.9 (qMC7). This demonstrates the adequacy of the RBMC for GLUE. Importantly, the values of RB were both positive and negative even for particular  $n_{sim}$ . This further reinforced the notion that the use of a single sampling scheme such as the well-known RND of LHS for GLUE comprises an uncertainty due to the choice of the parameter sampling method.

### 3.3. Uncertainty bounds

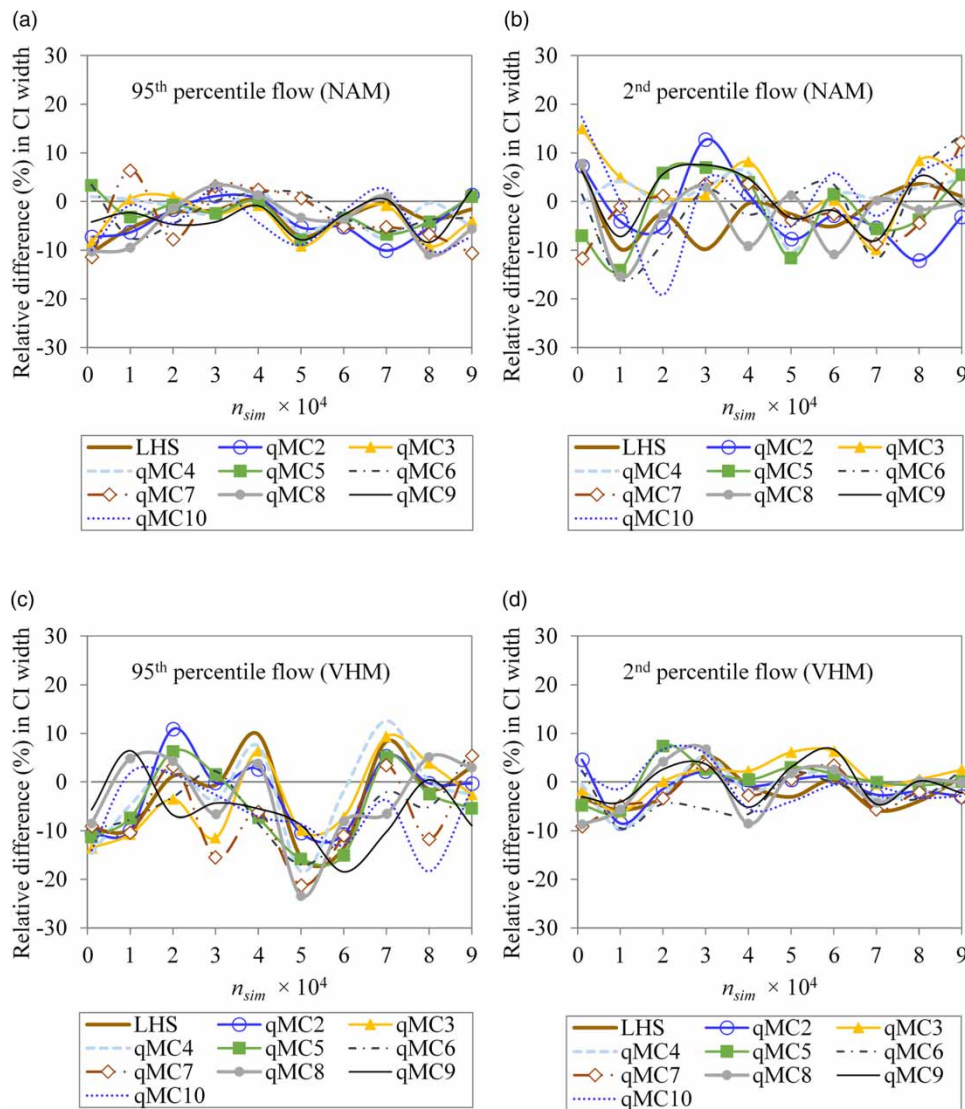
Figure 5 shows widths of 90% CI on the second percentile modelled flow quantile based on  $n_{sim}$  set to 1,000 for demonstration. The uncertainty bounds based on NAM (Figure 5(a) and 5(c)) were narrower than those of VHM (Figure 5(b) and 5(d)). Based on results from either VHM or NAM, the width of the CI varied from one sampling scheme to another. The largest CI widths for NAM and VHM were 21.5 and 48.7  $m^3/s$  based on qMC9 and qMC7, respectively. The modelled quantiles from both hydrological models were closer to the lower than upper CI limits. Assuming that the difference between the observed and modelled quantiles is minimal with the residuals being normally distributed, one would expect the modelled value to lie around the centre of the CI. However, the result from Figure 5 showed that many of the simulations from both



**Figure 5** | Uncertainty bound on the second percentile flow-based (a,c) NAM and (b,d) VHM.

models over-estimated the low flow quantile. Generally, calibration using objective functions such as NSE (as applied in this study) as a metric that is typically based on the comparison of overall water balance makes the hydrological model performance good for high flows compared to low flows. Nevertheless, the narrower the CI width, the more precise the prediction. The minimum CI widths of 16.2 and 27.9 m<sup>3</sup>/s for NAM and VHM were obtained using qMC7 and qMC8, respectively.

Figure 6 shows differences (RD) between RND and other sampling schemes in terms of the widths of the confidence intervals (CIs). The smaller the CI width, the more precise the method used. Here, the zero horizontal line can be taken as the reference indicating the bias of RND relative to itself. It is noticeable that values for a particular sampling scheme fluctuate about the reference. Positive RD indicates that the CI based on RND is more precise than that from the other sampling scheme under consideration. In the same line, negative RD values indicate improvement with respect to the precision of the prediction. The configurations qMC7 and qMC8 yielded the best improvements of CI on 95th percentile flow by 11.4 and 10.6%, respectively (Figure 6(a)). However, for the low flow quantile base from NAM, the improvements of CI relative to RND were best exhibited by qMC10 and qMC6 by 19.2 and 16.0%, respectively (Figure 6(b)). The top two improvements in CI on the 95th percentile flow from VHM were by 23.4 and 21.2% based on qMC8 and qMC7, respectively (Figure 6(c)).



**Figure 6** | Relative difference in the CI width between RND and other sampling schemes applied to model (a,b) NAM and (c,d) VHM. Due to the large difference between the first two  $n_{sim}$  values (i.e. 1,000 and 10,000), the horizontal axis starts from 0. However, the first value of  $n_{sim}$  is 1,000 and not 0.



low flow quantile from VHM, the largest improvement was by 9.5% based on qMC6, followed by 9.1% from qMC7 (Figure 6(d)). Fluctuations in NAM-based RD values for a given  $n_{\text{sim}}$  were larger for high flow than low flow quantiles (Figure 6(a) and 6(b)). However, the fluctuations in the RD values based on VHM were larger for low flow than high flow quantiles (Figure 6(c) and 6(d)). This contrast reflects the differences in the capabilities of the selected models in reproducing hydrological extremes.

#### 4. DISCUSSION

The likelihood value for the best simulation based on each RBMC configuration was shown to increase with the increasing  $n_{\text{sim}}$ . Similar results were also obtained with RND and LHS. It is worth noting that GLUE uses the 'less formal likelihoods'. Several studies (Mantovan & Todini 2006; Stedinger *et al.* 2008; Liu *et al.* 2022) argued that GLUE's use of informal likelihood functions (or soft rules) to determine the behavioural parameters normally results in very flat posterior distributions. Specifically, Mantovan & Todini (2006) demonstrated the incoherence of GLUE with Bayesian inference using a pseudo-Bayes experiment. They showed that 'less formal likelihoods' failed to add information to the conditioning process especially under the circumstance of gradually increasing the number of observations. Beven *et al.* (2007), in their response, clarified that GLUE was developed to be applied to actual calibration issues comprising errors from both model inputs and model structures. Furthermore, Beven *et al.* (2007) concluded that the demonstration by Mantovan & Todini (2006) of the incoherence of GLUE regarding the requirement that including every new observation should add information to the conditioning process could not hold.

The purpose of this paper was not to refute the argument that other MCMC methods could be more efficient and effective in obtaining a better posterior distribution of parameters than that based on GLUE. In fact, GLUE can be considered to be an extension of the Bayesian averaging approach to a less formal likelihood (Beven *et al.* 2000). The 'less formal likelihood' comprises the key aspect of the differentiation in the Bayesian inference thereby allowing for flexibility in the definition of the likelihood function to eliminate the need for strong assumptions on the error model (Jin *et al.* 2010). In the case when all the assumptions are satisfied, the use of the formal Bayesian technique becomes more acceptable given its linkage to its classical statistical theory and application of formal mathematical procedure and MCMC simulation for inferring parameter and model prediction distributions (Vrugt *et al.* 2009). It is worth noting that a direct comparison of GLUE and the formal Bayesian method is generally difficult for various reasons. Firstly, the formal Bayesian method's focus on unraveling the effects of errors due to model inputs, outputs, and model structures complicates statistical inference (Vrugt *et al.* 2009). On the other hand, GLUE does not consider separating these effects on the total uncertainty. Secondly, the formal Bayesian method makes use of an exact (or formal) likelihood function (assumed or transformed from an unknown form) to estimate the prediction precision of one-step ahead forecasting (Vrugt *et al.* 2009; Jin *et al.* 2010). However, GLUE applies an informal likelihood function to estimate prediction precision (or CI) on simulated variables (such as river flow in this case).

It is known that the sources of the total predictive uncertainty are numerous especially due to inputs (Renard *et al.* 2010; McMillan *et al.* 2018), calibration (Beven 2006), and model parameters or structure (Beven 1989; Butts *et al.* 2004; Renard *et al.* 2010). Attempts to reduce the total predictive uncertainty can be made by tackling the various sources of uncertainties. Beyond the calibration sphere, uncertainty due to the model structure becomes substantially dominant in the predictive uncertainty (Højberg & Refsgaard 2005; Rojas *et al.* 2008; Troin *et al.* 2018). Some sub-sources of calibration-related uncertainty include the choice of (i) a calibration method, (ii) an objective function, and (iii) an optimization approach. This study found that results (in terms of widths of CI, or number of behavioural solutions) of some RBMC configurations were below while others lied above those for RND. This demonstrated the influence from the choice of parameter sampling method as a sub-source of calibration uncertainty. Application of an array of the RBMC configurations presents a creditable capacity to quantify the total predictive uncertainty while offering the opportunity to understand the influence of the choice of a parameter sampling method. This property is not possessed by LHS or RND. Furthermore, the application of RND as a single parameter generation scheme conventionally used for GLUE makes it a difficult task to quantify the uncertainty due to the choice of sampling scheme. Other issues of the conventional sampling techniques especially RND and LHS are known. For instance, LHS assumes independence among the parameters (Petelet *et al.* 2010) and when some model parameters are dependent, the use of LHS in a GLUE framework can impact the number of behavioural solutions. Furthermore, the



pseudo-random sampling in LHS requires several samples to be accurate (Huntington & Lyrintzis 1998). The advantage of RND normally used in the conventional GLUE framework is that it is simple and faster to generate parameter values (Huntington & Lyrintzis 1998). However, RND has the limitation of tackling problems which require high-dimensional parameter estimation (Blasone *et al.* 2008). Given the composite nature of the introduced RBMC such that it comprises RND as one of its configurations, this paper puts forth the proposition for researchers to adopt RBMC for running the GLUE framework.

An important step in adopting RBMC for the GLUE framework is to decide on the number of RBMC configurations for application. In this study, only 10 configurations were used to demonstrate the acceptability of the introduced method. However, the number of RBMC configurations in the application of GLUE can be far larger than 10. The larger the number of RBMC configurations used, the better the uncertainty analysis. For each RBMC configuration, a separate set of uncertainty limits and modelled series can be derived. A combination of modelled series from the various RBMC configurations yields a single model ensemble. Various methods for obtaining model ensembles exist and they date back to the 1970s (Twedt *et al.* 1977) among which we have simple arithmetic averaging, and weighted mean approach (Baker & Ellison 2008).

## 5. CONCLUSIONS

Many models in hydrology are complex and make use of non-linear equations in their structures. Using analytical techniques to quantify the uncertainty in such models is a difficult task. GLUE is a calibration strategy which relies on RND. The use of RND reduces the effectiveness of GLUE in establishing behavioural solutions. This study introduced RBMC and investigated its use for GLUE. In the first step of the RBMC, the upper and lower limits of each parameter are stipulated. The next step consists of deciding on the number of intervals into which the full range of each parameter is divided. Over each interval, values of every parameter are separately generated. Here, the RND approach of generating parameter values can be used. Lastly, the various sub-blocks of the generated values of a given parameter are shuffled. During the shuffling process, the sequence of the generated values in each interval is ensured not to be affected.

The number of behavioural solutions based on RBMC was larger than that of RND, in some cases, by up to 33.6%. The widths of 90% CI on 95th percentile flow based on some RBMC configurations were smaller than those of RND by up to 23.4%. For a selected  $n_{sim}$  in the range 1,000–90,000, the numbers of behavioural solutions from RBMC fluctuated around those from RND. Similarly, the widths of 90% CI on a selected flow quantile were below and above that from RND. These findings revealed that the choice of a sampling scheme is a sub-source of calibration uncertainty. The use of RND or LHS as a single sampling scheme for GLUE is insufficient to support the quantification of the aforementioned sub-source of calibration uncertainty. In this line, the introduced method presents the key advantage that RND becomes nested within the RBMC approach. Furthermore, the introduced method takes into account the influence of the choice of a sampling scheme by offering an opportunity to select a particular number of RBMC configurations. Thus, RBMC given its demonstrated potential for uncertainty quantification is proposed to be adopted for GLUE instead of RND.

MATLAB codes for RBMC implemented in a simplified manner for calibrating HMSV lumped conceptual model (Onyutha, 2019) can be downloaded along with example modelling datasets via <https://doi.org/10.5281/zenodo.10702810> for illustration purpose. The following procedure is recommended for constructing uncertainty bounds on predictions from a model using RBMC while taking into account the influence of the choice of a sampling scheme on calibration results:

- (a) The upper and lower limits of each model parameter are stipulated, and the term  $b$  is also specified.
- (b)  $n_{sim}$  is set to a large number (e.g. 100,000) and as guided under section 2.1, relevant number parameter values are generated in each interval or sub-block of a given parameter with respect to  $n_{sim}$ ,  $n_{wa}$  and the term  $b$  stipulated in step (a). Here, a pseudo-random generator can be used to separately generate parameter values in each interval.
- (c) The total number of behavioural solutions ( $n_{bs}$ ) is also stipulated, for instance,  $n_{bs} = 2,000$ .
- (d) Threshold of the chosen objective function to generate behavioural solutions is specified (e.g.  $NSE = 0.6$ ).
- (e) The number of RBMC configurations ( $n_{con}$ ) is chosen (e.g.  $n_{con} = 20$ ). This means that RBMC is varied using qMC( $j$ ) where  $j = 2, 3, \dots, n_{con}$ .
- (f) By setting  $j$  in qMC( $j$ ) to 2 (or using qMC2), the model is set to run  $n_{sim}$  times using a *while loop*. This means that the *while loop* is terminated when the number of behavioural solutions is equal to  $n_{bs}$  stipulated in step (c).

- (g) For each of the remaining RBMC configurations (or from  $j$  in  $\text{qMC}(j)$  set to 3, 4 ... ,  $n_{\text{con}}$ ), step (f) is repeated. This leads to  $(n_{\text{con}} - 1)$  sets of behavioural solutions each of size equal to  $n_{\text{bs}}$ .
- (h) For each RBMC configuration, the series with the best value of the objective function (for instance, the highest NSE) is selected as the best-modelled series. The ensemble mean is obtained by averaging the best-modelled series based on all the considered  $(n_{\text{con}} - 1)$  RBMC configurations.
- (i) It is worth noting that for each RBMC configuration, there are  $n_{\text{bs}}$  modelled values in an attempt to reproduce each observed flow event. From these  $n_{\text{bs}}$  modelled values, the  $(100 - \alpha)\%$  CI on the modelled flow event under consideration is obtained as  $[0.005 \times \alpha\% \times n_{\text{bs}}]$ th and  $\{[1 - (0.005 \times \alpha\%)] \times n_{\text{bs}}\}$ th values, respectively. For all the selected RBMC configurations, it means that there are  $(n_{\text{con}} - 1)$  values of the upper limit of CI on a particular flow event. Similarly, there are  $(n_{\text{con}} - 1)$  values of the lower limit of CI on a flow value. To obtain the ensemble  $(100 - \alpha\%)$  CI, the  $(n_{\text{con}} - 1)$  values of the upper limit of CI on a particular flow event are averaged. Similarly, the  $(n_{\text{con}} - 1)$  values of the lower limit of CI on a flow value are averaged.

## ACKNOWLEDGEMENTS

The author acknowledges that this study made use of modelling data from Onyutha *et al.* (2021) which merged with reanalysis data published by Kobayashi *et al.* (2015), Funk *et al.* (2015), and Saha *et al.* (2014).

## FUNDING

This research received no external funding.

## DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information. Simplified MATLAB-based algorithm for RBMC sampling along with demonstration modelling datasets can be found via <https://doi.org/10.5281/zenodo.10702810>.

## CONFLICT OF INTEREST

The authors declare there is no conflict.

## REFERENCES

- Baker, L. & Ellison, D. 2008 The wisdom of crowds – ensembles and modules in environmental modeling. *Geoderma* **147**, 1–7.
- Beven, K. 1989 Changing ideas in hydrology – the case of physically-based models. *Journal of Hydrology* **105**, 157–172.
- Beven, K. 2006 A manifesto for the equifinality thesis. *Journal of Hydrology* **320**, 18–36.
- Beven, K. 2016 Facets of uncertainty: Epistemic uncertainty, non-stationarity, likelihood, hypothesis testing, and communication. *Hydrological Sciences Journal* **61** (9), 1652–1665.
- Beven, K. J. & Binley, A. 1992 The future of distributed models: Model calibration and uncertainty prediction. *Hydrological Processes* **6** (3), 279–298.
- Beven, K. & Binley, A. 2014 GLUE: 20 years on. *Hydrological Processes* **28**, 5897–5918.
- Beven, K. J. & Freer, J. E. 2001 Equifinality, data assimilation, and uncertainty estimation in mechanistic modelling of complex environmental systems using the GLUE methodology. *Journal of Hydrology* **249** (1–4), 11–29.
- Beven, K. J., Freer, J., Hankin, B. & Schulz, K. 2000 The use of generalised likelihood measures for uncertainty estimation in high order models of environmental systems. In: *Nonlinear and Nonstationary Signal Processing* (Fitzgerald, W. J., ed.). Cambridge Univ. Press, New York, pp. 115–151.
- Beven, K., Smith, P. & Freer, J. 2007 Comment on “Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology” by Pietro Mantovan and Ezio Todini. *Journal of Hydrology* **338**, 315–318.
- Blasone, R.-S., Vrugt, J. A., Madsen, H., Rosbjerg, D., Robinson, D. A. & Zyvoloski, G. A. 2008 Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. *Advances in Water Resources* **31**, 630–648.
- Boyle, D. P., Gupta, H. V. & Sorooshian, S. 2000 Toward improved calibration of hydrologic models: Combining the strengths of manual and automatic methods. *Water Resources Research* **36**, 3663–3674.
- Butts, M. B., Payne, J. T., Kristensen, M. & Madsen, H. 2004 An evaluation of the impact of model structure on hydrological modelling uncertainty for streamflow simulation. *Journal of Hydrology* **298**, 242–266.
- Christensen, S. 2004 A synthetic groundwater modelling study of the accuracy of GLUE uncertainty intervals. *Nordic Hydrology* **35** (1), 45–59.
- Dawdy, D. R. & O'Donnell, T. 1965 Mathematical models of catchment behaviour. *Journal of the Hydraulics Division ASCE* **91** (4), 113–137.

- Der Kiureghian, A. & Ditlevsen, O. 2009 *Aleatory or epistemic? Does it matter?* *Structural Safety* **31**, 105–112.
- Doherty, J. 2010 *PEST, Model-Independent Parameter Estimation – User Manual*, 5th ed.; with slight additions; Watermark Numerical Computing: Brisbane, Australia.
- Eckhardt, K., Fohrer, N. & Frede, H.-G. 2005 *Automatic model calibration*. *Hydrological Processes* **19** (3), 651–658.
- Freer, J., Beven, K. J. & Ambrose, B. 1996 *Bayesian estimation of uncertainty in runoff prediction and the value of data: An application of the GLUE approach*. *Water Resources Research* **32** (7), 2161–2173.
- Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A. & Michaelsen, J. 2015 *The climate hazards infrared precipitation with stations – a new environmental record for monitoring extremes*. *Scientific Data* **2** (1), 150066.
- Georgakakos, K. P., Seo, D. J., Gupta, H., Schaake, J. & Butts, M. B. 2004 *Characterizing streamflow simulation uncertainty through multimodel ensembles*. *Journal of Hydrology* **298** (1–4), 222–241.
- Gong, W., Gupta, H. V., Yang, D., Sricharan, K. & Hero, A. O. I. I. 2013 *Estimating epistemic and aleatory uncertainties during hydrologic modeling: An information theoretic approach*. *Water Resources Research* **49** (4), 2253–2273.
- Gupta, A. & Govindaraju, R. S. 2023 *Uncertainty quantification in watershed hydrology: Which method to use?* *Journal of Hydrology* **616**, 128749.
- Gupta, H. V., Sorooshian, S. & Yapo, P. O. 1999 *Status of automatic calibration for hydrologic models: Comparison with multilevel expert calibration*. *Journal of Hydrologic Engineering* **4** (2), 135–143.
- Hadka, D. & Reed, P. 2013 *Borg: An auto-adaptive many-objective evolutionary computing framework*. *Evolutionary Computation* **21**, 231–259.
- Højberg, A. L. & Refsgaard, J. C. 2005 *Model uncertainty-parameter uncertainty versus conceptual models*. *Water Science and Technology* **52**, 177–186.
- Hora, S. 1996 *Aleatory and epistemic uncertainty in probability elicitation with an example from hazardous waste management*. *Reliability Engineering and System Safety* **54** (2–3), 217–223.
- Huntington, D. E. & Lyrintzis, C. S. 1998 *Improvements to and limitations of Latin hypercube sampling*. *Probabilistic Engineering Mechanics* **13** (4), 245–253.
- Jin, X., Xu, C.-Y., Zhang, Q. & Singh, V. P. 2010 *Parameter and modeling uncertainty simulated by GLUE and a formal Bayesian method for a conceptual hydrological model*. *Journal of Hydrology* **383**, 147–155.
- Kavetski, D., Franks, S. W. & Kuczera, G. 2003 *Confronting input uncertainty in environmental modelling*. *Water Science and Application* **6**, 49–68.
- Kavetski, D., Kuczera, G. & Franks, S. W. 2006 *Bayesian analysis of input uncertainty in hydrological modeling: 1. Theory*. *Water Resources Research* **42**, W03407.
- Kobayashi, S., Ota, Y., Harada, Y., Ebata, A., Moriya, M., Onoda, H., Onogi, K., Kamahori, H., Kobayashi, C., Endo, H., Miyaoka, K. & Takahashi, K. 2015 *The JRA-55 reanalysis: General specifications and basic characteristics*. *Journal of Meteorological Society of Japan* **93** (1), 5–48.
- Kuczera, G. & Parent, E. 1998 *Monte Carlo assessment of parameter uncertainty in conceptual catchment models: The Metropolis algorithm*. *Journal of Hydrology* **211**, 69–85.
- Liu, Y., Fernández-Ortega, J., Mudarra, M. & Hartmann, A. 2022 *Pitfalls and a feasible solution for using KGE as an informal likelihood function in MCMC methods: DREAM(ZS) as an example*. *Hydrology and Earth System Sciences* **26**, 5341–5355.
- Mantovan, P. & Todini, E. 2006 *Hydrological forecasting uncertainty assessment: Incoherence of the GLUE methodology*. *Journal of Hydrology* **330**, 368–381.
- McMillan, H., Westerberg, I. K. & Krueger, T. 2018 *Hydrological data uncertainty and its implications*. *Wiley Interdisciplinary Reviews: Water* **5**, 1319.
- Moradkhani, H., Hsu, K.-L., Gupta, H. & Sorooshian, S. 2005 *Uncertainty assessment of hydrologic model states and parameters: Sequential data assimilation using the particle filter*. *Water Resources Research* **41** (5), 1–17.
- Nash, J. E. & Sutcliffe, J. V. 1970 *River flow forecasting through conceptual models part I – a discussion of principles*. *Journal of Hydrology* **10**, 282–290.
- Nearing, G. S., Tian, Y., Gupta, H. V., Clark, M. P., Harrison, K. W. & Weijs, S. V. 2016 *A philosophical basis for hydrological uncertainty*. *Hydrological Sciences Journal* **61** (9), 1666–1678.
- Nielsen, S. A. & Hansen, E. 1973 *Numerical simulation of the rainfall-runoff process on a daily basis*. *Nordic Hydrology* **4** (3), 171–190.
- Onyutha, C. 2019 *Hydrological model supported by a step-wise calibration against sub-flows and validation of extreme flow events*. *Water* **11** (2), 244. <https://doi.org/10.3390/w11020244>.
- Onyutha, C. 2022 *A hydrological model skill score and revised R-squared*. *Hydrology Research* **53** (1), 51–64.
- Onyutha, C., Turyahabwe, C. & Kaweesa, P. 2021 *Impacts of climate variability and changing land use/land cover on River Mpanga flows in Uganda, East Africa*. *Environmental Challenges* **5**, 100273.
- Petelet, M., Iooss, B., Asserin, O. & Lored, A. 2010 *Latin hypercube sampling with inequality constraints*. *Advances in Statistical Analysis* **94**, 325–339.
- Qi, W., Zhang, C., Fu, G., Sweetapple, C. & Liu, Y. 2019 *Impact of robustness of hydrological model parameters on flood prediction uncertainty*. *Journal Flood Risk Management* **12**, 12488.

- Renard, B., Kavetski, D., Kuczera, G., Thyer, M. & Franks, S. W. 2010 Understanding predictive uncertainty in hydrologic modeling: The challenge of identifying input and structural errors. *Water Resources Research* **46**, 1–22.
- Rojas, R., Feyen, L. & Dassargues, A. 2008 Conceptual model uncertainty in groundwater modeling: Combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resources Research* **44**, 44.
- Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y., Chuang, H., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M. P., van den Dool, H., Zhang, Q., Wang, W., Chen, M. & Becker, E. 2014 The NCEP climate forecast system version 2. *Journal of Climate* **27** (6), 2185–2208.
- Stedinger, J. R., Vogel, R. M., Lee, S. U. & Batchelder, R. 2008 Appraisal of the generalized likelihood uncertainty estimation (GLUE) method. *Water Resources Research* **44**, 1–17.
- Thiemann, M., Trosset, M., Gupta, H. & Sorooshian, S. 2001 Bayesian recursive parameter estimation for hydrological models. *Water Resources Research* **37** (10), 2521–2535.
- Troin, M., Arsenault, R., Martel, J.-L. & Brissette, F. 2018 Uncertainty of hydrological model components in climate change studies over two Nordic Quebec catchments. *Journal of Hydrometeorology* **19**, 27–46.
- Twedt, T. M., Schaake, J. C. & Peck, E. L. 1977 National weather service extended streamflow prediction. In: *Proc. 45th Western Snow Conference*, Albuquerque, New Mexico, pp. 52–57.
- Vrugt, J. A. 2016 Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB implementation. *Environmental Modelling and Software* **75**, 273–316.
- Vrugt, J. A. & Robinson, B. A. 2007 Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging. *Water Resources Research* **43**, W01411.
- Vrugt, J. A., Diks, C. G. H., Gupta, H. V., Bouten, W. & Verstraten, J. M. 2005 Improved treatment of uncertainty in hydrologic modeling: Combining the strengths of global optimization and data assimilation. *Water Resources Research* **41** (1), 1–17.
- Vrugt, J. A., ter Braak, C. J. F., Gupta, H. V. & Robinson, B. A. 2009 Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? *Stochastic Environmental Research and Risk Assessment* **23**, 1011–1026.
- Willems, P. 2014 Parsimonious rainfall-runoff model construction supported by time series processing and validation of hydrological extremes – part 1: Step-wise model-structure identification and calibration approach. *Journal of Hydrology* **510**, 578–590.
- Zhang, X. S., Liang, F. M., Srinivasan, R. & Van Liew, M. 2009 Estimating uncertainty of streamflow simulation using Bayesian neural networks. *Water Resources Research* **45**, W02403.
- Zhou, X., Liu, H., Pourpanah, F., Zeng, T. & Wang, X. 2022 A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications. *Neurocomputing* **489**, 449–465.

First received 9 August 2023; accepted in revised form 11 February 2024. Available online 22 February 2024